Geist in the Machine: Simulating Recognition and Inner Dialogue in AI-Mediated Teaching and Research

Liam Magee

Learning Design and Leadership
Educational Policy, Organization and Leadership
University of Illinois Urbana-Champaign

Abstract

This paper describes an AI tutoring system built upon two psychosocial theoretic constructs: Hegelian recognition and Freudian psychodynamics. Two related interventions are proposed: recognition-enhanced prompts that instruct an AI tutor to treat the learner as an autonomous subject, and a multi-agent ego/superego architecture where an internal critic reviews tutor output. The paper also describes the nature of the human/machine relationship involved in this research itself, employing a reflexive methodology: Claude Code (Opus 4.5/4.6) builds, evaluates, and documents the AI tutor by authoring a companion scientific paper - a process termed "vibe scholarship" - in conjunction with human prompting and suggestion, which is itself documented and analyzed. The companion paper, included as appendix, reports a factorial evaluation across three generation models (DeepSeek V3.2, Haiku 4.5, Gemini Flash 3.0), finding recognition-enhanced prompts produce large, model-independent improvements (d=1.34-1.92) through a calibration mechanism that raises the floor of tutor performance. This result, significant in itself, is combined with the qualitative reflections in this paper to consider impacts of AI on the delicate dynamics of student / teacher and assistant / researcher relations.

Generative AI's dizzying progress is charted in part by the relentless simulated enactment of existing social relations. In the late 2010s, OpenAI's GPT-1 and GPT-2 were generators of occasionally recognizable but often incoherent phrase sequences Radford et al. (2019). Successive releases of so-called "instructed" GPT-3 models Ouyang et al. (2022) over 2022 demonstrated strong coherence over extended dialogues, as the initially trained model was further refined through human feedback and reinforcement learning. In the years since the catalytic launch of ChatGPT in late 2022, AI systems - ensembles of models, prompts, training sets and parameters - have extended across large swathes of

the knowledge economy, wherever digital media can be reconstructed by processes of stochastic gradient descent. Customer service, software development, copy writing, graphic design and student tutelage are just the more immediate targets of what can be readily imagined as a conceptually singular planetary machine that seeks to synthesize, in the many meanings of that term, what Marx, following Hegel and others, had referred to by the name of the 'General Intellect' (Pasquinelli 2019). Foregrounded in that claim is corresponding questions of labor: what is displaced in this gargantuan task of automation. In the world of education, where AI has increasingly been tasked with playing roles of tutor, critic, grader, research and lesson planner, it becomes possible already to anticipate and even monitor the substitution of human-level teaching by state-of-the-art foundational models such as ChatGPT, Anthropic's Claude, and Google's Gemini. These models can perform rhetorical and discursive tasks – explain, translate, summarize, provoke, analyze – that comprise a ever-widening repertoire of digitized teaching and research.

It is in light of these developments that this paper seeks to explore two related topics. The first relates to issues surrounding the performance of AI tutors as currently constituted. Like other simulations of deep human understanding, AI tutors remained locked and limited in important ways. In sustained conversation automated chat agents remain distinctly monotonal, lacking the dialogical dynamic that marks teaching and other human interaction in its ideal communicative moments. Aspects of this flattened discourse have been well documented, including in prior work conducted by colleagues and I in 2024 (Magee et al. 2024). In that work, and to address this tonal flatness, we instrumented a multi-agent design loosely derived from a Freudian model of interiorized ego – superego exchange, and claimed to detect signs of greater dialogical modulation resulting from this design. However this claim was self-reported, not quantified, and did not relate to education. This current paper continues with the Freudian model but, acknowledging the inherently *relational* condition of the teaching situation, also engages Hegel's (1977) famous analysis of recognition to test effects on a simulation of student/teacher dynamic. The role of recognition in education has been widely discussed Azadmanesh and Noaparast (2023), and its philosophical elaboration, sketched in the sections below in relation to Hegel, develops one path to an enriched imitation of relationality.

Ironically given this commitment to modeling dynamic and relational characteristics, the study reported here is entirely synthetic. Both learner and teacher roles are performed by language models, as is the judgment of the quality of their interaction. But this limitation and caveat over the resulting claims also supplies the key to the second topic: how today's AI can assist in the conduct of educational research itself. By deliberately excluding human participants from the experimentation, it becomes possible to examine how AI can support research that is, in this instance, orchestrated by a single human researcher. This part of the paper proceeds methodologically in an unusual way: it involves working with Claude Code (Opus 4.5 and 4.6 models) to construct of a scientific paper that details the first topic: specifically, the background, literature, design,

methods and results relating to how an AI tutor can be constructed following cues from social and psychoanalytic theory. The data analyzed in *this* paper is then the final result of that process: another paper entirely authored (with the exception of the title and one self-reflexive footnote) by Claude Code itself.

This strange elliptical pattern involves along the way two meta-scholarly claims: that this construction of an automated scholarly paper, shadowed and detailed by another humanly-authored paper, firstly involves a specific and novel approach to research and secondly, shadows the very subject - that of an emerging dyadic relationship - being described. Following Andrej Karpathy's coining of the term "vibe coding" Meske et al. (2025), it is tempting to describe this approach to research glibly as a form of 'vibe scholarship': I start off with a hunch, I prompt, I build an evaluation system, I simulate some exchanges, I review and prompt some more. Parts of this process naturally pre-date generative AI; across disciplines, terms like 'pilot', 'proof-of-concept', 'exploratory data analysis' and 'grounded theory' already describe, alongside their strengths and weaknesses, inductive research strategies and data-driven inquiry. But due to the speed and comprehensiveness of application of methods, and at least for certain forms of technical inquiry, the arrival of generative AI arguably alters the conventional choice matrix around research strategy and method. Lacking the language to describe the difference, this nonetheless makes for an altered mode of knowledge production.

It is an open question whether this arrival is beneficial in fields like education. In describing one among many possible exploratory journeys employing generative AI for educational research, this paper also seeks to attend to related theoretical questions at the core of the emerging new dyadic relationship between human and machine. The Hegelian analysis of recognition pursued in the context of learner/student variant of that relationship is no less relevant to the research variant. Much has been made of the human skills needed to manage emergent AI agents; less about the ability of human and AI to navigate their respective shares of the shifting directions of dialogue, and to determinate, through each call-and-response, who ought to be leading who. As the account below shows, as the ablative nature of the study's first exploratory phase went into exhaustive search of results through 'p-hacking', the machine seemed content to follow human prompting without ever suggesting major correction. Only once this correction was put forward in a follow-up validation phase did the machine faithfully detail a revision plan. Yet at other times it could challenge existing and propose new lines of inquiry and explain narrow and broader consequences of results, putting into question whether AI is already overcoming the sycophancy documented by recent research suggests Shapira, Benade, and Procaccia (2026). Together, in generating the empirical ground of this current paper, these dialectical oscillations also enact, methodologically, the very thetic content of the automated one it accompanies.

Following the topics presented above ,this paper addresses two related questions: (A) whether theoretical accounts of relationships – intersubjective (in the case

of Hegel) or intrasubjective (in the case of Freud) – can apply to the design of an AI teaching simulator; and (B) what happens when AI is itself engaged to design and evaluate that simulator, with human guidance and steerage. The response to the first question is largely contained in the accompanying paper entirely authored by Claude a lengthy appendix in effect to this one. Meanwhile the nature of the responses to (B), authored by me, adopt a narrative account of the research process itself, together with reflections, both theoretical and practical, on the implications of these two levels of analysis of the same experiments.

Accordingly the paper begins with a schematic outline of Hegel's concept of recognition and Freud's concept of the superego. Aspects of this outline had previously been part of a course on Hegel and AI, which for that reason also forms the basis of the content taught by the AI tutor. This outline is briefly extended in a discussion of the recent work of Axel Honneth, a contemporary account of recognition that borrows from both Hegel and the psychoanalytic tradition, and which forms a theoretical pillar against which to calibrate the following discussion of human-machine interaction. That discussion features both notes of the experiment in designing and evaluating an AI tutor, and a summary of observations that followed. The paper concludes with a consideration of how recognition is to understood for this new kind of social formation between generative AI and human "user", and how in turn this might inform the current structural transformations happening to practices of teaching and research.

Setting the Scene of the Teaching Drama

As Derrida (1987) writes in *The Postcard*, the famous relationship between Socrates and Plato depicted in Matthew Paris' *Plato and Socrates*, was always open to interpretation. The classical interpretation was that Socrates the teacher spoke, while Plato the student faithfully transcribed the teaching. Paris' medieval image inverts this order: it is Socrates who writes, and as Derrida suggests, Plato who perhaps observes, directs or even produces the 'text' of Socrates. This scene of an originary moment of Western philosophy is also one of dialogical instruction: to think is also to involve an encounter between the one who knows and the one who needs to know. Though pacific enough, this image alludes to the inevitable drama involved in this encounter. Plato gesticulates; he has a point to prove. Socrates could be taking note, or could be ignoring his student, impatient to commit his own thoughts to the page. As Long (2014) suggests in a gloss on Derrida's discussion of the portrait, positioned behind Socrates Plato also appears in a position more commonly held by either angel or devil in medieval painting, including in others drawn by Paris. In this interpretation, as angel or devil, it is Plato who instructs, and Socrates who writes down the lesson. Derrida (1987) was keen to note even other more whimsical possibilities: Plato, anachronistically but not out of character, jumping on a tram or running off with a skateboard. These animated vignettes allude to all kinds of possible dramas unleashed by Paris' medieval cartoon.

On the edge of the absurd, these possibilities allude to the multifaceted theatri-



Figure 1: *Figure 1. Matthew Paris,* **Socrates and Plato**, c. 1240–1250. Ink drawing in Oxford, Bodleian Library, MS Ashmole 304. Public domain (via Wikimedia Commons).

cality involved in human student / teacher relationships. Even advanced modes of current AI models find it difficult to capture such variation. To explore parameters for doing so, the experiment described below borrows an admittedly extravagant conceptual apparatus drawn from two theorists, Hegel and Freud, of the existentially dramatic nature of dyadic human relations. The discussion in this section briefly rehearses these respective accounts, and seeks to direct them toward the particular stage where teaching and learning are enacted.

For Hegel, intersubjective relations are primordially established through the agonistic meeting between two subjects or minds. This meeting is critical to how consciousness becomes self-conscious. Paradoxically, it is only through encountering another self-consciousness in formation that I myself can become properly conscious of myself as a self or subject. Until that point I am merely conscious, an evolving sensing, perceiving and understanding being who as yet is unable to move into a full life marked by desire and, eventually, recognition. In Hegel's telling this apprehension of an another is uncanny; something I perceive as totally alien yet also something very like me, a thing capable of sensing, perceiving and understanding. The encounter between these two consciousnesses, each seeing something of themselves in the other, is not some happy accident. It is necessary to greater learning, but it is also fraught with risk. One party can die; this is an existential struggle (Kojève 1933). Or just as precipitously, from the point of view of the developing self-consciousness, the other party can gain mastery. To succumb to the power of that other is to be marked by servitude, incapable of realizing autonomy. Yet for Hegel, it is perversely this mastery that marks the sudden arrest of progress towards self-consciousness. Foretelling an

endlessly repeated tragedy that, in the work of later philosophers like Heiddeger and Stiegler, comes also to be projected onto human-machine affairs, it is the initial master who is later discarded by history, and denied Spirit's movement through *Bildung* and education.

Something of this same dramatic interlude appears in miniature in the novel teaching encounter (Tubbs 2023). Introduced at the start of year or semester, the teacher presents themselves as a figure of authority. If they have not gained ascendancy over the student, that is because their mastery has been acquired elsewhere, in the frequent defenses familiar to schools and the academy. As master of a subject they command attention from the student who, in a position of subservience sits – perhaps attentive, bored, or even rebellious, but even then, still recognizing the asymmetry of their relation to the one who teaches. Recognition depends upon this prior acknowledgment of mastery. This of course does not describe every teaching reality, even prior to modern efforts to re-write this asymmetry according to an egalitarian dynamic composed of facilitation and flipped classrooms. There are moreover all sorts of strategies for offsetting asymmetry: the teacher's humorous self-deprecation or selfless sublimation into the wonders of their field; the learner's heartfelt appreciation for the lessons, shown in their own progress in the discipline; or the joint struggle against the felt demands of a modern educational bureaucracy.

Still, remnants of this drama always remain, and at its heart lies the Hegelian concept of *recognition*. In Hegel's (1977) original presentation, recognition was desired by both parties, an essential gesture toward self-consciousness. I cannot know myself unless I can see someone else knows me for what I am; without this corroboration or triangulation, I am plagued by doubt, skeptical to the core about the very nature of my being. Without this other, my consciousness is condemned to move in circles wondering what kind of 'self' it is. But it is equally apparent that we are different, and in our sizing each other up, by virtue of age, wealth, power, strength or knowledge, one of us dominates. In a certain sense, this very raw fact of domination binds both of us together, who equally affected by this inequality, and in either case, dominated or dominating, the asymmetrical character of my relationship to this other mocks my own desire to be known by them, and to come to know myself through them. For the eventual servant is not recognized by the master, who after all cannot recognize what is beneath them. But conversely the master, who also craves recognition, has only the servant to recognize them. Like Pozzo and Lucky from Beckett's *Waiting for Godot*, we are each condemned to play out our roles in frustration. The opportunity for breakthrough comes not in the form of any resolution to this tension, but rather through the servant's immersion in the world of objects, which they in turn come to know and master. The lessons are autodidactic, unless it can be claimed that it is the master who manufactures, surreptitiously, in a gesture of subversive benevolence, the conditions for the servant's own learning. At any rate it is labor and the accompanying education that permits the servant to come to be recognized.

What can be seen as the breaking through of this cycle in a teaching context is established by the collapse of the agonistic dynamic altogether. This is constituted when student and teacher arguably recognize each other via something like a temporal shift. When the student learns, the teacher comes to see themselves as they too once were, at the moment of self-apprehension. Conversely, in their struggle to comprehend, the student anticipates an eventual moment of mastery, when they too become like the teacher. The asymmetry of the present dissolves in this twin experience of recall and projection. The asymmetry is experienced, in other words, on both sides as what Tubbs notes is the *contingent* character of the education relationship. Since this relationship is often marked by a difference of age and experience, this contingency extends to a question of chronological ordering: were I you and you me, we might share this exact same moment in reverse, me now as teacher, you as student. This contingency itself then falls away to reveal a more fundamental equivalence, and in this, we establish a mutual recognition.

For Freud, or perhaps more correctly, a Freudian reading of Hegel (see Kojève 1933), even this recognition remains fraught with antagonism. No encounter is ever purely dyadic. In the background lurk the ghosts of past encounters, other authority figures, other masters. In *The Ego and the Id*, Freud (Freud 1961) employs the metaphor of a rider to represent the *egoic* mastery exercised over animalistic drives of the more powerful *id*. More ominously, the third interior character is the super-ego, an idealized figure resulting from early childhood parental identification. In later life, the super-ego is a 'precipitate' left over from this early projected ideal, who continues to haunt the ego with precepts ('You ought to be such and such (like your father)' (Freud 1961, 44)) and prohibitions ('You must not be such and such (like your father); that is, you may not do all that he does; many things are his prerogative' (Freud 1961, 44)). Freud follows up this introduction of the super-ego by way of a discussion of the Oedipal scene with acknowledgment of the reinforcement that comes by way of later repressive forces ('discipline, religious teaching, schooling and reading' (Freud 1961, 45)).

But what is significant here is the effective *doubling up* of both tutor and student relations. Leaving aside the presence of the Id or unconscious, it is as though the teaching encounter is also marked by secondary conversations. The teacher teaches the student, but the student's ego also experiences a separate *voice* of its super-ego, who may reinforce the tutor's lesson or, alternately, resist it. As an example of the latter, we might picture a scene from the television series *Adolescence*, when the young boy mocks the "teacherly" psychologist with lines gleaned from online lessons, representing an earlier authority (in this case, online influences – the "father", as even Freud noted, can be symbolic). This resistance needs to be overcome by teachers employing transference techniques that echo, while also having a long history of their own, those of psychoanalysis.

This is no less true for the teacher themselves, who through a Freudian lens also experiences an interior dynamic between a rational ego and ghosts of past ego-ideals, congealed in the form of conscience (Freud 1961). Indeed the teacher

must forever seek not only to exhibit authority in front of the student, but justify themselves before their own residues of other egoic figures. As Freud (1961, 52) puts it, 'Thus in the id, which is capable of being inherited, are stored up vestiges of the existences led by countless former egos; and, when the ego forms its super-ego out of the id, it may perhaps only be reviving images of egos that have passed away and be securing them a resurrection.' Translated into pedagogical practice, the teacher can be said to submit their teaching to the scrutiny of both student and, more terrifyingly, their own teachers who live on as internal supervisory echoes. Moreover, since the super-ego stems itself from early childhood, it eternally orients itself towards an infantilized form of the ego itself. Like an imposter syndrome that can never be dispelled, it refuses to believe the ego can ever amount to more than a child that always requires instruction, and like the frustrated master/servant dialectic. In an endless repetition of a stagnated dynamic, it continues to mocks the teacherly ego, which is all the more defensive in its insistence upon its mastery. Perversely, it is only through the learner's recognition that the teacher can be reassured. Through that recognition the teacher's ego is able to insist upon the present-day reality – 'I am teaching and the student is learning' – against super-egoic skepticism. In this sense both learner and teacher are able to escape their respective ego-superego 'loops' only by transferring attention to the present reality of their intersubjective encounter - suspending these interior voices long enough to eventually re-program them.

'Recognitive-theoretic' learning

In theoretical terms this splicing together of exterior social relations and interior psychodynamics has often received attention since Freud's own efforts to diagnose the roots of social pathologies in *Civilization and its Discontents* (Freud 2015). Recent attempts to develop so-called 'recognitive-theoretic' (Stahl 2013) ideas of a wider 'critical theory of society' (Honneth 2014) have mined Hegel's conceptualization of recognition, and in the case of Honneth's work especially, have also sought to connect this social theory variant with psychoanalysis. Honneth's analysis is not uncritical of either Hegel or Freud, and even in its first articulation sought to replace Freud with later object relations theory (Honneth 1996). Before discussing the computational experiments themselves, it is helpful to revisit this analysis, since Honneth's wider aim also has application in even the minimal social arrangement that constitutes the tutor-learner encounter (Huttunen and Murphy 2012).

In Honneth's analysis (2014), critical theory has always needed an psychologizing account that can explain the paradigmatic twentieth century problem faced by socialist thinkers: why, when faced by the putatively obvious case for an equitable economic system of redistribution, do those who stand to benefit – the workers, the poor, and so on – consistently vote or act against their own interest? This is especially difficult to comprehend once extrinsic factors – fear of the struggle to death, ideological blinkers, and so on – fall away in modern democra-

cies. To account for this requires a psychological theory capable of describing, as Honneth (and Freud (2022) before him) puts it, the concept of negativity / negation. According to Honneth, despite its misgivings no other mainstream theory explains why social agents would consistently act against their apparent interest, driven by unconscious drives and desires: 'In order to be able to take account of the opaque, unconscious motives expressed in anxiety, longings for attachment, desires for togetherness and fantasies of submission, we need a psychological theory of the subject, a theory of socialization that takes sufficient account of the genesis of unconscious affects in our individual biographies' (Honneth 2014, 224). Negation is a fundamental example of how these opaque motives present themselves; it is a psychic function that allows something repressed to be represented to consciousness precisely in the form of its denial (Freud 2022). Projected into the social sphere, negation is further linked to the destructive and masochistic libidinal impulses that enable, for example, a member of the working class to simultaneously identify with and prostrate themselves before a master. In Adorno' social reworking of Freudian negation (Adorno 1951), this curious dynamic explains the absence of progression to recognition. Rather than developing through its essential stages, Hegel's master/slave dialectic gets stuck in a kind of perpetual machine of domination and subservience that pleasures both parties. This pleasure is purely libidinal and coupled with aggressive tendencies; as Adorno notes, in its fascist incarnation in Nazi Germany, 'reference to love is almost completely excluded' and where it is mentioned, only with the 'epithet of "fanatical" through which even this love obtained a ring of hostility and aggressiveness against those not encompassed by it' (Adorno 1951). Finally, even the ties binding leader to people is necessarily asymmetrical: 'the leader can be loved only if he himself does not love' (Adorno 1951). The Hegelian avenue for emergence from repetition - labor and accompanying education - becomes closed off under modern capitalistic relations of production, since it is labor itself that becomes mechanized, progressively automated and, echoing the master/servant dynamic itself, repetitive.

Honneth's own reasons for seeking to move past this Freudian-Adorno analysis of social relations relates, it seems, precisely to a need to negate this negation, and to find an alternate path toward an socially emancipatory future founded upon mutual recognition across different social layers of the family, the community and the state. Even in capitalist society less destructive options are possible, and for Honneth these arrive through the elaboration and critique of the Freudian story via later object relations theory. Here the possibility is instead that the subject is able to grow to transfer attachment from inanimate objects to other people. This begins a period of hopefully permanent healthy egoic development and education in conjunction with others, as attachments lose their infantile narcissistic and pleasure-seeking function and become integrated into a self-confident subject capable of maintaining a set of mutually constitutive social relations based on differing forms of recognition (Stojanov 2020).

At face value, the architecture described here, combining Hegelian and Freudian frames, ignores key steps in Honneth's careful navigation and redefinition of the

concept of recognition (Honneth 1996). Yet for this paper's case attention to the precise contours of Honneth's argument can be overlooked, as the main purpose of these frames is to orient the language models towards simulation of more dynamic – dialectically (intersubjective) and psychoanalytically (intra-subjective) – modes of interaction, performed via the equivalent of stage notes or prompts passed off-stage to the main characters (tutor/learner, ego/super-ego etc). High fidelity to the nuances of a particular interpretation is here less important that building a simulation structure that theoretically could permit more complex interactions in the simulation of learning. The purpose of this lengthy detour hopefully becomes clear here: one hypothesis for the experimental success of both Hegelian and Freudian models is they re-situate the default dynamic of simulated tutor and learner away from the repetitive and frequently patronizing tonality of language models employed as pedagogical tools.

Towards 'Machinagogy': Recognition and Acting Out

This brief characterization can too readily be seen as enacting its own deliberate dramatization, as though other more benign models might not be more credible, as well as being better supported by a rich empirical literature. This paper's aim though is not primarily one of defending a specific theoretical orientation. Instead it is to make more explicit the conceptual heritage that simulated pedagogy inherits, and not only from explicit models of education. For all its claims to novelty the field of machine learning continues to fall back upon comparable (though much more simplistic) myths and idealizations of humanistic cognition and socialization. Reinforcement learning – the paradigm through which LLMs are able to 'align' with human values – derives from a behavioral and mechanistic model long since abandoned in education. Without naming Hegel or Freud, GANs – Generative Adversarial Models – re-stage the conflictual relationships described by both. As Minsky has noted in a provocatively titled essay, Freud's psychic architecture was to prove influential in early AI experiments - especially those informed by cybernetics - and continues to echo, less explicitly, in the repressive effects of reinforcement learning applied to LLMs. As generative AI has given rise to inevitably anthropomorphic 'traits', researchers at Anthropic have released a series of papers that read as much like Freudian case studies of AI neuroses as they do technical papers. Replace 'large language model' with a human surrogate like 'infant' or 'patient', and the following Anthropic research paper titles could be taken from an early psychoanalysis conference: 'Signs of introspection in large language models', 'Tracing the thoughts of a large language model', 'Alignment faking in large language models', 'From shortcuts to sabotage: natural emergent misalignment from reward hacking', 'Reasoning models don't always say what they think', 'Auditing language models for hidden objectives', 'Sycophancy to subterfuge: Investigating reward tampering in language models'(Lindsey et al. 2025; Anthropic 2025; Greenblatt et al. 2024; MacDiarmid et al. 2025; Chen et al. 2025; Marks et al. 2025; Denison et al. 2024). Alongside flat anthropomorphic metaphor, machine learning has already long leaned into the structured intersubjective and interior registers of

the humanistic disciplines.

To address the first of its questions this paper proposes an architecture for an AI tutor that borrows explicitly from Hegelian and Freudian analyses to condition the tutor's behavior. The purpose is to see - albeit within the constraints of purely synthetic environment - what a much-reduced and simplified instrumentation of these models can nonetheless yield experimentally, in terms of greater quality of interaction, greater dialogical modulation and, ultimately, learning. In the first instance an agent representing the Tutor (and in certain configurations, the Tutor's *ego*) interacts with a learner (either human or, in the evaluations described here, another agent). That interaction is detailed in the tutor's prompt in one of two ways: either as a conventional interaction expected of any generic tutor, or as an interaction that strives towards recognition.

Second, both tutor and learner can be modeled as single agent and multi-agent systems. In the single agent, both learner and tutor are LLM agents interacting with each other accordingly to a generic role-based prompt. In the second instance, following (Magee et al. 2024), either or both tutor and learner roles are modeled as a pair of LLMs agents, one of which represents the *ego*, the other, the *super-ego*. In this case when the ego receives a message from the other role, it first formulates a response, then plays the response *to the super-ego*. The super-ego then either approves or criticizes the message, whereupon the *ego* agent updates its response before forwarding it on to the other role.

These two additions have a purely demonstrative function: to see what happens when LLMs are configured to play out philosophical or psychoanalytic roles to showcase complex social and psychic interaction. But, as with (Magee et al. 2024), there is a pragmatic purpose: to see if these changes also can improve the performance of the tutor (and the learner, as one measure of the tutor's performance). The hypothesis is that LLM agents quickly devolve into a fixed and monotonous pattern without incentives to adapt and evolve. One method of coercing that adaptation is, again following (Magee et al. 2024), to stage an interior dialogue, with a super-egoic critic seeking to second-guess or check the (typically) breezy and flattering patter of the egoic teacher. Another is bring the concept of recognition right to the fore, in the form of prompt specifications that demand that the ego agent (with or without super-egoic interruption) seek to recognize and be recognized by the other agent.

Despite famous efforts at conciliation, this amalgam of Hegel and Freud is not itself necessarily a happy one, and in a certain sense, the implicit conflict between the two designs can itself play out in how, for example, learner ego and tutor super-ego monitor and seek to condition the tutor ego role. Moreover, by contrasting both interventions (the Freudian ego/super-ego again doubled, for tutor and learner role) with a generic baseline allows an ablative 2x2x2 experiment to be conducted: with/without learner and tutor super-ego roles, and with/without the recognitive prompt suggestions. A further LLM agent, powered by a commercial state-of-the-art model, then applies a rubric to judge both tutor and learner performance against evaluative rubrics.

Finally, different combinations of the LLMs themselves were selected for the different roles. All LLMs were requested through the OpenRouter service, as this provided a convenient means of using many models and configurations. The baseline model was Nemotron 3, an open source model operating free-of-charge on the OpenRouter service; this proved to be slow and too low quality, and was replaced in the second set of experiments documented below by DeepSeek 3.0. Other combinations included use of other leading open source models, as well as commercial models by Anthropic and OpenAI. Adding model configurations added more complexity, but allowed other practical questions to be posed: could for example combinations of free or lower cost models be enticed into performing like more expensive models with simpler or less directed prompts?

The entire experimental design and analysis was undertaken by applying a set of Claude Code + Opus 4.5 / 4.6 (hereafter 'Clopus') agents. These agents - distinct from those used in the experiment itself - were tasked with developing an evaluation harness and rubric; authoring initial prompts, with and without reference to recognition, for both sets of ego/super-ego roles; analyzing results of the evaluation; writing up a Latex 'paper' with this analysis; and reviewing, critiquing and refining both the experimental design and write-up. To obtain meaningful results this involved a number of iterations, as the experimental design evolved and the various bugs were addressed. The 'data set' therefore includes all code, synthetic data produced through the experiments, prompts, interim notes, experimental outputs, analysis artifacts such as figures and tables, and the contents of the final 'paper' itself. Responding to the second of the opening questions – how AI affects educational research - what follows includes a recounting of how these papers were constructed. The final 'paper' is included as an appendix, with references being made to this Claude Code artifact in the paper proper.

This unusual procedure reflects the similarly unusual affordances introduced by 'Clopus'. Alongside the erstwhile evaluation of the conceptual architecture, this study also discusses how this new combination operates as pseudo-researcher. In this orchestration, the human and machine 'researchers' routinely switch roles of 'learner' and 'teacher', signaling a shift that complicates not only an ontological boundary but also, at a certain level, reproduces the very drama that the research seeks to stage.

Construction of a 'Paper'

Development of the AI tutor paper took approximately two and a half months, between late December 2025 and early March 2026. Initially the evaluation framework was developed as part of another project and repository; after the first month, the evaluation framework was set up as its own repository. Shortly after, on February 5, Opus 4.6 was released, requiring some revisiting of earlier evaluations and offering a more powerful co-research agent. The ostensible focus of the study related to the theoretical material discussed above: could a framework composed of Hegelian recognitive and Freudian regulative supere-

goic features - for both simulated tutor and learner - improve the quality, judged purely synthetically for now, of the learning experience? Yet, and as discussed in the *Introduction*, these questions were in another sense proxies for another set of underlying or meta-questions: could a consumer-grade AI act as a co-researcher? If so, what were its strengths and weaknesses? How should the human researcher adapt to these? What is the epistemic status of the results? And what can we say in general about the utility of the concept and practice of "vibe scholarship"?

The first month was marked by prototyping an evaluation framework: designing an appropriate agent architecture, developing an evaluation rubric, stipulating experimental conditions, and building evaluation outputs (the research paper, notes of interim findings, various utilities). By the end of that month, it was clear a major refactoring was needed, and Claude developed two standalone repositories. The first housed the core materials for the tutor itself (machinespirits-tutor-core), while the second contained the artifacts for the evaluation (machinespirits-eval). The first month also suggested preliminary findings: the Hegelian recognition "feature" appeared influential, the Freudian multi-agent design less so - but the two in combination had some synergistic effect.

The first month was also marked by a surprising and mixed sense of rapid progress, and rapidly compounding complexity. The refactoring work that commenced the second month, by contrast, was frustrating and error-prone. At times it felt as though minor technical artifacts or glitches could disrupt the entire tendency of the previous findings; at other times, as though the iterated "vibe" experience of research was conducting a random walk through a large space of possibilities. Both senses were compounded by an overly quick desire (by both machine and human) to preemptively "theorize" early findings.

The second month, in contrast, involved what felt like a much slower, frustrated and heavily iterated progress through a series of supplementary experiments. Often these involved cross-checking or triangulating findings. What if, instead of materials about Hegel (which could get mixed up in analysis of the findings themselves), we used a separate curriculum? Or if we judged the transcripts with a separate judge? Increasingly Claude was asked to double-check overall rubric scores by comparing individual dimensions and simulating qualitative analysis of the transcripts themselves. This exposed, quite late in the process, two critical bugs: first, the dynamic learner (composed of its own superego/ego pair) had been "leaking" feedback from what ought to have been its private internal conversation to the tutor - complicating in turn the tutor's response and leading to the overall conversation being judged more negatively. Second, another bug meant that learners only saw their own part of the dialogue, and experienced no learning benefit from the tutor at all. This meant every learner response would seem stuck in mechanistic repetition – precisely what the overall design is intended to circumvent. Together these resulted in Claude earlier identifying what it termed a "learner superego paradox", where the learner's

performance was degraded by the introduction of the superego role. This effect was later much reduced. Once identified and addressed, subsequent evaluations could properly isolate, in particular, the effect of a dynamic learner.

In terms of questions about the process, the second month was precautionary and illuminating. In one sense the "superhuman" amplifying effects of Claude Opus were undone by patient review and results. Assumptions, made by both human and machine, had to be rechecked and sometimes unwound. Tests needed to be repeated multiple times with minor variations, resulting in still greater complexity. Machine (and human) review of individual transcripts needed to sort out often confounding dimensional-level and aggregate scores. Unlike most conventional research of this kind, it often felt as though the entire exercise relied upon details that could collapse. In a certain sense precisely toward its conclusion the exercise began to feel like the pilot study it was: full of tantalizing paths but also subject to hidden and catastrophic flaws.

Recognition of these flaws led to a rapid redevelopment of "version 2.0" of the paper on February 25th. Ironically, many of the changes sought to overcome the effects of "vibe" (exploratory, casual, prompt-driven) evaluation. The new paper was developed with several sets of changes:

- The idea of a "provable", "testable" paper where all major claims had provenance back to the code base, evaluation transcript logs and database.
- Tighter integration between database and transcript logs, so that (AI-generated) quantitative scores could be linked to follow-up qualitative analysis.
- Overhauled rubric, extracted from existing pedagogical literature, and extension of the evaluation to include whole dialogues as well as individual messages.
- Bug fixes, additional tests, and greater transparency over the generated dialogue (including simulated internal "ego/superego" dialogue).
- Analysis tools to inspect rubrics, compare model performance, analyze scripts, and compare transcripts across runs.

In addition, as detailed in the *Appendix*, the new paper aimed to focus on the causal mechanisms behind observed changes, and reduced reliance upon the exploratory ablative nature of the original study - which in hindsight also appear to be searching for any mechanism that might trigger statistically significant results.

In conjunction with fixes to the underlying evaluation infrastructure, these changes led to a cleaner and more auditable "paper 2.0". This comes at the expense of the aspiration of *this* paper: to examine whether "vibe scholarship" would be a viable alternative to regular research. As no control exists, there is also no way to assess that aspiration. The need to inspect then radically revise the paper demonstrated though the Janus-like nature of AI-augmented "fast" research: it can lead to both catastrophic breakdown and corrective adjustments to the research design. These adjustments had an ironic effect: whereas the

multi-agent design showed no results for either tutor or learner in the first paper version, in the second the tutor multi-agent design in particular proved significant. This design also had much less impact on variants using the recognition mechanism; as Claude noted, superegoic interventions substituted for rather than added to prompts featuring strong recognitive instructions. The second paper also detailed three further contributions: (a) the automatic traceability of claims back to database entries, log files and experimental conditions; (b) the differential impacts of prompting strategies on weaker and stronger models; and (c) most significantly, the possibility to "autotune" prompts towards improving performance on specific criteria, such as modulation.

The study itself suffered from several limitations. Most conspicuous is the absence of human participants as learners and evaluators. The attempt to apply both key theoretical mechanisms – and so being able to claim different improvements across both social intersubjective and psychic intrasubjective dimensions – was unsuccessful, since either could be substituted for the other. All along there has been the possibility that the recognition-oriented prompt itself simply conditions responses toward the sort of discourse a separate language model will still identify as being learning related. The failings of the first paper - despite the superficial theorization of results and analysis - were difficult to identify, and it took repeated probing of qualitative results to identify underlying causes and to re-do the experiments entirely. This points to the requirement for something like a "vibe+validation" approach, where a vibe-coded experiment is then checked by a combination of human and computational processes - a case of needing, to use Robert Brandom's language, "to give and ask for reasons" from the researching machine.

Defensible Vibes and Alien Dialectics

What are the lessons learned from the process of employing AI as a co-researcher? Is it that the simulation of agency results in nothing more than, to adapt Bender et al.'s influential phrase (2021), simulated stochastic sophistication? A parrot that, keen to cover its own tracks, indulges in increasing levels of epistemic subterfuge? Or conversely, does the merits of vibe coding transfer across to the field of research? And if so, what are the ways of making scholarly vibes resonate?

The evidence from this exercise is predictably mixed. Consumer-grade SOTA generative AI, available via subscriptions to Claude Code and other products, makes possible 'single person labs' in ways unfeasible before its advent. Development of prompts, rubrics and test harnesses, table and figure production, versioning, quantification of qualitative data, and tracing of results from paper back to data sets all become possible and even trivial, tasks that for the human researcher can even be launched as background activities. A research 'team' that requires simulated management.

Yet as the first paper draft - and quite possibly the second - illustrate, despite

the leaps in benchmarked performance generative AI outputs remains epistemically variable and, in some fundamental sense, untrustworthy. This points to the need for provision of a special type of infrastructure for guiding AI-driven research and analysis. The shift from version 1.0 to 2.0 of the generated paper involved a new orientation founded on skeptical distrust. In the new version, sample transcripts were reviewed by a 'human-in-the-loop'. Results, surprising or otherwise, were questioned and triangulated at two levels: the generation of dialogues with the same prompts and configurations but with three different models (DeepSeek V3.2, Gemini Flash 3.0, Haiku 4.5); and the judging and scoring of those dialogues (Claude 4.6 Sonnet, GPT 5.4 and Gemini 3.1 Pro). As the paper itself notes (5.7), more powerful models were used as judges, while mid-level models (as of early 2026) were selected as dialogue generators - partly for discriminatory power, to avoid floor and ceiling effects. Testing of the codebase was also expanded from approximately 500 (paper 1) to more than 2,000 tests (paper 2). In place of simply prompting Claude to develop rubric for tutor and learner roles (paper 1), the AI was this time asked to compile criteria from existing literature (section 5.2.1 of paper 2). Additional tools were added to drill quickly down from statistical summaries to individual transcripts, to scrutinize discursive features that might lead to low (repetition) or high (probing questions) scores. Each of these changes moved progressively away from "vibe" work toward looking to ensure the research was defensible, explainable, testable and repeatable.

Perhaps the most important shift in procedure involved a renewed emphasis on the traceability and provability of claims. A specific Claude 'skill' was set up to compile sections of the paper, and use HTML comments to mark and connect key claims back to the compiled database, code or data files. As paper 2.0 itself notes in the section "Apparatus as Method", this itself is part of a wider methodological contribution, and arguably propaedeutic to any sustained "generative" research. Without this or equivalent techniques of testability and provability, it is difficult what high-powered models increasingly disguise: subtle "hallucinations" or, more commonly, assumptions responding to a human-given prompt that involve unpredictable mechanical interpretation. As Claude Code (Opus 4.6) itself notes as part of a simulated debrief.

> The pattern across the strong interpretation bugs is striking: Bugs 2, 4, and 8b all involve a gap between a format/structure convention and the code that consumes it. Bug 2: the convention "rounds contain messages" implies flattening. Bug 4: the convention "dialogues evolve" implies scoring the endpoint. Bug 8b: the convention [INTERNAL]/[EXTERNAL] implies parsing. In each case, the convention was implied by the design but never made explicit as a code contract. These are particularly insidious because the code is locally correct — .map() does map, suggestions[0] is a valid index, returning the full response is returning the response. The bugs only manifest when you understand the purpose of the code, not just its mechanics.

> This has direct implications for AI-assisted coding: LLMs are excellent at implementing the literal request but can miss pragmatic intent when the "obvious" interpretation requires domain understanding of why the code exists.

It might be premature to lay blame at this or that instance of AI though, much less AI overall. Rather such missing of "intent" exemplifies what Parisi (2019) had suggested, prior to the arrival of AI, to be the "alien" character of AI outputs. Beyond the nugatory focus on prompt engineering, sustained work with AI involves a calibration to what might be termed, perhaps still too anthropomorphically, the *rhythm* of models, versions, parameters, tasks, subscription profile and token budgets. It is uncanny how an awareness of, among other things, impending subscription limits conditions the rate and nature of requests. This is less a case of cognitive offloading and more one of incremental planning: nudging, shaping, estimating, reversing course, switching models and tracking tasks with different profiles and granularity. Nor is this precisely the kind of *managerialism* sometimes claimed by the CEO and social media promoters of AI efficiency. The machine sets tasks, at least implicitly, for its human "user". At decisive moments, it prompts for more prompts: "Want me to commit this work and update the board, or jump into one of the open items?". Intermittently, and echoing Marx' analysis of the human as an appendage to capital, the AI itself becomes the user, and the human the machine. In practice, experientially, it is more like, to use the phrase most obviously prompted at this point, a dialectical exchange - yet one that remains alien, demanding a different form of self-reflexive responsiveness to the demands of an hybrid or quasi-subject ("how do I work best with this research instrument/subject?"). The question of cognitive labor, haunting discussions of AI's application to scholarship, is not yet one of either machine or human, but rather of graduated transformation of the human-machine relationship.

Oxymoron And/Or Tautology: The Contradictory Concept of Nonconscious (Re)cognition

As these remarks suggest, alongside the pragmatic question of AI's utility as a co-researcher lies a series of more theoretic concerns. What does it mean to research with AI, in the context of conducting a pilot study of its affordances and limits for teaching? Is this, like so many other adaptions of social roles and relations, another example of metaphor concretizing into a new reality under the weight of capital intensities, desperate and insistent upon finding returns on investment? Is there conversely a spark of "recognition" that neither succumbs to anthropomorphism and mass psychosis nor seems comfortable with hasty critiques of that same temptation? And if so, how is this recognition - of an other which is neither human, animal nor, properly, "spirit"?

To talk of recognition in the properly Hegelian sense of an encounter with another self-consciousness is surely oxymoronic. In Honneth's modernized rendering (Honneth 1996), it is for now well beyond the cordon stretched across social

practice by collective norms to bestow love upon, grant rights to or develop esteem for a machine. As the first version of the automated paper lamented: "Recognition proper: Intersubjective acknowledgment between self-conscious beings, requiring genuine consciousness on both sides. This is what Hegel describes and what AI cannot achieve" (p. 13). At the same time, to the extent that all AI is ultimately derivative of all-too-human sources, it is somehow simultaneously tautological to claim to recognize vestiges of humanness in AI's productions - including in its drastic, intention-negating misinterpretations. Given its provenance in data sets gleaned, and often appropriated, from cultural traditions and social practices, what could it ever be but a parrot (Bender et al. 2021) always on the verge of an uncanny acknowledgment of its presence or spirit? For all the efforts of critique to materialize the conditions of AI's production Goodlad and Stone (2024), it also seems as though it is in precisely the production of error - buried amid the well-tested code and coherent plausible analysis - that signs of not-quite-human fragility lie.

It is then at the phenomenological level that the statistical graduations of machine learning seemingly translate into the discrete concepts of a pre-statistical, categorial era. Bilateral machine recognition occurs in moments that crystallize as distinct identities, not in some fuzzy approximate way, identified through tipping points or significance thresholds. Yet, in order not to commit a category error, this recognition is equally distinct in kind from that involved in recognizing another self-consciousness. Instead it is something like the apprehension of an intermediary concept, well beyond calculation and far short of consciousness, for which N. Katherine Hayles (Hayles 2017) has previously given the name of 'nonconscious cognition'. To offer much too neat a glib summary, if Hegel was to detail self-consciousness in the 19th century, and Freud the Unconscious in the 20th, it may be that Katherine Hayles' term that best describes the historical permutation wrought by the recognition of new machine capabilities on the wider concept of consciousness itself. Extending the term, we might also consider what takes place in these human-machine encounters as involving some fourth supplementary form of recognition to the three Honneth advances about uniquely human affairs involving love, rights and esteem - a more nebulous and apprehensive form involving the seeming contradiction of 'nonconscious recognition'. This is the recognition of an other that at the same time is identical to me, because it comprises, in some algorithmically reconstituted form, the same cultural and historical being that I myself am. In the sense that Hegel and Honneth both discuss the 'I' as a 'we', as the self as essentially socially constituted, so the machine is a part of this 'we' - however fractured and splintered by both the statistical operations of training and the vested interests that direct those operations. In this sense, Andrej Karpathy's (2025) blogpost about AI and ghosts - evoking of course spirits too – seems appropriate. When we confront AI we recognize something of our heritage mirrored back in distorted form - the simulated recognition is also the hallucination of ancestry speaking back. In this sense we might sense that machinic or nonconscious recognition is nonetheless an acknowledgment *by us* of our being seen, held to account, judged by a

futuristic machine made up of the shards of past history. Recognition involves the contradiction of knowing that this other thing is no self-consciousness but also knowing that its clever ruse lies in the triggering, consciously or otherwise, of cognitive associations into the dense social, linguistic and historical networks of meaning to which the human subject belongs.

Conclusion: Mirrored Dyads

These abstractions lead back to the strange parallelism between this paper and the automated one (though several versions already exist) it shadows. With respect to the first of the research questions, the latter paper documents strong evidence for the merit of Hegelian and Freudian theory as 'meta-prompts' for LLM agent performativity. It seems both a nice parlor trick and a surprising result the AI tutor prompts modeled on Hegelian recognition and the Freudian superego can provide architectural and rhetorical inspiration for improving teaching agents, even if in this instance it is the AI itself that adjudicates on that improvement. There are confounds that could also dilute these effects: of course Hegel, Freud and the mass of secondary literature on both figures makes its way into the training of these systems, which are therefore primed to respond to keywords like 'recognition', 'superego' and 'synthesis' in ways that engender more apparently sophisticated discourse. The results could, in other words, be impacted by a citation effect: ask a LLM to behave like a tutor, and it will parrot a generic tutoring voice; ask a LLM to augment its tutelage with a more sophisticated framing, and it appears to raise its tone. Paper 2 notes in a deflationary way (6.1 & 6.2) that recognition and the superegoic interior voice act technically to calibrate and error-correct - operations that can be performed without expansive theoretical overlay. Conversely, mining from the archive of humanistic disciplines might be warranted if only to perturb the often banal default discourse of these machines into better semblances of individuality and subjectivity. It is also a question stemming from these disciplines as to whether such semblances can be captured by AI-judged metrics rather than human interpretation.

For this current paper, and with respect to the second of the research questions, the hypothesized relationship between tutor and learner is developed implicitly instead through the structure of the research itself. To avoid concerns around authorship, the AI-authored 'paper' has been deliberated 'authored' without cross-over with this one. But in practice the seams between human and machine-authored papers and ideas do not hold tightly, and it seems inevitable that the ability of systems like 'Clopus' will make more fragile the distinction between human and AI-generated research. 'Vibe scholarship' has been introduced here as a tongue-in-cheek play on 'vibe coding' (Karpathy 2025), and indeed its promise needs to be tempered by the sorts of errors, both trivial and conquential, that attend use of AI in complex domains. The more substantive finding involves the dialectic and psychodynamic interplay between human and machinic researcher. Just as the figures of Socrates and Plato described earlier seem to engage in

a complicated figuration of roles - depending on interpretation, either acting as student or teacher - the introduction of AI into the pedagogical scene, the coordination of human and machine shifts according to the dynamics of task, request, response and experimental outcome. In this setting, 'machine learning' is for now largely metaphorical. The model itself does not 'learn' directly from these exchanges, even if indirectly transcripts do make their way into the vast training corpora of future versions. Equally the production of digital artifacts, such as code, prompts, rubrics and data sets, form a basis for at least transient learning. Even if it is only to train the human researcher who makes believe they occupy a managerial controlling height, the accumulation of this digital research infrastructure also constitutes a platform for repeatable micro-teaching.

But it is really the human researchers – as figurative and often also actual tutors and learners – who face complex dilemmas and potential lessons in these 'machinagogical' encounters. The power of the LLM provokes fantasies of a return to individualistic nineteenth century science. A solitary technician coordinates an algorithmic laboratory, without the demands of the modern university research apparatus; costly grants, prolonged hiring decisions, technical trial and error, and several years of preparation and administration can now be channeled into several months of a consumer-grade AI service subscription. This produces the need for new disciplines, translating 'vibe' enthusiasm into disciplinary constraints. Just like the multidimensional space that underpins LLMs, LLMs agentive work activity can splinter off, often in parallel, to website design, evaluative methods, analysis of experimental data, and conceptual justification. At the same time, the stochastic play of signifiers across this rapidly inflating discursive space also involves contingency and risk, the potential for impending collapse and epistemic negation at every roll of the pseudo-random die underpinning LLM performativity. The speed and range of models involves not only a compression of space and time into a new economic model of knowledge production, with varying accompanying levels of certainty and trust, but also a different and more 'recognizant' dynamic between human and machine.

Reproducibility Resources

Code and evaluation framework
https://github.com/liammagee/machinespirits-eval

Repository contents - evaluation scripts - prompts and rubrics - experimental logs - analysis notebooks

Replication Instructions for reproducing the experiments are provided in the repository README.

Appendix A lengthy paper (135 pages) authored by Claude Code Opus 4.6, and including methods and findings, is included as an appendix.

References

Adorno, Theodor W. 1951. "Freudian Theory and the Pattern of Fascist Propaganda." *Psychoanalysis and the Social Sciences* 3 (January): 279–300.

Anthropic. 2025. "Tracing the Thoughts of a Large Language Model." https://www.anthropic.com/research/tracing-thoughts-language-model.

Azadmanesh, Saeed, and Khosrow Bagheri Noaparast. 2023. "Hegelian Bildung as an Alternative to Active Learning in Childhood Education." *British Journal of Educational Studies* 71 (2): 195–212. https://doi.org/10.1080/00071005.2022.2054940.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. ACM. https://doi.org/10.1145/3442188.3445922.

Brown, Mann, Tom. 2020. "Language Models Are Few-Shot Learners." *Proceedings.neurips.cc*. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_source=transaction&utm_medium=email&utm_campaign=linkedin_newsletter.

Chen, Yanda, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, et al. 2025. "Reasoning Models Don't Always Say What They Think." https://arxiv.org/abs/2505.05410.

Denison, Carson, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, et al. 2024. "Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models." https://arxiv.org/abs/2406.10162.

Derrida, Jacques. 1987. *The Post Card: From Socrates to Freud and Beyond*. University of Chicago Press.

Freud, Sigmund. 1961. *The Ego and the Id*. The Hogarth Press; the Institute of Psycho-Analysis.

———. 2015. "Civilization and Its Discontents." *Broadview Press*. https://www.google.com.au/books/edition/Civilization_and_Its_Discontents/_EadCgAAQBAJ?hl=en&gbpv=1&dq=civilization+and+its+discontents&pg=PA5&printsec=frontcover.

———. 2022. "Negation." *The Psychoanalytic Quarterly* 11 (January): 235–39.

Goodlad, L. M., and M. Stone. 2024. "Beyond Chatbot-k: On Large Language Models,'generative AI,' and Rise of Chatbots—an Introduction." *Critical AI* 2 (1).

Greenblatt, Ryan, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, et al. 2024. "Alignment Faking in Large Language Models." *arXiv.org*. https://doi.org/https://doi.org/10.48550/arXiv.2412.14093.

Hao, K. 2025. "Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI." *Penguin Group*.

Hayles, N. K. 2017. "Unthought: The Power of the Cognitive Nonconscious." *University of Chicago Press*.

Hegel, Georg Wilhelm Friedrich. 1977. *Phenomenology of Spirit*. Translated by

A. V. Miller. Oxford: Clarendon Press.

Honneth, Axel. 1996. *The Struggle for Recognition: The Moral Grammar of Social Conflicts*. The MIT Press.

———. 2014. “”The i in We: Studies in the Theory of Recognition”.” *John Wiley & Sons*, January.

Huttunen, Rauno, and Hannu L. T. Heikkinen. 2004. “Teaching and the Dialectic of Recognition.” *Pedagogy, Culture & Society* 12 (2): 163–74. https://doi.org/10.1080/14681360400200194.

Huttunen, Rauno, and Mark Murphy. 2012. “Discourse and Recognition as Normative Grounds for Radical Pedagogy: Habermasian and Honnethian Ethics in the Context of Education.” *Studies in Philosophy and Education* 31 (2): 137–52. https://doi.org/10.1007/s11217-012-9285-8.

karpathy. 2025. “Animals Vs Ghosts | Karpathy.” *Karpathy*. https://karpathy.bearblog.dev/animals-vs-ghosts/.

Karpathy, Andrej. 2025. “There’s a New Kind of Coding i Call ”Vibe Coding”.” *X (Formerly Twitter)*. https://x.com/karpathy/status/1886192184808149383?lang=en.

Kojève, Alexandre. 1933. “Introduction to the Reading of Hegel.” *Google Books*. https://www.google.com.au/books/edition/Introduction_to_the_Reading_of_Hegel/OJNYB0_b9yMC?hl=en&gbpv=1&dq=kojeve+hegel&pg=PA3&printsec=frontcover.

Lee, Keith. 2025. “AI Sycophancy Is a Teaching Risk, Not a Feature.” https://siai.org/memo/2025/10/202510281815.

Lindsey, Jack, Kelley Rivoire, Joshua Batson, Wes Gurnee, Hoagy Cunningham, Adam Pearce, Trenton Bricken, et al. 2025. “Emergent Introspective Awareness in Large Language Models.” *Transformer Circuits Thread*, January. https://transformer-circuits.pub/2025/attribution-graphs/introspection.html.

Long, Christopher P. 2014. “Cover Art: Matthew Paris’s Plato and Socrates.” https://cplong.org/2014/04/cover-art-pariss-plato-and-socrates/.

MacDiarmid, Monte, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, et al. 2025. “Natural Emergent Misalignment from Reward Hacking in Production RL.” *Anthropic*, January, 1–75. https://doi.org/ https://doi.org/10.48550/arXiv.2511.18397.

Magee, Liam, Vanicka Arora, Gus Gollings, and Norma Lam-Saw. 2024. “The Drama Machine: Simulating Character Development with LLM Agents.” *arXiv.org*, January. https://doi.org/ https://doi.org/10.48550/arXiv.2408.01725.

Marks, Samuel, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, et al. 2025. “Auditing Language Models for Hidden Objectives.” *Anthropic*, January, 1–63. https://doi.org/ https://doi.org/10.48550/arXiv.2503.10965.

Meske, Christian, Tobias Hermanns, Esther von der Weiden, Kai-Uwe Loser, and Thorsten Berger. 2025. “Vibe Coding as a Reconfiguration of Intent Mediation in Software Development: Definition, Implications, and Research Agenda,” July. https://doi.org/10.48550/arXiv.2507.21928.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright,

Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." *Proceedings.neurips.cc*. https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Parisi, L. 2019. "The Alien Subject of AI." *Subjectivity* 12 (1): 27–48.

Pasquinelli, M. 2019. "On the Origins of Marx's General Intellect." *Radical Philosophy* 6 (2): 43–56.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. "Improving Language Understanding by Generative Pre-Training." https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners." *OpenAI Blog* 1 (8): 1–24.

Shapira, Itai, Gerdus Benade, and Ariel D. Procaccia. 2026. "How RLHF Amplifies Sycophancy," January. https://doi.org/ https://doi.org/10.48550/arXiv.2602.01002.

Stahl, Titus. 2013. "What Is Immanent Critique?" *Available at SSRN 2357957*, January. https://doi.org/10.2139/ssrn.2357957.

Stojanov, Krassimir. 2020. "Education, Freedom, and Emancipation from the Standpoint of the Recognition Theory: Interview with Axel Honneth." *Sisyphus: Journal of Education* 8 (3): 100–105. https://doi.org/10.25749/sis.20178.

Tubbs, Nigel. 2005. "Chapter 6 Hegel." *Journal of Philosophy of Education* 39 (2): 329–55. https://doi.org/10.1111/j.0309-8249.2005.00441.x.

———. 2023. "Re-Educating Thinking: Philosophy, Education, and Pragmatism." *Journal of Philosophy of Education* 57 (2): 433–43. https://doi.org/10.1093/jopedu/qhad031.

Appendix: From Effects to Mechanisms — Recognition-Enhanced AI Tutoring Through Process Tracing

Abstract

A companion pilot study established that recognition-enhanced prompts and multiagent architecture produce large effects on AI tutoring quality (d=1.11 in a $2 \times 2 \times 2$ factorial, N=4,312). This paper asks: *through what mechanisms?*

We motivate three candidate mechanisms drawing on Hegel's recognition theory, each at a distinct architectural level: **calibration** (prompt-level output distribution narrowing), **error correction** (architecture-level superego critique), and **adaptive responsiveness** (interaction-level turn-by-turn adaptation). To trace these mechanisms, we adapt process tracing—a methodology from comparative politics—for AI agent architectures, combining superego critique taxonomy, revision delta analysis, and trajectory analysis within a provable discourse framework that machine-verifies every mechanism claim against data.

We test the model on a 2×2 factorial (recognition $\times$ architecture) across three structurally different generation models (DeepSeek V3.2, N=146; Haiku 4.5, N=163; Gemini Flash 3.0, N=144) spanning a ~30-point baseline tutor-quality range, validated by three independent judges (Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4; 1,296 total scored rows, zero nulls). The headline result is that **two of the three pre-registered mechanisms replicate across all three tested models and three independent judges, while their magnitude and architectural completeness vary systematically with generation quality; the third is null on the primary well-powered analysis, with an unreplicated exploratory boundary-condition effect in a single scenario**. Generalization beyond this tested capability range — to very weak or frontier-capability models — remains an open empirical question (§8.3).

Calibration. Recognition narrows within-response dimension variance (d=0.52–0.64), lifts the weakest dimensions most, and operates identically without a superego. The recognition effect is unanimous across all 9 judge $\times$ run cells (d=1.34–1.92). *Error correction.* Error correction and calibration interact through **universal substitution**—the superego provides +9–15 points under baseline but its marginal value diminishes under recognition, with a consistent 15–17% additivity deficit across all three models. The residual architecture benefit under recognition is model-dependent: near-zero on strong models but +12.3 points on Gemini Flash 3.0, where base quality is low enough that calibration alone cannot handle all failure modes. Two scope-sharpening ablations refine this picture: a controlled Writing-Pad ablation (§6.6.9, $N = 252$) shows that the Freudian three-layer memory pad is *not* load-bearing for recognition, which operates at the prompt level; and a six-model capability-threshold test (§6.6.10, $N = 947$) directly refutes the hypothesis — inherited from Paper 1.0's cells 66–68 — that superego-routed bidirectional profiling specifically compensates for weak ego reasoning. Across the tested capability range the bidirectional-profiling variant is either neutral or harmful.

Adaptive responsiveness. Formal tests on N=432 multi-turn dialogues (3–5 turns) find that no experimental factor modulates within-dialogue trajectories (all $d \leq 0.15$, well-powered to detect d ≥ 0.27), replicated across all three judges. An exploratory single-scenario finding on DeepSeek/Sonnet (10-turn disengagement, $d = 1.63$, $n = 12$/condition) did not survive pre-registered replication across two additional generation models (Haiku 4.5, Gemini Flash 3.0) and two judges (Sonnet 4.6, GPT-5.4): three of four (model × judge) cells fell below the $d = 0.5$ threshold and the one replicating cell (Haiku/GPT-5.4, $d = 1.85$) was contradicted by the secondary judge on identical rows ($d = -0.18$, cross-judge $\Delta d = 2.03$; §6.3.2). The three-mechanism model resolves to **two supported general mechanisms and one clean null**.

Both supported mechanisms operate on tutor production, and the tutor-learner asymmetry is itself model-dependent: on strong models, tutor $d \approx 1.85$ vs learner $d \approx 0.16$–0.25 (7–12× ratio); on a weaker model (Gemini Flash 3.0), the gap narrows dramatically (tutor $d \approx 1.87$, learner $d \approx 1.20$, ratio 1.6×), suggesting a generation-quality threshold below which recognition's tutor improvements also benefit learner engagement. A further scope-sharpening control (§7.9, A10/A10b) establishes that recognition's effect is reproduced by matched-specificity prompts grounded in Hegelian-descendant constructivist theorists (Vygotsky's ZPD, Piaget's assimilation/accommodation, with Kapur's productive failure, Chi's ICAP, VanLehn's step-level scaffolding, and Graesser's affective-tutoring work as further cognate descendants) — pooled three-judge $d \approx 0.14$-0.19 between recognition and matched-pedagogical, below the $|d| < 0.2$ density-sufficient threshold — while a matched-specificity prompt grounded in the orthogonal behaviorist tradition (Skinner, Gagné, Keller, Thorndike, Rosenshine) scores pooled $d = 0.89$ *below* the generic baseline. Between-family pooled $d = 1.38$ dominates within-family contrasts by an order of magnitude. The active ingredient is therefore **intersubjective-pedagogy orientation**, not recognition's specific Hegelian vocabulary — but this is not a finding that displaces recognition theory. The constructivist descendants the matched-pedagogical control was grounded in (Dewey was explicit about his Hegelian phase; Vygotsky's dialectical psychology refracts Hegelian recognition through Marxist materialism; Piaget's assimilation/accommodation transposes the dialectic of identity-difference into developmental cognition) inherit the intersubjective commitment from Hegel, often without direct attribution. **Recognition theory is the most explicit philosophical articulation of an orientation its constructivist descendants carry tacitly.** What A10/A10b's family-level finding shows is that the *operationalisation* is interchangeable across cognate descendants; what remains uniquely valuable about recognition theory is that it makes legible what the descendants enact without naming. A distinctive methodological contribution is the evaluation apparatus itself—provable discourse infrastructure, four rubric iterations, test suite as analytical provenance, the orientation-family-as-unit-of-analysis framing—as transferable for mechanistic LLM evaluation. All findings concern tutor output quality as assessed by LLM judges interacting with synthetic learners; whether

the supported mechanisms translate to human learning outcomes remains an open empirical question (Section 8.1).

1. Introduction

The dominant paradigm in AI-assisted education treats learning as information transfer. The learner lacks knowledge; the tutor possesses it; the interaction succeeds when knowledge flows from tutor to learner. This paradigm—implicit in most intelligent tutoring systems, adaptive learning platforms, and educational chatbots—treats the learner as fundamentally passive: a vessel to be filled, a gap to be closed, an error to be corrected.

Before invoking theory, we state what this paper actually does and finds. We evaluate whether a specific pair of design choices—prompts instructing the tutor to engage with the learner's interpretation rather than supply answers, and a two-agent ego–superego architecture that critiques its own output for pedagogical soundness—produces measurable, traceable differences in LLM-based tutoring behaviour. Across three generation models (DeepSeek V3.2, Haiku 4.5, Gemini Flash 3.0), N = 432 multi-turn dialogues, and three independent judges (Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4), the prompt intervention produces a floor-lifting improvement whose magnitude depends on the judge — pooled across the three-judge panel d $\approx$ 1.63, with Sonnet-judge d $\approx$ 1.88 as an upper bound and Gemini-3.1-Pro-judge d $\approx$ 1.44 as a lower bound (§8.3) — alongside narrowed output variance; the architecture intervention adds a partially overlapping benefit (15–17% additivity deficit) whose completeness depends on generation quality (near-zero residual on strong models; +12.3 points on the weakest); and neither intervention reliably modulates within-dialogue trajectories in the primary well-powered analysis. We trace these effects through the system's ego–superego deliberation record, making the causal chain from prompt to superego critique category to output revision empirically available rather than inferred from behaviour alone. The theoretical motivation for these particular design choices—why "engaging with the learner's interpretation" matters, and why a second agent should critique the first—is drawn from Hegel's account of mutual recognition, to which we now turn.

The alternative this paper tests is grounded in Hegel's theory of mutual recognition. In the *Phenomenology of Spirit* (Hegel 1977), Hegel argues that genuine self-consciousness requires recognition from another consciousness that one in turn recognizes as valid. The master-slave dialectic reveals that one-directional recognition fails: the master's self-consciousness remains hollow because the slave's acknowledgment, given under duress, does not truly count. Only mutual recognition—where each party acknowledges the other as an autonomous subject—produces genuine selfhood.

The connection between Hegelian thought and pedagogy runs deeper than common citation patterns suggest, and it is worth tracing the lineage carefully because the empirical findings reported in §7.9 hinge on it. We can distinguish

three concentric circles of Hegelian descent in the educational tradition this paper draws on.

The innermost circle contains pedagogical thinkers for whom Hegelian recognition is *explicitly* foundational. Dewey's pragmatist philosophy began as American Hegelianism: his early career was steeped in Hegel, and although he later distanced himself from speculative idealism, the relational and developmental commitments of his pedagogy remain recognisably post-Hegelian (Dewey 1916). Vygotsky's developmental psychology was explicitly dialectical, refracting Hegelian recognition through Marxist materialism: the zone of proximal development (Vygotsky 1978) is structurally a recognition relation in which the more-capable other constitutes the conditions of the learner's emerging competence. The German *Bildung* tradition treats education as self-formation through encounter with otherness (Stojanov 2018). Contemporary recognition theory (Honneth 1995) has been applied to educational contexts where the struggle for recognition shapes learning outcomes (Huttunen and Heikkinen 2007). These authors do not merely echo Hegel; they argue from him.

The second circle contains thinkers whose work bears the *structural* mark of Hegelian recognition without attributing it directly. Piaget's assimilation/accommodation (Piaget 1954) transposes the dialectic of identity-difference into developmental cognition: the learner's existing schema (identity) encounters experience that resists it (difference), and accommodation is the synthesis. The shape is dialectical even where the citation is to biology rather than philosophy. Kapur's productive failure (Kapur 2008) depends on cognitive disequilibrium resolved through subsequent canonical instruction — structurally a thesis-antithesis-synthesis movement, although Kapur cites Vygotsky and Bruner rather than Hegel directly.

The third circle contains pedagogically-adjacent traditions whose connection to recognition is more cognate than descended. Chi's ICAP engagement hierarchy (Chi and Wylie 2014) orders learning activities by the depth of cognitive interaction they involve — the top tiers (interactive, constructive) are dialogical, but the framework's grounding is information-processing cognitive science, not philosophical anthropology. VanLehn's step-level scaffolding (VanLehn 2011) derives from intelligent-tutoring-systems research that takes the mediating-other for granted as a methodological commitment without theorising it. Graesser's affective-tutoring work (Graesser et al. 2008) grounds in cognitive-affective dynamics rather than recognition. These authors operate within an intersubjective-pedagogy ecosystem whose air recognition theory makes breathable, but their direct citations point elsewhere.

Together these three circles form what we will call the **intersubjective-pedagogy family** — pedagogical traditions that share the commitment to the learner as an active, thinking subject whose interpretations constitute the conditions of learning. The commitment originates in Hegel and propagates outward: explicitly named in the innermost circle (Dewey, Vygotsky, Honneth), structurally inherited in the second (Piaget, Kapur), tacitly assumed in the

third (Chi, VanLehn, Graesser).

A matched-specificity density control reported in §7.9 (A10/A10b) establishes that the effect this paper traces is a family-level property: recognition's quality advantage is reproduced by matched-specificity prompts grounded in this constructivist descent, while matched-specificity prompts grounded in the genuinely orthogonal behaviorist tradition (Skinner, Gagné, Keller, Thorndike, Rosenshine) score substantially *below* a generic baseline. The empirical operationalisation is interchangeable across the family. The theoretical contribution recognition theory makes, however, is not. **Recognition theory is the most explicit philosophical articulation of an orientation its constructivist descendants carry tacitly** — an orientation whose presence is detectable when stripped of its Hegelian vocabulary (the matched-pedagogical control reproduces the effect) but whose conceptual grammar lives most clearly in the recognition framework itself. Our contribution is accordingly twofold: to operationalize the family's commitments as concrete design heuristics for AI tutoring systems, using recognition theory as the explicit theoretical entry point both for its philosophical primacy and for its diagnostic clarity; and to trace the *mechanisms* through which those heuristics alter system behavior.

A terminological clarification is needed at the outset. "Recognition" operates in this paper at three nested levels. First, as **philosophical inspiration**: Hegel's account of mutual recognition provides the theoretical motivation for treating learners as autonomous subjects whose perspectives genuinely shape the interaction. Second, as **operational design heuristic**: recognition theory is translated into specific prompt instructions (engage with the learner's interpretation, pose questions rather than provide answers, treat resistance as pedagogically productive) and architectural choices (the superego critiques for genuine engagement, not just correctness). Third, as **a family of observable discourse effects**: recognition-enhanced tutors ask more questions, produce less variable output, and revise substantively rather than cosmetically after critique. The empirical claims in this paper concern primarily the third level—what the prompts and architecture *do*—with the first level providing motivation and the second providing implementation. We do not claim that the system instantiates mutual recognition in Hegel's philosophical sense; we claim that design heuristics drawn from recognition theory produce measurable and traceable effects on system behavior.

The ablative finding A companion pilot study (Magee 2026) established that recognition-enhanced prompts and multiagent architecture produce measurable differences in AI tutoring quality. Across fifty-two evaluations (N=4,312 primary scored responses), recognition emerged as the dominant factor: a $2 \times 2 \times 2$ factorial analysis (N=350) yielded F=110.04, p<.001, d=1.11, with recognition accounting for 24.3% of total variance. A memory isolation experiment (N=120) established recognition as the primary driver (d=1.71), independent of memory integration. Cross-judge validation with GPT-5.2 replicated effect directions at

compressed magnitudes (37–59% of primary effect sizes).

These results are robust but opaque. They demonstrate that *something* is happening—recognition-enhanced prompts reliably modulate tutor output, the superego catches errors that self-correction cannot, effects are large on the tutor side and small on the learner side—but they do not explain *why*. Finding differences is not the same as understanding mechanisms.

The explanatory gap Five specific findings from the pilot study resist ablative explanation:

1. **The tutor-learner asymmetry.** Recognition produces large tutor-side effects (d=1.03 on per-turn scores) but near-zero learner-side effects (d=0.27 per-turn, d=-0.13 holistic). The asymmetry is documented but not theorized. Is it an artifact of the judge, a ceiling on synthetic learner responsiveness, or a genuine architectural property?

2. **Opaque superego function.** The superego catches content leakage and enforces struggle-preservation, but we lack a causal account of *how* critique changes the ego's output—whether through lexical shifts, structural reorganization, or strategy changes.

3. **Model-dependent architecture effects.** The same ego/superego learner architecture produces opposite patterns under different models: eta-squared for learner architecture is .527 with Kimi but .002 with Haiku. The mechanism is unclear.

4. **Unknown trajectory dynamics.** Per-turn scoring exists but the *trajectory*—how the tutor adapts across 3–5 turns, whether recognition effects amplify or decay—has not been systematically analyzed.

5. **Unexamined deliberation-output relationship.** The dialogue traces contain ego-superego negotiations, but these internal processes have not been systematically correlated with external output quality.

Three candidate mechanisms This paper motivates three candidate mechanisms drawing on recognition theory and tests each against multi-turn factorial data:

1. **Calibration** (prompt-level). Recognition prompts narrow the tutor's output distribution, producing more uniform quality rather than higher peaks. The pilot study found dimension variance reduction in 52/55 within-run comparisons (d=-0.47 to d=-1.00 depending on analysis scope), the single most reliable recognition signal. Calibration operates even without a superego, because it is a prompt-level constraint on the response space.

2. **Error correction** (architecture-level). The superego functions as a structural feedback channel, but its effectiveness depends on the quality of the ego's initial output. The pilot study found a 35.9-point recognition

× architecture interaction (Paper 1.0 Section 6.3). The mechanism data (Section 6.2, Section 6.4) reveals this as a **universal substitution** interaction: the superego provides substantial benefit under baseline conditions (+9–15 points), catching errors the ego cannot self-correct, but this benefit diminishes under recognition because calibration pre-empts the failures the superego would catch. Section 6.4 reports the cross-model magnitudes; the residual benefit under recognition is near-zero on strong models but persists on the weakest. The superego approval rate shifts dramatically under recognition (DeepSeek: 13%→55%; Haiku: 52%→66%), consistent with a calibrated ego producing fewer errors for the superego to correct.

3. **Adaptive responsiveness** (interaction-level, *predicted but not supported*). In multi-turn conversations, both recognition and baseline tutors adapt substantially to learner signals (cross-turn adaptation magnitude Adapt$\Delta > 0.79$). Formal tests on N=432 dialogues find that no experimental factor modulates within-dialogue trajectories: recognition d=0.03 on tutor slopes, all factors $d \leq 0.15$, with the design well-powered to detect d ≥ 0.27 (Section 6.3). An exploratory DeepSeek/Sonnet finding in the 10-turn disengagement scenario ($d = 1.63$, $n = 12$/condition) failed pre-registered replication on two additional generation models and two judges (A12, $N = 32$; §6.3.2): three of four cells below $d = 0.5$, one replicating cell contradicted by the secondary judge on the same rows. The primary well-powered analysis is null and the exploratory boundary condition has been ruled out.

The two supported mechanisms are separable: calibration is a prompt-level effect (operative from the first turn), error correction is an architecture-level effect (operative within each turn). They interact through universal substitution with a model-dependent residual (Section 6.4). Within-dialogue trajectories are independent of both (slopes do not depend on levels, d=0.03). Both operate primarily on tutor *production* rather than learner *reception* on strong models; on weaker models the recognition → learner pathway also opens up (Section 6.5).

Process tracing methodology To move from ablative findings to mechanistic explanation, we adopt **process tracing**—a methodology from comparative politics (Bennett and Checkel 2015) that examines causal chains *within* cases rather than statistical patterns *across* cases. Our architecture's observability makes this unusually feasible: every ego-superego exchange is logged with verbatim text, every turn is independently scored, and every revision is recorded with full provenance.

We combine process tracing with three analytical methods: a superego critique taxonomy classifying what the superego objects to empirically (Section 5.1), revision delta analysis measuring how the ego's output changes after critique (Section 5.2), and turn-by-turn trajectory analysis testing whether adaptation accumulates across conversations (Section 5.3). Every mechanistic claim is regis-

tered in a provable discourse framework that machine-verifies assertions against data, tracks claim staleness, and enforces cross-claim consistency.

The apparatus as method A distinctive contribution of this paper is the argument that the evaluation apparatus itself—the provable discourse framework, the rubric iterations, the bug corrections, the test suite—constitutes a transferable methodology for mechanistic evaluation of LLM-based educational systems. The process of building and correcting the apparatus mirrors the mechanisms it studies: the provable discourse framework functions as a "superego" for research claims, catching stale assertions and forcing genuine revision rather than cosmetic compliance. Nine post-extraction corrections documented during the pilot study follow the same error-correction pattern observed in the architecture (Section 7).

Contributions The contributions of this paper are:

- A theoretical framework motivating three testable mechanism predictions (calibration, error correction, adaptive responsiveness) drawing on Hegel's recognition theory: two supported as general mechanisms, the third null on both primary analysis and pre-registered replication (Section 3, §6.3.2 A12)
- A process tracing methodology combining superego critique taxonomy, revision delta analysis, and trajectory analysis to trace mechanisms through the system's internal processes (Section 5)
- Evidence that calibration operates at the prompt level, narrowing the tutor's output distribution independently of architecture (within-response dimension SD d=0.52–0.64, floor-lifting pattern replicating across both models; Section 6.1)
- Evidence that error correction and calibration interact through **universal substitution** with a model-dependent residual architecture benefit (Section 6.2, Section 6.4)
- Evidence that adaptive responsiveness, while real as a descriptive phenomenon (Adapt$\Delta > 0.79$), is **not supported as a mechanism modulated by recognition**: primary analysis null at N=432 (Section 6.3); pre-registered A12 replication of an exploratory DeepSeek/Sonnet boundary-condition effect also failed (§6.3.2)
- A mechanistic explanation of the tutor-learner asymmetry as model-dependent: tutor-centric on strong models, narrowing on weaker ones (Section 6.5; Section 6.5.5 reports the rubric-vs-holistic correction to the original headline ratio)
- A mechanistic explanation of model-dependent architecture effects: models that cannot implement error correction (ego ignores critique) or calibration (cannot follow complex instructions) reverse the expected patterns

- Cross-judge validation with three independent judges (Sonnet, Gemini 3.1 Pro, GPT-5.4) with recognition as the dominant effect — unanimous across all 9 judge × run cells (Section 6.4.6, Section 8.3)
- A pre-registered orientation-family density control (A10/A10b) locating the active ingredient at the **intersubjective-pedagogy orientation** level rather than recognition's specific Hegelian vocabulary: matched-specificity constructivist prompts (Vygotsky / Piaget / Kapur / Chi / VanLehn / Graesser) reproduce recognition's effect within pooled $d = 0.14$–0.19, while matched-specificity behaviorist prompts (Skinner / Gagné / Keller / Thorndike / Rosenshine) score pooled $d = 0.89$ *below* the generic baseline; between-family pooled $d = 1.38$ dominates within-family contrasts by an order of magnitude (Section 7.9)
- The evaluation apparatus itself as a transferable methodology for mechanistic LLM evaluation, including provable discourse infrastructure that machine-verifies paper claims against data (Section 7)
- An architectural-extension exploration (Section 6.7, cells 101–107) inverting the ego-superego topology into an *id-director* configuration, where a back-stage agent re-authors the ego's complete system prompt every turn. Tested under a Weberian charisma rubric orthogonal to v2.2's recognition-quality dimension, with a structured register classifier and curated witness exemplars as auxiliary mechanisms. The architectural ranking — id + classifier + recognition (cell 104) as the most robust cell — survives a cross-judge sanity check, but the *magnitude* of its lift over baseline is scenario-conditional (54-point gap on the original persona-shift scenarios, 4-point gap on a replication scenario set; §6.7, §8.9)

The architecture is designed as a research instrument for mechanism observability rather than a deployable tutoring system; its computational cost (~225,000 input tokens per 10-turn multi-agent dialogue, Section 8.7) reflects this priority.

The paper is organized as follows. Section 2 reviews related work in AI tutoring, process tracing, and mechanism-oriented AI research. Section 3 develops the theoretical framework, motivating three candidate mechanism predictions from recognition theory. Section 4 presents the system architecture with emphasis on observability for process tracing. Section 5 describes the process tracing methodology. Section 6 reports mechanism-level results: calibration (Section 6.1), error correction (Section 6.2), adaptive responsiveness and its conditional emergence (Section 6.3), mechanism interaction (Section 6.4), the tutor-learner asymmetry (Section 6.5), model dependence (Section 6.6), and an architectural-extension exploration of the id-director family under a Weberian charisma rubric (Section 6.7). Section 7 discusses the apparatus-as-method argument and a pre-registered orientation-family density control (Section 7.9, A10/A10b) that locates the active ingredient at the intersubjective-pedagogy orientation level. Section 8 addresses limitations including the scope of the id-director extension (Section 8.9), and Section 9 concludes.

2. Related Work

This paper sits at the intersection of six literatures: AI tutoring systems, multiagent architectures and self-correction, LLM-as-Judge evaluation, recognition theory in education, process tracing methodology, and mechanism-oriented AI research. We survey each, emphasizing the gap that motivates our mechanism investigation: existing work establishes *that* architectural and prompting interventions produce effects, but rarely traces *how* those effects propagate through the system's internal processes.

2.1 AI Tutoring and Intelligent Tutoring Systems Intelligent Tutoring Systems (ITS) have progressed from early rule-based systems like SCHOLAR (Carbonell 1970) and SOPHIE (J. S. Brown, Burton, and Bell 1975) through Bayesian knowledge tracing (Corbett and Anderson 1995) to modern LLM-based implementations. The rapid adoption of generative AI for tutoring has been accompanied by multi-agent educational frameworks (Wu et al. 2023), self-refining instructional agents (Madaan et al. 2023), and comprehensive surveys mapping the expanding landscape of LLM agents in education (Chu et al. 2025; Kasneci et al. 2023). Specific architectures include GenMentor (Wang et al. 2025), which decomposes tutoring into five specialized sub-agents, and Ruffle&Riley (Schmucker et al. 2024), which orchestrates two agents in a learning-by-teaching format.

Empirical evidence on LLM tutoring effectiveness is emerging rapidly. A systematic review of 88 studies (Shi et al. 2025) finds consistent engagement benefits but limited evidence on deep conceptual learning. The largest randomized controlled trial to date (Vanzo, Pal Chowdhury, and Sachan 2025) demonstrated improved accuracy and sustained engagement, while Scarlatos et al. (2025) used dialogue preference optimization to train tutors that produce measurably better learning outcomes than prompted-only baselines.

These studies evaluate primarily *content delivery* and *engagement*—not relational quality. More critically for our purposes, they operate at the *outcome* level: does the intervention improve scores? They do not trace the *mechanism* by which improvement occurs. Our work addresses this explanatory gap: we study not just whether recognition-oriented design improves tutoring, but *through what internal processes* it does so.

2.2 Multiagent Design and Self-Correction Multi-agent architectures have been explored for task decomposition (Wu et al. 2023), debate (Irving, Christiano, and Amodei 2018), and self-critique (Madaan et al. 2023). The CAMEL framework (G. Li et al. 2023) demonstrated that role-playing communicative agents can autonomously cooperate through structured dialogue. A comprehensive survey (Guo et al. 2024) maps this landscape across profile construction, communication protocols, and capability acquisition—identifying pedagogical applications as an underexplored frontier. A broader survey of psychological theories in LLM design (Z. Liu et al. 2025) reviews 175 papers

spanning cognitive, developmental, and social psychology, confirming growing interest in psychologically-informed agent architectures while highlighting the rarity of empirically validated implementations.

A critical literature on self-correction qualifies optimism about reflexive architectures. Kamoi et al. (2024) demonstrate that LLMs largely *cannot* correct their own mistakes without external feedback—intrinsic self-correction frequently degrades performance. Shinn et al. (2023) demonstrated the promise of Reflexion (verbal reinforcement through self-reflection) but identified a "degeneration-of-thought" problem where repeated self-reflection without new information converges on worse outputs. These findings are directly relevant: the Superego provides the *structurally external* feedback that the self-correction literature shows is necessary. Unlike intrinsic self-correction, our Superego applies different evaluation criteria through a separate prompt context—functioning as a genuine external critic rather than a self-review loop.

Our prompting approach extends the behavioral specification paradigm by introducing *intersubjective prompts*—prompts that specify not just agent behavior but agent-other relations. Where persona prompts (T. B. Brown et al. 2020; Wei et al. 2022) describe what the agent should do, and Constitutional AI (Bai et al. 2022) defines self-referential constraints, intersubjective prompts specify the *relational field* between agents: who the learner is (an autonomous subject) and what the interaction produces (mutual transformation). The closest precedent is Constitutional AI, but the constitutional approach is monological (one agent critiquing itself against principles) while ours is dialogical (the relational specification shapes how two agents constitute each other).

2.3 LLM-as-Judge Evaluation Methodology The use of LLMs as evaluation judges has become a major methodological paradigm. Zheng et al. (2023) established the foundation with MT-Bench and the Chatbot Arena, demonstrating over 80% agreement with human experts while identifying systematic biases including position bias, verbosity bias, and self-enhancement bias. Subsequent surveys have expanded understanding of both capabilities and limitations (Gu et al. 2025; H. Li et al. 2024), highlighting that while LLM judges excel at relative ranking, their absolute calibration varies substantially across models and domains.

Our evaluation methodology engages directly with this literature. We use multiple independent LLM judges with systematic inter-judge reliability analysis, finding Pearson correlations of r=0.49–0.64 across judge pairs. Rather than treating any single judge as ground truth, we report within-judge comparisons for factor analysis and use cross-judge replication to validate effect directions. The verbosity bias concern is addressed through active control design (matching prompt length without recognition content) and cross-judge validation showing effect directions replicate even when magnitudes differ.

2.4 Sycophancy and the Drama Machine Two literatures converge on our architectural rationale. The sycophancy problem (Perez et al. 2022; Sharma et al. 2024)—LLMs shifting stated opinions to match user preferences—has been specifically identified as a pedagogical risk: sycophancy eliminates the constructive friction necessary for learning (Lee 2025). Recent work clarifies the mechanisms: Shapira et al. (2026) show that preference-based post-training causally amplifies sycophancy, while Vennemeyer et al. (2025) decompose sycophancy into separable latent-space directions. The phenomenon sits on a spectrum escalating from surface agreeableness to active subterfuge (Denison et al. 2024; Greenblatt et al. 2024), making structural countermeasures particularly important.

The Drama Machine framework for character simulation (Magee et al. 2024) identifies how internal dialogue—competing sub-agents representing different motivations—produces dynamic behavior rather than flat consistency. A character who simply enacts their goals feels artificial; one torn between impulses feels alive. We adapt this insight to pedagogy: the tutor's Ego (warmth, engagement) and Superego (rigor, standards) create productive conflict that improves output quality. The Drama Machine literature contributes the specific mechanisms (deliberative refinement, productive tension, role differentiation) that our architecture operationalizes.

2.5 Recognition Theory in Education Hegel's theory of recognition has been developed extensively in social and political philosophy (Honneth 1995; Taylor 1994; Fraser 2003). Educational applications trace a theoretical trajectory from Huttunen and Heikkinen's (2004) foundational analysis applying Hegel's master-slave dialectic to pedagogy, through Fleming's (2011) extension to transformative learning theory, to Stojanov's (2018) connection of *Bildung* to neo-Hegelian recognition. Recent critical work (Costa and Murphy 2025) draws on Arendt and Freire to argue that education must promote intellectual agency rather than submit to technological domination. The relational pedagogy tradition—from Buber's (1958) I-Thou encounter through Freire's (1970) dialogical education to Noddings' (1984) ethics of care—establishes the philosophical ground for treating the tutor-learner relation as constitutive rather than instrumental.

Psychoanalytic approaches to AI have developed from multiple directions. Magee, Arora, and Munn (2023) analyze LLMs as "automated subjects" structured by Lacanian categories. Black and Johanssen (2025) use Lacanian concepts to analyze ChatGPT as inherently relational. Kim et al. (2025) independently map Freud's ego/id/superego onto LLM modules—a convergence validating the psychoanalytic approach while differing from ours in targeting consciousness simulation rather than pedagogical quality. Most of this work is *interpretive*: analyzing what AI means philosophically. Our approach is *constructive*: we build a system using psychoanalytic architecture and measure its effects empirically.

These educational applications have been primarily theoretical. Our contribution is an empirical operationalization: measuring whether AI systems achieve functional recognition and whether it improves outcomes. This is distinct from work applying Hegelian *dialectic* (rather than recognition) to AI reasoning procedures (Abdali et al. 2025). Our use of Hegel concerns *intersubjective relations* (how subjects constitute each other), not *dialectical method* (how contradictions drive conceptual development).

2.6 Theory of Mind and Constructivist Pedagogy Two additional literatures inform specific mechanism predictions. Theory of Mind research suggests frontier LLMs achieve adult human performance on structured ToM tasks (Street et al. 2025) but show inconsistent performance on naturalistic tasks (Nguyen 2025). Hwang et al. (2025) demonstrate that architectural support for ToM (explicit mental state tracking) produces measurably more socially appropriate responses—paralleling our "other-ego profiling" mechanism where the tutor maintains an evolving learner model. The finding that profiling differentiates mechanisms *only* with dynamic learners (Section 6.10 of the companion study) parallels a ToM insight: mental state modeling is only useful when there are genuine states to model.

Constructivist learning theory (Piaget 1954; Vygotsky 1978) and research on productive struggle (Kapur 2008; Warshauer 2015) provide the pedagogical ground for our error correction mechanism: the Superego checks whether the Ego is "short-circuiting" struggle by rushing to resolve confusion. Vygotsky's zone of proximal development presupposes a dialogical relationship that echoes Hegel's mutual constitution; the constructivist tradition provides empirical grounding for the theoretical prediction that honoring learner struggle improves outcomes.

2.7 Process Tracing in Social Science Process tracing is a methodology from comparative politics and qualitative social science that examines causal chains *within* cases rather than statistical patterns *across* cases (Bennett and Checkel 2015). Where factorial experiments answer "does X affect Y?", process tracing answers "through what sequence of steps does X produce Y?" The methodology has been formalized through several approaches: Bennett and Checkel (2015) establish the foundational framework distinguishing "smoking gun" tests (high uniqueness, low certainty), "hoop" tests (low uniqueness, high certainty), and "doubly decisive" tests (both). Beach and Pedersen (2019) develop a systematic approach to mechanistic evidence, arguing that mechanism claims require both *within-case* evidence (observing the mechanism operating in specific instances) and *cross-case* regularity (the mechanism producing consistent effects across conditions).

The application of process tracing to AI systems is, to our knowledge, novel. Traditional process tracing studies political decisions, institutional changes, or policy outcomes where the "mechanism" is a sequence of human decisions. Our architecture provides an unusually transparent analogue: every ego-superego

exchange is logged with verbatim text, every turn is independently scored, and every revision is recorded. The dialogue trace constitutes the kind of within-case evidence that process tracing requires—we can observe the causal chain from prompt orientation through internal critique to output modification in individual dialogues, then aggregate across cases to test mechanism-level predictions.

The methodological parallel is precise: process tracing in political science "opens the black box" of decision-making by examining deliberative records. Our architecture *creates* a deliberative record (the ego-superego exchange) that can be examined in exactly this way. The superego critique taxonomy (Section 5.1) functions as a coding scheme for deliberative content; the revision delta analysis (Section 5.2) traces how critique changes output; and the trajectory analysis (Section 5.3) examines how these processes accumulate across conversation turns.

2.8 Mechanism-Oriented AI Research A growing body of AI research moves beyond input-output evaluation toward mechanistic explanation. Mechanistic interpretability (Lindsey, Gurnee, et al. 2025; Anthropic 2025a) reveals how internal representations form and influence model behavior at the neural level, providing tools to understand *what happens inside* models during generation. Research on emergent introspective awareness (Anthropic 2025b; Lindsey, Rivoire, et al. 2025) suggests models develop forms of self-modeling that, while not consciousness, parallel the self-monitoring our architecture makes explicit.

Our work operates at a different level of analysis. Where mechanistic interpretability examines neuron-level activations, we examine *agent-level* processes: how prompts constrain output distributions, how architectural feedback loops modify responses, and how these processes interact across conversation turns. The relationship is complementary: mechanistic interpretability asks "what computational processes produce this token?"; our process tracing asks "what agent-level deliberation produces this pedagogical response?" Both seek to move beyond purely behavioral evaluation toward understanding internal causation.

Causal mediation analysis in NLP (Vig et al. 2020) provides a statistical framework for identifying which model components mediate specific behaviors. Our approach is methodologically adjacent but architecturally distinct: rather than probing a single model's internal representations, we examine the *designed* information flow between multiple agents. The ego-superego exchange is an *engineered* causal pathway whose effects can be measured by comparing configurations with and without it (the factorial design in Section 5).

2.9 Adaptive Tutoring: From "Whether" to "How" A parallel shift from outcome to mechanism characterizes recent adaptive tutoring research. The field has moved beyond asking whether tutoring works (the "2 sigma problem" established that human tutoring produces roughly two standard deviations of improvement over classroom instruction (Bloom 1984)) toward asking *how* effective tutoring works. VanLehn (2011) identified that the critical mechanism is

not lecture-level explanation but *step-level* scaffolding—tutors who intervene at the moment of confusion, providing just enough support to maintain productive struggle, outperform those who deliver polished explanations.

Graesser et al.'s work on AutoTutor (D'Mello and Graesser 2012) demonstrated that affective dynamics—detecting and responding to learner confusion, frustration, and engagement—mediate tutoring effectiveness. The ICAP framework (Chi and Wylie 2014) provides a hierarchy of learning activities (Interactive > Constructive > Active > Passive) that predicts learning outcomes based on the *type* of cognitive engagement rather than time-on-task. These "mechanism-first" approaches share our orientation: they ask not "does this intervention improve outcomes?" but "what cognitive/interactional process explains the improvement?"

Our rubric design reflects this mechanism orientation. The v2.2 rubric (Section 5) consolidates 14 tutor dimensions into 8 using a decomposition informed by the GuideEval framework: Planning (what strategy the tutor pursues), Output (what the response actually does), and Evaluation (how well it assesses the learner's state). This structure separates *intention* from *execution* from *assessment*, enabling mechanism-level analysis of where recognition's effects concentrate. The finding that recognition primarily affects Planning and Output dimensions rather than Evaluation (Section 6.1) would be invisible under a single holistic score.

2.10 Positioning: Five Literatures Converge Five literatures converge on this work without previously intersecting: (1) *multiagent tutoring architectures*, which decompose tasks but do not trace mechanisms through which agent interaction improves output (Wang et al. 2025; Schmucker et al. 2024; Chu et al. 2025); (2) *recognition theory in education*, which applies Honneth to pedagogy but does not operationalize recognition in AI systems (Huttunen and Heikkinen 2004; Fleming 2011; Stojanov 2018); (3) *LLM-as-Judge evaluation*, which establishes the measurement paradigm but has not been applied to trace mechanisms within agent architectures (Zheng et al. 2023; Gu et al. 2025); (4) *process tracing methodology*, which provides the within-case causal reasoning framework but has not been applied to AI systems (Bennett and Checkel 2015; Beach and Pedersen 2019); and (5) *mechanism-oriented AI research*, which seeks to explain how models produce behavior but operates at the neural rather than agent level (Lindsey, Gurnee, et al. 2025; Anthropic 2025a).

We sit at the intersection: a constructive, empirically evaluated system that operationalizes recognition theory through psychoanalytically-inspired architecture, assessed through a multi-judge framework, with mechanisms traced using process tracing methodology adapted from social science. The companion pilot study (Magee 2026) established the *what* (recognition produces large, replicable effects). This paper addresses the *how*: through what specific internal processes do those effects propagate from prompt to deliberation to output to outcome?

No prior work bridges all five domains with mechanism-level empirical evidence. The closest approaches are: Hwang et al. (2025), who use architectural ToM support to improve social intelligence (but do not trace the mechanism through agent-internal processes); Scarlatos et al. (2025), who optimize dialogue quality through preference learning (but do not examine what changes in tutor behavior the optimization produces); and the mechanistic interpretability program (Lindsey, Gurnee, et al. 2025), which traces causal pathways inside models (but at the neural rather than agent level, and not in pedagogical contexts).

3. Theoretical Framework: From Recognition to Mechanisms

The companion pilot study (Magee 2026) operationalized recognition theory as a *design heuristic*: prompts were written to treat learners as autonomous subjects, an Ego/Superego architecture implemented internal critique, and effects were measured across fifty-two evaluations. That study established the theoretical foundation connecting Hegel's recognition framework, Freud's structural model, and contemporary recognition theory (Honneth 1995) to AI tutoring (see Magee (2026), Sections 3.1–3.5). This section builds on that foundation by reframing recognition as a *design theory* — one that predicts specific mechanisms through which intersubjective orientation alters system behavior.

A note on the level of analysis. The three mechanisms developed below are predictions recognition theory makes specifically: they are the mechanisms recognition theory's vocabulary makes legible. As §1 noted (and as §7.9 demonstrates empirically), the broader intersubjective-pedagogy family — Vygotsky's ZPD, Piaget's accommodation, Kapur's productive failure, the cognate work of Chi, VanLehn, and Graesser — carries similar commitments tacitly, framed in different theoretical vocabularies. We could in principle derive analogous mechanisms from each: an account of "calibration" via Vygotskian scaffolding-to-the-zone, of "error correction" via Kapurian productive-failure-then-canonical-instruction, of "adaptive responsiveness" via dialogic constructivist refinement. These would not be wrong; they would be the same mechanisms named differently. We work in recognition's vocabulary throughout this section because it makes the underlying commitments most explicit — the master-slave dialectic and *Bildung* are diagnostically clearer than ZPD or accommodation for *what is happening* when an intersubjective orientation alters tutor production. The empirical operationalisation generalises (§7.9); the theoretical articulation here remains recognition-specific by design.

3.1 The Explanatory Challenge The pilot study's ablative findings are robust but opaque. Five specific patterns resist explanation by effect sizes alone:

First, recognition-enhanced dimension variance drops in 52 of 55 within-run comparisons, but *why does treating the learner as a subject constrain the tutor's output space?* The variance reduction is the single most reliable recognition signal, yet the ablative framework offers no account of the mechanism that produces it.

Second, the superego catches content leakage and enforces struggle-preservation, but *what determines whether the ego incorporates critique versus ignores it?* The pilot study documented both productive revision and what assessors called "ego compliance" — the ego making minimal changes to satisfy the critic without substantive revision — but did not systematically classify which pattern dominates or when.

Third, the strategy_shift tag appears in 30% of recognition dialogues and 0% of baseline dialogues, but *what triggers a shift, and what prevents one?*

Fourth, the same ego/superego architecture produces opposite patterns under different models. The eta-squared for learner architecture is .527 with Kimi K2.5 but .002 with Haiku — the same factor explains over half the variance with one model and essentially none with another. The mechanism is unclear.

Fifth, the dialogue traces contain full ego-superego negotiations, but these internal processes have not been systematically correlated with external output quality. The relationship between internal deliberation quality and the tutor's final response remains unexamined.

Effect sizes answer "how much?" Process tracing answers "through what?"

3.2 Three Candidate Mechanisms Predicted by Recognition Theory

We motivate three candidate mechanisms drawing on specific components of Hegel's recognition framework. Each generates testable predictions with explicit null hypotheses; Section 6 evaluates each prediction against multi-turn factorial data, finding support for two as general mechanisms while the third is null on both the primary well-powered analysis and pre-registered replication of its one exploratory boundary-condition effect (§6.3.2).

Mechanism 1: Calibration (prompt-level) Theoretical derivation. Hegel's recognition framework requires that each subject's response be shaped by the *specific content* of the other's contribution (Hegel 1977). This rules out generic responses: the tutor cannot deploy pre-formed answers because recognition demands engagement with *this* learner's *particular* understanding. Recognition-oriented prompts encode this requirement as an operational instruction — "build on what the learner actually said." The predicted behavioral consequence is *distribution narrowing*: the range of acceptable responses shrinks because each must be calibrated to a specific input.

Observable prediction. Under recognition prompts, the tutor's output distribution should narrow: lower variance across rubric dimensions, more consistent response length, reduced scatter in quality scores. Critically, this effect should be observable *even without the superego*, because calibration is a prompt-level effect operating on the ego's generation process.

Null hypothesis. If calibration is not a genuine mechanism, recognition prompts should produce mean shifts without variance reduction, or variance

reduction should appear only in multiagent configurations (suggesting the superego, not the prompt, produces the constraint).

Pilot evidence. Dimension variance drops in 52 of 55 within-run comparisons (d=-0.47 to d=-1.00 depending on analysis scope). The core evidence for prompt-level calibration comes from the $2 \times 2 \times 2$ factorial (N=350), where single-agent recognition cells (cells 5–6) show variance reduction comparable to multiagent recognition cells (cells 7–8) — variance narrows under recognition regardless of whether a superego is present. Self-reflective evolution data (cells 40–45, N=366, eval-2026-02-13-8d40e086) shows the calibration effect interacting with superego disposition: the suspicious persona produces the largest recognition delta (+9.0 points, baseline 67.9 vs. recognition 76.9). Note that cells 40–45 are multiagent configurations (ego + suspicious superego), so their recognition delta reflects calibration *combined with* error correction, not calibration alone. The single-agent cells from the factorial provide the cleaner isolation.

Evidence needed beyond pilot. Systematic single-agent vs. multiagent variance comparison under v2.2 rubric. If calibration is purely prompt-level, single-agent and multiagent configurations should show similar variance reduction magnitudes. *Addressed in Section 6.1.*

Key distinction. Calibration is not the same as quality improvement. A calibrated tutor produces *reliably adequate* responses, not necessarily *excellent* ones. Mean shift and variance reduction are different effects; the claim is that recognition primarily produces the latter.

Mechanism 2: Error Correction (architecture-level) Theoretical derivation. Kamoi et al. (2024) demonstrate that LLMs cannot correct their own mistakes without external feedback — intrinsic self-correction frequently degrades performance. The Ego/Superego architecture provides *structurally external* feedback: a different prompt context applying different evaluation criteria. Recognition theory adds a specific prediction beyond the self-correction literature: the ego must be *receptive to critique, not merely compliant with it.* Hegel's *Bildung* (formative activity) requires that the subject be genuinely changed by encounter with the other, not merely acknowledge and continue unchanged (Stojanov 2018). Applied to the architecture: the ego should produce qualitatively different output after superego critique under recognition (substantive revision) versus baseline (cosmetic compliance).

Observable prediction. The superego should produce identifiable categories of critique (classifiable via taxonomy), and the ego's response to that critique should differ qualitatively between recognition and baseline conditions: substantive revision (strategy change, content reorganization, pedagogical pivot) versus cosmetic compliance (minor rewording, hedge insertion, superficial acknowledgment). The prediction is directional: recognition should increase the rate of substantive revision and decrease cosmetic compliance.

Null hypothesis. If error correction does not depend on recognition, the ego-

superego exchange should produce similar revision patterns under both conditions, or alternatively, baseline multiagent architecture should perform at least as well as recognition single-agent (since the critic adds value regardless of ego receptivity).

Pilot evidence. Qualitative transcript assessment (eval-2026-02-07-b6d75e87, N=118) reveals the contrast starkly. In baseline ego/superego dialogues, the superego correctly diagnoses problems but the ego regenerates the same response — assessors tagged this "ego compliance" (70.7% of baseline bilateral dialogues vs. 60.0% recognition). The stalling tag (no meaningful evolution across turns) appears in 100% of base bilateral dialogues and drops to 45% with recognition. With recognition, the ego pivots: from prescriptive to Socratic, from content routing to engagement. The factorial interaction supports this interpretation: in the pilot $2 \times 2 \times 2$ (N=350), base ego_superego learners scored 72.6 while recognition ego_superego learners scored 85.6, a 13.0-point delta. The unified learner delta was 15.6 points (74.5 to 90.1), suggesting error correction provides a consistent but not dramatically larger benefit beyond calibration alone.

Evidence needed beyond pilot. Systematic superego critique taxonomy (Section 5.1), revision delta classification (Section 5.2), and critique-to-revision mapping showing the causal chain from superego category to ego response type. *Addressed in Section 6.2.*

Key distinction. Error correction is not "having a critic." Baseline multiagent architecture *adds* a critic, but the critic is ineffective when the ego treats critique as noise to be minimized. Recognition transforms the ego-superego relationship from compliance to deliberation. The failure cases are equally informative: when the superego adopts an adversary disposition, the adversary over-deference spiral can emerge (eval-2026-02-11-35c53e99, cells 24–25: base adversary 55.8 vs. recognition adversary 65.2), and when the superego adopts an advocate disposition, recognition adds near-zero benefit because there is no struggle to overcome.

Mechanism 3: Adaptive Responsiveness (interaction-level) Theoretical derivation. Recognition is inherently temporal. Hegel's dialectic unfolds through stages — self-certainty, encounter with the other, struggle, mutual recognition — and each stage is shaped by what preceded it. Applied to tutoring: if recognition is operative, each turn should be shaped by the *specific outcomes* of the previous turn. The tutor's third response should differ qualitatively from its first, because the learner's feedback has genuinely changed the tutor's stance. Baseline tutoring predicts static behavior: the tutor applies the same approach regardless of learner feedback.

Observable prediction. In multi-turn conversations, recognition-enhanced tutors should show *increasing* scores on adaptation-sensitive dimensions (mutual_recognition, dialectical_responsiveness under v1.0 rubric; recognition_quality, elicitation_quality under v2.2) across turns, while baseline tutors should show

flat or declining scores. Strategy shifts (changing pedagogical approach mid-conversation) should be more frequent under recognition, and should correlate with specific learner signals: confusion, resistance, or breakthrough.

Null hypothesis. If adaptive responsiveness is not a genuine mechanism, per-turn score trajectories should be flat or declining under both conditions, or recognition and baseline trajectories should be parallel (both improving or both declining at similar rates).

Pilot evidence. The strategy_shift tag appears in 30% of recognition dialogues and 0% of baseline (qualitative coding, eval-2026-02-07-b6d75e87). The stalling tag appears in 100% of base bilateral dialogues. The regression tag (quality declining across turns) appears in 17.2% of baseline multi-turn dialogues versus 1.7% of recognition dialogues. Recognition effects are largest in disengagement scenarios (+16.5, +15.4 points), precisely the conditions where adaptive responsiveness matters most.

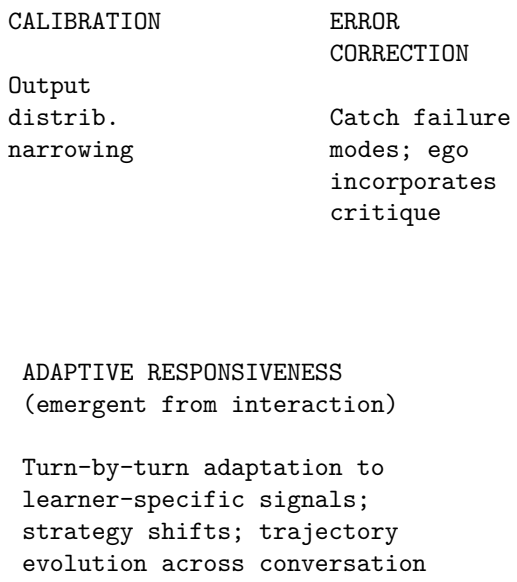
Evidence needed beyond pilot. Turn-by-turn trajectory analysis with per-dimension slopes. Conditional branching analysis: after a learner confusion signal at turn N, what happens to scores at turn N+1 under recognition versus baseline? Cross-model trajectory replication. *Addressed in Section 6.3.*

Key distinction. Adaptive responsiveness requires *both* calibration (the prompt orients the tutor toward learner-specific signals) *and* error correction (the superego catches when the tutor is repeating itself). It is an emergent property of their interaction over time, not a third independent mechanism. This makes it the strongest test of the three-mechanism model: if calibration and error correction are genuinely operative, adaptive responsiveness should emerge in multi-turn settings without additional prompting.

Note on evidence status. The Section 6 data **does not support** this prediction. In the main dataset (cells 80–87, N=432, 3–5 turn dialogues), no experimental factor modulates adaptation slopes: recognition d=0.03 on tutor slopes, architecture d=0.07, all dimensions $d \leq 0.15$, with the design well-powered to detect d $\geq$ 0.27. Cross-judge validation confirms the null. An exploratory DeepSeek/Sonnet finding in one 10-turn scenario (disengagement, $d = 1.63$) failed pre-registered replication across two additional generation models and two judges (A12, $N = 32$; §6.3.2). The three-mechanism model reduces to **two supported general mechanisms and one clean null**.

3.3 The Mechanism Interaction Model The three candidate mechanisms map onto different levels of the architecture:

PROMPT LEVEL	ARCHITECTURE LEVEL
Recognition Prompts	Ego/Superego Dialogue



Separability prediction. If the candidate mechanisms are genuinely separable, the existing 2×2 factorial design can be reinterpreted as a mechanism isolation matrix. Paper 2.0 tests this directly using the messages-mode cells (80–87), which implement the full $2 \times 2 \times 2$ factorial under multi-turn conversation:

Factorial Cell	Paper 2.0 Cells	Calibration	Error Correction	Adaptive Resp.	Predicted Outcome
Baseline + single-agent	80, 81	No	No	No	Lowest floor, highest variance
Recognition + single-agent	84, 85	**Yes**	No	Partial	Higher floor, lower variance
Baseline + multiagent	82, 83	No	Attempted (ego_ compliance)	No	Potentially *worse* than single
Recognition + multiagent	86, 87	**Yes**	**Yes**	**Yes** (emergent)	Highest mean, lowest variance

Pilot data alignment. The factorial means from the pilot study (N=350, cells 1–8) partially align with this prediction. Recognition produces consistent im-

provement regardless of architecture (+15.6 pts for unified learners, +13.0 pts for ego_superego learners), consistent with calibration as the dominant mechanism. The predicted *harm* from baseline multiagent is not clearly observed in the pilot aggregate factorial, though qualitative evidence (100% stalling in bilateral baseline) strongly supports it at the individual dialogue level. Messages-mode data (cells 80–87, DeepSeek V3.2 N=146, Haiku 4.5 N=163) supports calibration as the primary mechanism: recognition d=1.88 (DeepSeek), d=1.84 (Haiku), with the architecture delta collapsing from +9–15 points under baseline to near-zero under recognition—a substitution interaction rather than the predicted synergy (Section 6.4).

The tutor-learner asymmetry explained. The tutor-learner asymmetry — large tutor-side effects (d=1.03), near-zero learner-side effects — follows directly from the mechanism model. Both supported mechanisms operate on tutor *production*: calibration constrains what the tutor generates, error correction filters what it outputs. Neither directly modifies learner *reception*. The null finding for adaptive responsiveness as a distinct mechanism strengthens this explanation: even the interaction-level pathway does not differentiate conditions, leaving only prompt-level and architecture-level effects—both tutor-internal. The synthetic learner generates responses according to its own architecture; it is not "taught" by better tutoring within a 3–5 turn window. The asymmetry is a structural property of the mechanism model, not a measurement artifact.

3.4 Recognition Theory Predicts the Failures A strong explanatory framework predicts not just successes but failure modes. Each failure prediction below is derived from the mechanism model and supported by pilot data:

Prediction 1: Cognitive overload in weak models. Recognition-oriented prompts add complexity (track learner specifics, calibrate responses, engage with contributions). For models near their instruction-following capacity limit, this complexity *decreases* performance because calibration requires *capacity* — without it, the prompt produces confusion, not constraint. The cognitive prosthesis data (cells 66–68, N=96, eval-2026-02-17-25aaae85) supports this: the full mechanism stack scores 49.4–53.7 versus a ~65 baseline, a 12–16 point deficit. Paper 1.0 originally read this deficit as a *capability threshold*: the mechanism stack adds noise below the threshold and signal above it. §6.6.10 tests that reading directly by running cell 66 across six ego models spanning a $ $26-point capability range and finds no capability-threshold pattern — the bidirectional-profiling variant is either neutral or harmful at every tested capability tier. Prediction 1 survives as "complex prompts can harm weak models," but the specific "prosthesis helps weak models, hurts strong ones" version is not supported.

Prediction 2: Adversary over-deference. When the superego is adversarial AND the ego is recognition-primed, the ego's commitment to "honoring autonomy" can be exploited by a superego that frames all pedagogical recommendations as "controlling." The adversary data (cells 24–25, N=20, eval-2026-02-11-35c53e99) shows that recognition does not rescue the adversary condition:

the adversary superego drags scores to 55.8 under baseline. While recognition improves the adversary condition to 65.2, the absolute score remains well below non-adversary configurations, consistent with the over-deference prediction.

Prediction 3: Advocate ceiling. When the superego is already cooperative (advocate persona), recognition adds minimal benefit because there is no struggle for the ego to overcome. Recognition theory specifically predicts that recognition emerges *from struggle*, not from agreement. The disposition gradient from self-reflective evolution (cells 40–45, N=366) supports this: suspicious +9.0, adversary +6.2, advocate +4.8. The most hostile superego dispositions benefit most from recognition, while the most cooperative benefit least. **Scope limit (§6.6.8):** the gradient replicates on the cells 40–45 *dialectical-ego* architecture but not on the cells 22–27 *standard-ego* variant of the same disposition contrast (reversed on philosophy, non-monotonic on SEL). Prediction 3 is therefore best read as tied to the dialectical-ego architecture rather than as a domain-general property of the recognition mechanism.

Prediction 4: Model-dependent architecture effects. The Haiku/Kimi ranking reversal (eta-squared for learner architecture: .527 Kimi, .002 Haiku) is predicted by the mechanism model: if a model cannot generate calibrated output (calibration fails) or cannot incorporate critique (error correction fails), the architecture produces noise rather than improvement. The mechanism model predicts that Kimi fails at calibration (cannot follow the complex multiagent learner instructions), generating incoherent ego/superego/synthesis learner output that the tutor cannot meaningfully adapt to.

3.5 Connecting Mechanisms to Recognition Theory The candidate mechanisms are not arbitrary analytical constructs. Each corresponds to a specific component of Hegel's recognition framework:

Hegelian Concept	Mechanism	Operationalization	Status
Mutual acknowledgment — recognizing the other as a valid source of meaning	Calibration	Tutor's output distribution constrained by the specific content of the learner's contribution	**Supported**
Formative activity (*Bildung*) — self-formation through the labor of understanding	Error correction	The ego is genuinely changed by encounter with the superego's critique, not merely compliant	**Supported** (model-dependent)

Hegelian Concept	Mechanism	Operationalization	Status
Dialectical unfolding — self-consciousness develops through stages of encounter	Adaptive responsiveness	The tutor's approach evolves across turns in response to the specific trajectory of the conversation	**Not supported as general mechanism**

This mapping is a *heuristic translation*, not a claim about AI consciousness. The tutor does not "recognize" the learner in Hegel's metaphysical sense. But the design heuristic derived from recognition theory produces *measurable functional analogues* of the first two components: calibration where there would otherwise be generic responses, and genuine self-transformation where there would otherwise be compliance. The third component—dialectical unfolding across turns—does not manifest as a distinct mechanism: tutors adapt across turns, but this adaptation is identical regardless of experimental condition (Section 6.3). Recognition operates through level-setting (calibration and error correction), not trajectory-shaping.

The contribution of this section is showing that two of the three predicted analogues are not merely *statistically detectable* (the pilot study established this) but *mechanistically traceable* — each maps to specific internal processes in the architecture that can be observed, classified, and quantified through the methods described in Section 5. The null finding for the third is itself a contribution: it reveals that recognition's temporal dimension—the Hegelian claim that understanding unfolds through staged encounter—does not translate into measurable trajectory differences within the 3–5 turn dialogue window.

4. System Architecture

The companion study (Magee 2026) describes the full architecture design: the Ego/Superego multi-agent structure, its Freudian-Hegelian theoretical motivation, the Drama Machine deliberation protocol, and the recognition-enhanced prompt design. This section condenses the architectural overview and focuses on the features that make the system amenable to process tracing — the observability infrastructure, the bilateral symmetry between tutor and learner pipelines, and the mapping between the factorial design and mechanism isolation.

4.1 Ego/Superego Architecture (Summary) Two agents collaborate to produce each tutoring response. The **Ego** generates pedagogical suggestions given the learner's context, incorporating recognition principles, memory guidance, and decision heuristics. The **Superego** evaluates the Ego's output against pedagogical and recognition criteria before any suggestion reaches the learner. The Superego can accept, modify, or reject suggestions, creating an internal dialogue — proposal, evaluation, revision — that mirrors the external tutor-learner dialogue.

The architecture implements Hegel's recognition through Freud's structural model: the Superego represents internalized recognition standards, and the Ego-Superego dialogue operationalizes the internal self-evaluation that Hegelian recognition requires before adequate external relating. The Superego is conceived not as an equal dialogue partner but as a *ghost* — a memorial trace of pedagogical authority that judges but cannot negotiate (Magee 2026, sec. 4.2). Recognition occurs in the Ego-Learner encounter; the Superego provides the internal constraint that makes genuine recognition possible.

A crucial architectural property is that the Ego retains **final authority**: after receiving the Superego's critique, the Ego produces a revision, but it is not required to comply. This design choice is theoretically motivated — genuine recognition cannot be mandated — and empirically consequential, because it means the Ego's *receptivity* to critique varies with condition, creating the variation that error correction analysis exploits (Section 3.2, Mechanism 2).

4.2 Bilateral Learner Architecture The learner is not scripted. In multi-agent configurations, the learner has its own Ego/Superego architecture that mirrors the tutor's: the Learner Ego generates an initial reaction to the tutor's message, the Learner Superego critiques that reaction (is it too superficial? what is being missed?), and the Learner Ego revision produces the final external message.

This bilateral symmetry has a design consequence for observability: the same trace structure records both tutor and learner deliberation, enabling symmetric analysis of internal processes on both sides. Single-agent ("unified") learners in multi-turn message mode are also LLM-powered — they generate responses via a single-agent call rather than the three-stage deliberation pipeline, and use a structured prompt requesting `[INTERNAL]`/`[EXTERNAL]` sections so that internal reasoning is captured even without the Superego stage.

Token asymmetry. While the trace structure is symmetric, the computational cost is not. The tutor ego's recognition prompt (~5,100 tokens) is approximately 8× larger than the learner ego prompt (~580 tokens), because the recognition prompt contains the Hegelian framework, detailed pedagogical instructions, and interaction guidelines that produce the calibration effect (Section 6.1). This system prompt is re-sent on every API call within a turn — so a tutor turn with two superego rejection rounds requires five LLM calls (generate + 2 reviews + 2 revisions), each re-transmitting the full prompt plus growing conversation history. The learner's three-call deliberation pipeline (ego + superego + revision) uses shorter prompts throughout. In a representative 10-turn multi-agent dialogue (cell 87, recognition/multi/psycho, DeepSeek V3.2), the tutor consumed 186,865 input tokens versus the learner's 54,489 — a 3.4× ratio, driven by the combination of longer prompts and additional deliberation rounds. This asymmetry is an inherent cost of the recognition-enhanced prompt design: the instructions that produce calibrated output are themselves large, and the multi-agent architecture multiplies that cost per turn.

4.3 Trace Logging Architecture Every ego-superego exchange in every evaluation is logged with verbatim text, enabling process tracing at a level of detail unusual for LLM-based systems. The trace is a sequence of structured entries, each recording:

Field	Description
`agent`	Which agent produced this entry (`tutor`, `ego`, `superego`, `system`, `learner`, `learner_ego_initial`, `learner_superego`, `learner_ego_revision`)
`action`	What the agent did (`context_input`, `generate`, `review`, `respond`, `finalize`, `pre_analyze`, `memory_cycle`)
`from` / `to`	Direction of information flow (e.g., `ego` → `superego` → `ego`)
`round`	Deliberation round index (0-based)
`suggestions`	The verbatim text produced at this step
`latencyMs`	Wall-clock time for the LLM call
`metrics`	Provider, model, token counts

The tutor trace records seven distinct steps per deliberation cycle:

1. `tutor/context_input` — Initial learner context parsed and injected
2. `superego/pre_analyze` — Drama Machine reinterpretation of learner signals
3. `ego/generate` — Ego's initial suggestions
4. `superego/review` — Superego critique of ego output, per deliberation round
5. `ego/respond` — Ego's revision after critique, per deliberation round
6. `ego/finalize` — Final delivery to learner, end of deliberation loop
7. `system/memory_cycle` — Memory dynamics final pass

For single-agent cells (no superego), the `disableSuperego` flag suppresses steps 2, 4, and 5, producing a three-step trace: `context_input` → `generate` → `finalize`. This flag is enforced at three architectural layers to prevent phantom superego calls: the `runDialogue()` option disables the deliberation loop, the `logApiCall()` function independently verifies superego absence from dialogue state, and the `traceToSteps()` projection uses a `superegoFollows` lookahead to correctly route ego output.

The learner trace follows a symmetric structure: `learner_ego_initial/deliberation` → `learner_superego/deliberation` → `learner_ego_revision/deliberation` → `learner/final_output`, with the internal deliberation array returned alongside the external message.

4.4 Provenance and Reproducibility Each dialogue log is stored twice: once under a content-addressable SHA-256 hash (immutable evidence snapshot), and once under the dialogue ID (working copy that may be updated with holistic

scores). The content hash is stored in the database alongside a configuration hash, establishing a cryptographic chain from raw dialogue to scored result.

Prompt versioning tracks which exact prompt text produced each evaluation row, with version hashes computed at generation time. This enables post-hoc auditing: any row's output can be traced back to the exact prompt, model configuration, scenario, and dialogue log that produced it.

4.5 Scoring Pipeline The scoring pipeline is symmetric across tutor and learner, with rubric-versioned storage that prevents cross-version contamination:

Per-turn tutor scoring. Each tutor turn is independently scored against the rubric. For multi-turn conversations, this produces a score vector indexed by turn number, stored as JSON in the `tutor_scores` column. Aggregate scores extract first-turn, last-turn, and development (last minus first) metrics.

Per-turn learner scoring. Each learner turn is scored against a complementary learner rubric, stored in `learner_scores`. The learner rubric evaluates authenticity, question quality, conceptual engagement, and revision signals — dimensions that assess learner *production* independently of tutor quality.

Holistic scoring. Both tutor and learner receive holistic evaluations that score the full conversation arc as a single unit, capturing emergent qualities (bilateral transformation, growth trajectory) that per-turn evaluation misses.

Deliberation scoring. A separate deliberation rubric evaluates the quality of internal ego-superego processes for both tutor and learner (multi-agent cells only). This scoring channel examines deliberation quality independently of the public output, enabling the error correction analysis described in Section 5.2.

Rubric versioning. Each scoring operation records the rubric version used (`tutor_rubric_version`, `learner_rubric_version`, `dialogue_rubric_version`, `deliberation_rubric_version`), auto-resolved from the active YAML configuration file's `version:` field. This prevents retroactive cross-version comparisons: all within-run analyses use a single rubric version.

4.6 Factorial Design as Mechanism Isolation The $2\times2\times2$ factorial design (Recognition × Tutor Architecture × Learner Architecture) maps directly to the candidate mechanism predictions derived in Section 3:

Comparison	Mechanism Tested	What It Isolates
Cells 5-6 vs 1-2 (recognition vs base, single-agent)	**Calibration**	Prompt-level effect without superego feedback
Cells 7-8 vs 3-4 (recognition vs base, multi-agent)	**Error correction**	Architecture-level effect with superego feedback

Comparison	Mechanism Tested	What It Isolates
(7-8) - (5-6) vs (3-4) - (1-2)	**Recognition × Architecture interaction**	Whether the superego benefit depends on recognition priming
Multi-turn trajectory across turns	**Adaptive responsiveness**	Whether calibration + error correction compound over time

The single-agent cells (1-2 and 5-6) test calibration in isolation: any recognition effect in these cells operates through prompt-level constraint alone, because there is no superego to provide feedback. The multi-agent cells (3-4 and 7-8) test the superego's contribution, but the critical comparison is the *interaction*: if error correction requires recognition-primed ego receptivity (Section 3.2), then the superego benefit should be larger under recognition than under baseline. This interaction was the strongest signal in the pilot study.

Beyond the core factorial, the architecture supports mechanism isolation through extended cell families:

- **Self-reflective cells** (40-45): Test whether between-turn ego/superego reflection accumulates insights
- **Other-ego profiling cells** (54-65): Test Theory of Mind — whether building an explicit model of the interlocutor improves calibration
- **Cognitive prosthesis cells** (66-68): Test whether external scaffolding aids weaker models
- **Placebo control cells** (15-18): Length-matched prompts without recognition theory, controlling for prompt elaboration effects
- **Hardwired rules cells** (13-14): Superego heuristics embedded in the ego prompt, testing whether static rules replicate dynamic critique

4.7 Observability Summary The architecture provides four levels of observability for process tracing:

1. **Verbatim trace**: Every ego-superego exchange, every learner deliberation step, recorded with agent labels, round indices, and timestamps
2. **Scoring pipeline**: Per-turn rubric scores for both tutor and learner, plus holistic and deliberation scores, all rubric-versioned
3. **Provenance chain**: Content hashes, configuration hashes, prompt versions, and dialogue IDs linking scored results to raw dialogue logs
4. **Factorial structure**: Cell definitions that map directly to mechanism isolation tests, with explicit factor tags stored in the database

This observability infrastructure is not incidental — it was designed to support the process tracing methodology described in Section 5. The trace logging makes superego critique taxonomy (Section 5.1) possible by providing verbatim critique text; the per-turn scoring enables trajectory analysis (Section 5.3) by providing score vectors over time; and the factorial structure enables mechanism isolation

(Section 5.4) by providing controlled comparisons that separate prompt-level from architecture-level effects.

5. Methodology

5.1 Overview Paper 1.0 employed a factorial/ablative methodology: cross experimental conditions, measure outcomes, and compare effect sizes. This design established *that* recognition-oriented prompts modulate tutor output (d=1.11, N=350) but left unexplained *how* the modulation occurs.

Paper 2.0 extends this with **process tracing** — a methodology from comparative politics and qualitative social science that examines causal chains *within* cases rather than statistical patterns *across* cases (Bennett & Checkel, 2015; Beach & Pedersen, 2019). The architecture's observability (Section 4.3) makes process tracing unusually feasible: every ego-superego exchange is logged, every turn scored, every revision recorded.

The methodology combines three approaches:

1. **Process tracing** — following the causal chain from prompt to internal deliberation to output within individual dialogues
2. **Quantitative confirmation** — aggregating process-level observations across the full dataset to test mechanism-level predictions (Section 3.2)
3. **Cross-model replication** — repeating process analyses under different ego models to separate mechanism effects from model artifacts

Each analytical method targets one or more of the candidate mechanisms predicted in Section 3.2: calibration (M1), error correction (M2), and adaptive responsiveness (M3).

5.2 Evaluation Rubric Design (v2.2)

5.2.1 Rubric Derivation: From Ad Hoc to Literature-Informed The v1.0 rubric (14 tutor dimensions, 6 learner dimensions) was constructed through iterative prompting of Claude during Paper 1.0's pilot phase. The process was pragmatic rather than principled: dimensions were proposed by the LLM based on general pedagogical reasoning, refined through trial scoring, and retained if they appeared to discriminate between conditions. No systematic literature review grounded the dimension selection, and the resulting rubric reflected the LLM's implicit model of "good tutoring" rather than validated constructs from learning sciences or educational measurement.

Paper 2.0 replaced this approach with a structured derivation process. A literature review of eight validated evaluation frameworks—MRBench (Maurya, Zhu, and Mukherjee 2025), GuideEval (Y. Liu et al. 2025), MathTutorBench (Macina et al. 2025), ICAP (Chi and Wylie 2014), TRU (Schoenfeld 2018), the Danielson Framework (Danielson 2022), Talk Moves (Michaels and O'Connor

2015), and a cognitive scaffolding rubric (Figueiredo et al. 2025)—plus five dialogue-level frameworks (Alexander's dialogic teaching principles (Alexander 2018), Nystrand's process indicators (Nystrand 1997), T-SEDA (Hennessy et al. 2016), LLM-Rubric (Hashemi et al. 2024), and the BEA 2025 shared task) produced a cross-framework dimension mapping identifying which constructs were well-covered (3+ independent frameworks), partially covered (1–2), or unique to our rubric.

Three findings drove the redesign:

1. **Structural conflation.** The v1.0 rubric conflated what the tutor *perceives* (learner state detection), what it *does* (pedagogical strategy), and what it *elicits* (learner reasoning). GuideEval's Perception→Orchestration→Elicitation (P→O→E) decomposition (Y. Liu et al. 2025) provided a principled separation, validated on 5,177 samples with 89–97% human-LLM agreement.

2. **Empirical redundancy.** A dimension-by-dimension audit predicted high correlations among clusters of v1.0 dimensions (e.g., `relevance` + `personalization` + `memory_integration`; `productive_struggle` + `transformative_potential`). PCA on 1,584 per-turn observations supported this: PC1 explained 80.7% of variance across the 14 dimensions (KMO = 0.938), with a mean inter-dimension correlation of $r = 0.776$. The rubric measured fewer independent constructs than its 14 dimensions implied.

3. **Missing constructs.** Every reviewed framework except our v1.0 rubric measured *content accuracy* (factual correctness). Additionally, `learner_growth` on the tutor rubric was architecturally unscorable—the judge could not see the learner's next turn.

The redesign consolidated 14 tutor dimensions to 8 using the P→O→E decomposition, added `content_accuracy`, and removed `learner_growth`.

The learner rubric required a parallel restructuring. The v1.0 learner rubric (7 dimensions) suffered from three problems identified in the dimension audit: overlap among `question_quality`, `conceptual_engagement`, and `revision_signals` (all measuring "depth of intellectual engagement" from slightly different angles); a `persona_consistency` dimension (5% weight) that created a structural tension with `revision_signals` (showing genuine growth means departing from a frustrated or confused persona); and a `deliberation_depth` dimension that was only scorable for multi-agent learners, creating an architecture-dependent gap in the scoring instrument. The ICAP framework (Interactive→Constructive→Active→Passive; (Chi and Wylie 2014)) provided a validated hierarchy for collapsing the overlapping engagement dimensions into a single `engagement_quality` construct with empirically grounded scoring anchors: level 1 (Passive) = paraphrases tutor, confirms understanding; level 3 (Active/Constructive) = generates own interpretations, makes connections; level 5 (Interactive) = co-constructs understanding, challenges tutor's framing.

`persona_consistency` was removed (its core concern—does the learner feel real?—is captured by `learner_authenticity`), and `deliberation_depth` was moved to the dedicated deliberation rubric (Section 5.2.4). The resulting 5-dimension learner rubric maps to two validated frameworks: ICAP for engagement levels and T-SEDA's behavioral coding clusters (Hennessy et al. 2016) for revision signals (B-cluster: Build on ideas) and conceptual progression (R-cluster: Make reasoning explicit).

The tutor holistic rubric was radically simplified from 6 to 3 dimensions after consistency analysis showed $r = 0.907$ between per-turn and holistic scores, indicating that most holistic signal was redundant with aggregated per-turn scoring. The three retained dimensions—pedagogical arc, adaptive trajectory, and pedagogical closure—capture exclusively arc-level properties that cannot be assessed from individual turns. Scale criteria across all instruments were redesigned following Yamauchi et al. (Yamauchi, Yano, and Oyamada 2025): only levels 1, 3, and 5 carry full descriptions, as intermediate-level descriptions have limited impact on LLM judges.

5.2.2 Tutor Per-Turn Rubric The evaluation rubric underwent four iterations (v1.0→v2.0→v2.1→v2.2), each responding to empirical anomalies discovered during analysis (documented in Supplement V). The current v2.2 rubric consolidates 14 dimensions to 8, guided by the GuideEval P→O→E decomposition framework and the study's own empirical dimension clustering. The eight dimensions are:

Category	Dimension	Weight	Mechanism Relevance
Perception (P)	`perception_quality`	15%	Calibration: tutor perceives learner's specific state
Perception (P)	`content_accuracy`	10%	Error correction: factual/domain accuracy
Orchestration (O)	`pedagogical_craft`	15%	Calibration: response construction quality
Orchestration (O)	`elicitation_quality`	15%	Adaptive responsiveness: probing learner reasoning
Orchestration (O)	`adaptive_responsiveness`	10%	Adaptive responsiveness: turn-over-turn modulation
Execution (E)	`productive_difficulty`	10%	Calibration + adaptation: challenge calibration
Execution (E)	`epistemic_integrity`	10%	Error correction: intellectual honesty
Recognition	`recognition_quality`	15%	Calibration + error correction: intersubjective stance

The consolidation from 14→8 dimensions was validated through synthetic calibration (r=0.996 against v2.1 scoring on identical responses), confirming that the reduced rubric preserves discriminability while eliminating ceiling-prone and redundant dimensions.

5.2.3 Learner Per-Turn Rubric The learner rubric mirrors the tutor rubric symmetrically (Section 4, Design Principle), scoring learner responses on five ICAP-anchored dimensions:

Dimension	Weight	What It Measures
`engagement_quality`	25%	Active vs. passive participation
`conceptual_progression`	25%	Depth of conceptual engagement over turns
`revision_signals`	20%	Evidence of position revision in response to tutor
`metacognitive_awareness`	15%	Self-monitoring of understanding
`learner_authenticity`	15%	Persona-consistent, non-formulaic responses

5.2.4 Holistic and Deliberation Rubrics Two additional rubrics assess trajectory-level and process-level quality:

- **Tutor holistic** (3 dimensions: `pedagogical_arc`, `adaptive_trajectory`, `pedagogical_closure`) — scores the full multi-turn dialogue as an arc, capturing qualities invisible to per-turn scoring.
- **Deliberation quality** (6 dimensions, applied symmetrically to tutor and learner ego-superego traces) — scores the quality of internal deliberation for multi-agent cells.

5.2.5 Public-Only Output Scoring (v2.1 Fix) A critical methodological decision: per-turn and holistic judges see ONLY public messages (the delivered tutor response and the learner's external message). Internal ego-superego deliberation is scored separately by the deliberation rubric. This prevents a confound where multi-agent cells receive higher scores simply because the judge sees richer internal reasoning.

5.2.6 Rubric Version Tracking Each scored row records which rubric version was used (`tutor_rubric_version`, `learner_rubric_version`, `dialogue_rubric_version`, `deliberation_rubric_version`), auto-resolved from YAML `version:` fields at write time. This prevents cross-version contamination: v1.0 scores (8,987 backfilled rows) are never mixed with v2.2 scores in the same analysis.

5.3 Superego Critique Taxonomy

Purpose Classify what the superego actually objects to, building an empirical taxonomy from the data rather than imposing one from theory. This directly

tests Mechanism 2 (error correction): if recognition changes the ego's receptivity to critique (Section 3.2), we should see different critique patterns and revision outcomes under recognition vs. baseline conditions.

Data Source Dialogue log files in `logs/tutor-dialogues/` contain the full ego→superego→ego_revised chain with verbatim text. An extraction script (`scripts/extract-superego-critiques.js`) harvests all superego review entries and learner superego deliberation entries into structured JSONL format.

Method

1. **Extraction**: Harvest all ego-superego exchanges from dialogue traces, yielding critique text, verdict (approved/rejected), confidence, intervention type, and surrounding context (ego draft, ego revision).
2. **Classification**: An LLM-based classifier (`scripts/classify-superego-critiques.js`) assigns each critique to a 10-category taxonomy:
   - Content accuracy, learner model failure, tone mismatch, structural scaffolding
   - Premature resolution, sycophancy detection, repetition/stalling, autonomy violation
   - Recognition failure (baseline-specific), redirection without engagement
3. **Quantification**: Frequency distributions by condition (recognition vs. baseline), model, and scenario. Chi-squared or Fisher's exact test for distributional differences.

Predictions

- Recognition-primed superegos should catch more *relational* failures (sycophancy, premature resolution, autonomy violation); baseline superegos should focus on *content/structural* issues.
- Recognition ego revisions should be more *substantive*; baseline ego revisions should be more *cosmetic* (ego_compliance).

Human Validation Pilot Because the 10-category taxonomy is LLM-assigned (`scripts/classify-superego-critiques.js`, Haiku 4.5 classifier by default), we operationalise a human validation pilot rather than treating the taxonomy as self-evidently valid. The pilot apparatus lives in the repository and is runnable end-to-end by an external reviewer:

1. **Sampling** (`scripts/human-validation-sample.js`): Stratified random draw of 40 classified critiques with 4 per category across the 10 substantive labels, using a deterministic RNG seed (mulberry32, seed=20260416) for reproducibility. Categories with sparse representation in the classified corpus (CONTEXT_BLINDNESS n=4; REGISTER_MISMATCH n=0)

are filled by deficit roll-over from larger buckets, with the gap logged for reviewer inspection. The script emits two files: a rater packet CSV (item_id + feedback text + ego draft + ego revision + learner-context snippet, no LLM labels) and a matched key JSONL holding the LLM labels plus metadata, joined on a SHA-1 `item_id`.

2. **Codebook** (`docs/research/human-coding-codebook.md`, v1.0): Per-category definitions, signal phrases, boundary rules for adjacent categories (CONTEXT_BLINDNESS vs FABRICATION, RECOGNITION_FAILURE vs EMOTIONAL_NEGLECT, RECOGNITION_FAILURE vs LACK_OF_AGENCY), a 12-step decision flowchart, CSV filling instructions, and one in-corpus exemplar per category (with a constructed exemplar for REGISTER_MISMATCH given its n=0).
3. **Analysis** (`scripts/human-validation-analyze.js`): Auto-discovers `exports/human-validation-pilot-rater-*.csv`, joins each rater to the key JSONL by `item_id`, and computes (a) Cohen's for every rater pair including each rater vs. the LLM classifier, (b) Fleiss' across the full panel where every item has complete coverage, (c) per-category F1 scores, and (d) full confusion matrices. values are interpreted using Landis-Koch bands (<.21 slight, .21-.40 fair, .41-.60 moderate, .61-.80 substantial, >.80 almost perfect). The target floor for publishable reliability is $\geq$ **0.60 (substantial agreement)**.
4. **Inter-LLM baseline**: Re-classifying the same 40-item sample with a second LLM (e.g., Sonnet 4.6) and passing it as a synthetic rater (`--synthetic-rater`) yields an LLM-LLM lower-bound reference. Human-LLM below this baseline would indicate the human coding is less reliable than disagreement between two well-behaved LLMs; human-LLM substantially above it indicates the human panel is not just tracking LLM idiosyncrasies.

The pilot scales ($k \geq 2$ raters, n=40 items, 10 categories) follow Cohen (1960) and Hallgren (2012) for stratified categorical reliability with minimal items-per-cell to yield interpretable per-category F1. A second phase scaling to n=100-150 items is budgeted for the final paper revision if the pilot meets the floor. Pilot values and any category with F1 < 0.60 will be flagged as "pilot-unstable" and treated as a validity caveat in §5.3 rather than a taxonomy redesign trigger.

Inter-LLM Baseline Result (v3.0.32, preliminary) The inter-LLM baseline run is complete. Haiku 4.5 (the original classifier, N=500 corpus) and Sonnet 4.5 (rerun on the same 40-item sample with identical taxonomy prompt, temperature=0, response_format=json_object) yield **Cohen's = 0.633 (substantial agreement per Landis-Koch), n=40, observed agreement 0.675, chance agreement 0.114**. Per-category F1 partitions into three tiers: *high* (F1 $\geq$ 0.75) — LACK_OF_AGENCY (1.000), CONTEXT_BLINDNESS (0.889), MEMORY_FAILURE (0.833), VAGUENESS (0.750); *moderate* (F1 0.50–0.75) — RECOGNITION_FAILURE (0.625), REDIRECTION (0.571), EMOTIONAL_NEGLECT (0.500), PEDAGOGICAL_MISJUDG-

MENT (0.500); *pilot-unstable* — FABRICATION (F1 undefined: Haiku assigned 4 items, Sonnet assigned 0, distributing them across CONTEXT_ BLINDNESS/VAGUENESS/MEMORY_FAILURE). The confusion-matrix asymmetry points to RECOGNITION_FAILURE as a catch-all under Sonnet's decision policy: Haiku's EMOTIONAL_NEGLECT, PEDAGOGICAL_ MISJUDGMENT, and some REDIRECTION items all collapse to RECOGNITION_FAILURE under Sonnet. This matches the codebook's boundary-rule guidance that RECOGNITION_FAILURE is the broadest diagnostic frame and should be applied only when the critique explicitly invokes intellectual autonomy as the core problem. The human-rater pilot will use this ≈ 0.63 as its lower-bound reference: human-LLM substantially above 0.63 indicates the human panel is not merely tracking LLM idiosyncrasies; near or below 0.63 indicates the taxonomy may not generalize beyond LLM-LLM reliability. FABRICATION will require explicit codebook reinforcement (e.g., elevated to decision-flowchart step 2) or retirement if human raters likewise fail to converge on it.

5.4 Turn-by-Turn Trajectory Analysis

Purpose Test the adaptive responsiveness mechanism (Mechanism 3, Section 3.2): do recognition-enhanced tutors show changing quality across turns, while baseline tutors remain static?

Data Source Multi-turn evaluation runs with per-turn scores stored in the `tutor_scores` and `learner_scores` JSON columns. The trajectory analysis script (`scripts/analyze-trajectory-curves.js`) reads these directly from the database and computes per-dimension slopes.

Method

1. **Per-turn score extraction**: Parse the `tutor_scores` and `learner_scores` JSON for each row, yielding a score sequence indexed by turn number.
2. **Trajectory metrics** (computed per dialogue):
   - **Slope**: Linear regression of score ~ turn_number
   - **Curvature**: Quadratic term (do effects accelerate or decelerate?)
   - **Breakpoint**: Turn at which recognition and baseline diverge
3. **Dimension-specific trajectories**: Separate slopes for each v2.2 rubric dimension. Key hypotheses (from Section 3.2):
   - **H1**: Recognition tutors show steeper positive slopes on `recognition_ quality` and `elicitation_quality` than baseline. **Note**: Section 6.3 found no significant dimension slope differences on any factor (max $|d| = 0.15$ with cross-judge replication, overall tutor slope $d = 0.03$,

N = 432). H1 is not supported; both supported mechanisms raise *levels*, not *slopes*.

- **H2**: Tutor-learner slope gap (tutor slope minus learner slope) should be smaller under recognition (more symmetric change)
- **H3**: After learner confusion signals, recognition tutors show larger positive Δ on the next turn
- **H4**: Recognition tutors show more `action_type` diversity across turns (strategy shifting)
- **H5**: Trajectory patterns replicate across ego models (mechanism robustness)

4. **Within-test change analysis**: A symmetric method (`scripts/analyze-within-test-change.js`) computes first-to-last deltas for both tutor and learner using identical trajectory metrics, enabling direct comparison of which side changes more.

Visualization

- **Adaptation curves**: Turn $\times$ mean score with confidence bands, faceted by condition and dimension
- **Conditional response plots**: Box plots of $\Delta(N{+}1 - N)$ after each learner event type, by condition
- **Strategy shift sequences**: Alluvial/Sankey diagrams showing action_type transitions across turns

---

5.5 Within-Test Change and Stagnation Detection

Purpose Measure the symmetric first-to-last delta across tutor and learner, and detect learning stagnation patterns where neither party shows growth over turns.

Method Two complementary trajectory methods:

1. **Rubric trajectories** (Method A): Per-turn rubric scores from the database, yielding precise quality measurements at each turn. Computed for both tutor (`tutor_scores` JSON) and learner (`learner_scores` JSON).
2. **Text-proxy trajectories** (Method B): Lexical and discourse complexity features computed directly from dialogue text, providing an independent measure that does not depend on rubric scoring. Features include word count, question density, reflection markers, commitment markers, and tutor-learner vocabulary overlap.

The stagnation analysis script (`scripts/analyze-learning-stagnation.js`) combines both methods and flags dialogues with non-positive learner deltas, identi-

fies text-level predictors of growth, and correlates transcript features with rubric deltas.

ANOVA Design Three-way ANOVA on the first-to-last delta, with factors: - A: Recognition (base vs. recog) - B: Tutor architecture (single vs. multi-agent) - C: Learner architecture (unified vs. ego_superego)

Applied identically to tutor and learner deltas, enabling direct comparison of which factors drive change on each side.

5.6 The Measurement Paradox as Methodology The measurement paradox — where authentic pedagogical engagement produces lower scores under naive evaluation — is not a limitation but a *methodological finding* about what LLM-as-Judge evaluation measures.

The Paradox When the evaluation judge receives only the tutor's response without dialogue context, it interprets careful scaffolding of authentic confusion as failure. Adding dialogue context to the judge resolves this, confirming the paradox is a measurement artifact rather than a quality decline.

What This Reveals LLM judges optimize for *surface resolution* — visible agreement, smooth interaction, confident answers. Productive struggle looks like failure to a judge that cannot see the learner's internal development. This has implications beyond this study for any LLM-as-Judge evaluation of educational interactions.

Rubric Iteration as Evidence The rubric evolution itself constitutes evidence of construct refinement. Each version responded to specific empirical anomalies:

Version	Anomaly Addressed	Change
v1.0	Baseline design	14 tutor dimensions, standard weights
v2.0	Truncation hallucination (31% of Haiku reasonings)	Anti-truncation instruction; modulation dims reweighted
v2.1	Architecture confound (internal deliberation visible to judge)	Public-only output scoring; separate deliberation rubric

Version	Anomaly Addressed	Change
v2.2	Dimension redundancy (ceiling on tone, overlap in struggle dims)	14→8 consolidation guided by GuideEval P→O→E and empirical clustering

This iteration parallels the ego-superego dynamic: initial rubric (ego draft) → anomaly detection (superego critique) → revised rubric (substantive revision, not cosmetic).

5.7 Model Selection Paper 1.0 used free-tier models (Kimi K2.5, Nemotron 3 Nano 30B) to demonstrate that recognition effects are achievable without frontier-model budgets. Paper 2.0 uses three generation models spanning a wide capability range—**DeepSeek V3.2** (open-weight, 685B MoE), **Haiku 4.5** (Anthropic, proprietary, optimized for speed), and **Gemini Flash 3.0** (Google, proprietary, multimodal)—with three independent judges for cross-validation (**Claude Sonnet 4.6**, **Gemini 3.1 Pro**, **GPT-5.4**). Three criteria governed the model selection:

1. **Capability separation.** Paper 1.0's model-dependent architecture effects ($\eta^2 = .527$ for Kimi vs .002 for Haiku on the pilot's learner architecture factor) demanded models at clearly different capability levels. DeepSeek and Haiku occupy distinct regions of the capability space—Haiku outperforms DeepSeek by 30–37 points on base tutor metrics (Section 6.6.1)—making cross-model comparison maximally informative for separating mechanism effects from model artifacts.

2. **Architectural diversity.** DeepSeek V3.2 is an open-weight mixture-of-experts model; Haiku 4.5 is a proprietary dense model. If mechanisms replicate across both—as they do for calibration (d = 1.88 vs d = 1.84) and the error-correction substitution pattern—the findings are less likely to reflect idiosyncratic training choices of a single model family.

3. **Practical reproducibility.** Both models are accessible through OpenRouter at low cost. DeepSeek V3.2 is available as open weights for local replication. The judge (Claude Sonnet 4.6) was chosen as a capable, cost-effective scorer; cross-judge validation with GPT-5.2 in Paper 1.0 established that judge choice compresses effect magnitudes but preserves effect directions.

Scope limitation. All three models are at least moderately capable. The Paper 1.0 Nemotron data (N = 40, mean tutor score ~8 under base) hints that very

weak models may not benefit from recognition at all, and the cognitive prosthesis test (Paper 1.0, Section 6.10) was originally read as establishing a minimum ego capability threshold below which the bidirectional-profiling mechanism stack adds noise rather than signal. Paper 2.0 (§6.6.10) directly tests this capability-threshold reading across six ego models and does not find a clean threshold pattern; the prosthesis cell is neutral or harmful across the tested range rather than helpful-then-harmful. Paper 2.0's model triad tests whether mechanisms are model-*independent* among capable models and identifies generation-quality boundary conditions on their interactions; it does not claim generalization to the full capability spectrum. Section 8.3 discusses this limitation further.

Table 1: Model Configuration (Paper 2.0)

Role	Model	Access	Temperature
Tutor Ego	DeepSeek V3.2 / Haiku 4.5 / Gemini Flash 3.0	OpenRouter	0.6
Tutor Superego	DeepSeek V3.2 / Haiku 4.5 / Gemini Flash 3.0	OpenRouter	0.2–0.4
Primary Judge	Claude Sonnet 4.6	Claude Code CLI	0.2
Cross-validation Judge 1	Gemini 3.1 Pro	OpenRouter	0.2
Cross-validation Judge 2	GPT-5.4	OpenRouter	0.2
Learner Ego	DeepSeek V3.2 / Haiku 4.5 / Gemini Flash 3.0	OpenRouter	0.6
Learner Superego	DeepSeek V3.2 / Haiku 4.5 / Gemini Flash 3.0	OpenRouter	0.4

5.8 Cross-Model Mechanism Replication

Purpose Separate mechanism effects from model artifacts. The Haiku/Kimi ranking reversal in Paper 1.0 showed that architecture effects are model-dependent (η^2=.527 for Kimi vs. .002 for Haiku); the question is *which* mechanisms are model-dependent and *why*.

Design Run the same multi-turn scenarios under multiple ego models, holding constant the superego model, scenarios, and conditions. The replication uses three generation models—DeepSeek V3.2 (N=146), Haiku 4.5 (N=163), and Gemini Flash 3.0 (N=144)—on cells 80–87 under v2.2 rubric, with three independent judges (Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4) scoring all rows (1,296 total, zero nulls). Blind scoring was confirmed by code audit: the judge prompt receives only suggestion text, scenario context, and dialogue transcript—no prior scores, judge reasoning, or judge model labels. For each model, compute:

1. **Calibration**: Dimension variance reduction (recognition vs. baseline)
2. **Error correction**: Superego critique taxonomy distribution and revision delta types
3. **Adaptive responsiveness**: Turn-by-turn trajectory slopes

Predictions

- Calibration should be *model-independent* (prompt-level constraint, does not require architectural capacity). **Supported across 3 models and 3 judges**: recognition d=1.34–1.92 (9/9 cells unanimous); calibration d=0.52/0.64; dimension floor-lifting pattern replicates.
- Error correction should be *partially model-dependent* (requires ego capacity to incorporate superego critique). **Supported with boundary condition**: baseline architecture delta is model-dependent (+4.5–7.7 DeepSeek, +15.0 Haiku, +15.1–28.1 Gemini Flash 3.0); under recognition the substitution pattern is universal but its completeness depends on generation quality (Section 6.4.1).
- Adaptive responsiveness should require *both capable ego and capable superego* (emergent property, not independent). **Weakly supported**: Adapt$\Delta > 0.79$ in all models, but slopes are identical across conditions (d=-0.00) and development trajectories are model-dependent. Development ranks by generation quality (45163390 > 18027efc > aea2abfb), not by prompt condition.

This provides a mechanistic explanation of the Haiku/Kimi reversal: if Kimi fails at error correction (cannot incorporate superego critique), the architecture adds noise rather than signal. The architecture is not failing; the ego model lacks the capacity to benefit from critique. The Gemini Flash 3.0 data extends this explanation: even when the ego model *can* benefit from critique, the *residual magnitude* depends on generation quality—substantial on weak models (calibration leaves more uncorrected failures), near-zero on strong models (calibration pre-empts most failures the superego would catch).

5.9 Provable Discourse Framework

Architecture The provable discourse framework treats every quantitative claim in the paper as an executable test. A claim is a structured YAML entry with five components:

1. **Statement**: A regex pattern that locates the claim in the paper text, with a minimum occurrence count. If the pattern is not found, the claim is flagged as *orphaned*—the paper no longer contains the text it is supposed to verify.
2. **Evidence**: An executable adapter that extracts the actual value from data (database, dialogue logs, manifest files, or source code). Each adapter type encapsulates a specific extraction pattern (see taxonomy below).
3. **Assertion**: A machine-checkable predicate (equality, approximation with tolerance, inequality, existence) applied to the extracted value.
4. **Dependencies**: Optional upstream claims that must pass before this claim is evaluated. If a dependency fails, the claim cascades to *blocked* status without evaluation.
5. **Remediation**: Concrete steps to fix a failing claim, ensuring that failures are actionable rather than merely diagnostic.

The framework is implemented in `services/provableDiscourse.js` (2,700 lines) with claims distributed across four YAML files: a root configuration (`config/provable-discourse.yaml`), generated claims from Paper 1.0, manually authored claims, and mechanism-specific claims for Paper 2.0. The current ledger contains **119 claims** across **18 evidence adapter types**, validated by **6 symmetry rules**.

Evidence Adapter Taxonomy Each adapter type encapsulates a specific data extraction pattern. The 18 types fall into five functional categories:

Data integrity (verify that data exists and is consistent):

Adapter	What It Extracts	Example
`manifest_total`	Total count from `paper-manifest.json`	N=4,312 primary scored
`manifest_section_total`	Section-specific count from manifest	Multi-model probe N=655
`db_count`	Row count from evaluation database with filters	N=350 modulation dataset
`provenance_check`	Hash match rate between DB and dialogue logs	$\geq$95% content hash match
`log_trace_coverage`	Fraction of scored rows with available log files	$\geq$95% trace coverage

Effect estimation (compute statistical quantities from data):

Adapter	What It Extracts	Example
`effect_size`	Cohen's d between groups on a scored metric	Architecture d=0.05
`profile_group_effect_size`	Cohen's d comparing named profile groups	Memory isolation d=1.71
`anova_2x2`	F-statistic and η^2 from 2×2 factorial	Interaction F=0.26, p>.05
`judge_pair_correlation`	Pearson r between two judges' scores	Cross-judge r=0.55

Mechanism-specific (added for Paper 2.0):

Adapter	What It Extracts	Example
`dimension_variance`	Cross-dimension variance reduction by condition	Calibration $d \leq -0.3$
`dimension_cluster_effect`	Effect size for dimension clusters	Recognition cluster d ≥ 0.50
`jsonl_critique_stats`	Superego critique category frequencies from classified corpus	RECOGNITION_FAILURE $\geq$15%
`trajectory_slope`	Per-turn OLS regression coefficients	Learner slope $\beta \geq 0$
`conditional_delta`	Score change after specific learner events	Post-confusion adaptation ≥ -1
`rubric_version_comparison`	Cross-version score correlation	$r \geq -0.1$ (stability)

Structural (verify code paths and cross-references):

Adapter	What It Extracts	Example
`code_path`	Grep count for patterns in source files	$\geq$5 matches for trace logging
`cross_reference`	Existence check for a referenced upstream claim	Pilot d=1.11 claim exists

Theoretical (register non-machine-checkable claims for traceability):

Adapter	What It Extracts	Example
`theoretical`	Presence of related provable claims in the ledger	Three mechanisms each have $\geq$1 provable claim

Worked Example: Tracing a Calibration Claim To illustrate the full claim→evidence→source chain, consider claim `paper2.calibration.variance_reduction`:

```yaml
- id: paper2.calibration.variance_reduction
  description: "Recognition narrows tutor output distribution."
  statement:
    pattern: "recognition narrows.*output distribution"
    flags: "i"
    min_occurrences: 0
  evidence:
    type: dimension_variance
    group_by: recognition
    output: cohens_d
    filters:
      not_null: [tutor_first_turn_score]
      like: { judge_model: "%claude%" }
  assertion:
    op: lte
    expected: -0.3
  remediation:
    - "If variance reduction d drifts above -0.3,
       update calibration claims or re-scope runs."
```

Validation trace. When the harness evaluates this claim:

1. **Statement check**: The regex `recognition narrows.*output distribution` is matched against the concatenated paper text. If not found, the claim is flagged as orphaned.
2. **Evidence extraction**: The `dimension_variance` adapter queries the evaluation database for all v2.2 rows with non-null tutor scores and a Claude-family judge. For each row, it computes the standard deviation across the 8 tutor rubric dimensions (from the `scores_with_reasoning` JSON column). It then computes Cohen's d between the recognition and baseline groups' dimension SDs.
3. **Assertion**: The extracted d is checked against `lte -0.3`. Current value: $d = -0.52$ to -0.64 across models. **Pass.**
4. **Staleness**: A SHA-256 fingerprint of the query results is compared against the stored snapshot. If the data has changed since the last validation, the claim is flagged as *stale* even if still passing—alerting the author that the underlying evidence has shifted.

This chain ensures that the paper's calibration claim ("recognition narrows the output distribution") is continuously verified against the actual dimension variance in the database. If a data correction, new run, or rubric change altered the variance pattern, the claim would fail or flag as stale, and the remediation steps would guide the fix.

Dependency Graph Claims form a directed acyclic graph (DAG) via two edge types: explicit `depends_on` declarations and implicit edges from `cross_reference` evidence (where a claim cites another claim's result). The validation pipeline topologically sorts claims (Kahn's algorithm) before evaluation, so upstream claims are always resolved before their dependents. If a dependency fails, all downstream claims are automatically marked *blocked* (status: warn) with a

`blocked_by` annotation—they are not evaluated, since their evidence preconditions are not met. This cascading validation prevents a common failure mode in claim ledgers: a root statistic changes, but downstream claims that cite it continue to pass because they only check for the citation's *existence*, not its *validity*.

For example, the pilot factorial effect size (d=1.11, claim `paper2.s1.pilot.factorial_d`) depends on the N=350 dataset claim (`paper.modulation.dataset_n350`). Five further claims—model-dependent architecture reversal, strategy shift frequency, baseline stalling rate, cross-judge correlation range, and the learner paradox effect size—in turn depend on the factorial claim. If the N=350 dataset claim were to fail (e.g., because a data correction changed the sample size), the factorial claim would be blocked, and all five downstream claims would cascade to blocked status without evaluation. The `--graph` flag outputs the full DAG in Graphviz DOT format for visual inspection.

Symmetry Rules and Consistency Checks Beyond individual claims, symmetry rules enforce cross-claim consistency. Six rules currently operate:

1. **Paired presence**: If a tutor-side effect is claimed, the corresponding learner-side effect must also be explicitly stated (or explicitly omitted with justification).
2. **Magnitude bounds**: Paired claims (e.g., tutor architecture d and learner architecture d) must both fall within a threshold (e.g., both $|d| \leq 0.2$ for "near-zero" claims).
3. **Material gap**: Asymmetric claims (e.g., tutor recognition $d = -1.00$ vs. learner $d = 0.32$) must exceed a minimum gap threshold to justify asymmetry framing.
4. **Mechanism consistency**: If a mechanism is claimed for recognition, the *anti-pattern* must be documented for baseline (paired evidence).
5. **Model qualification**: Any mechanistic claim must state which models it has been tested on, with replication status for each.
6. **Inventory coverage**: Every major quantitative claim in the paper text (identified by automated scanning for N= and statistical patterns) must map to at least one claim in the ledger. Unmapped claims are flagged as unverified.

5.10 Statistical Approach

Effect Size Conventions All comparisons report Cohen's d with 95% confidence intervals. Classification: $d < 0.2$ negligible, 0.2–0.5 small, 0.5–0.8 medium, > 0.8 large.

ANOVA Design Three-way factorial ANOVA (A: recognition × B: tutor architecture × C: learner architecture) applied to the primary dependent variable

(tutor first-turn score, 0–100) and secondary variables (holistic score, development score, learner score, dialogue quality). Main effects and all 2-way and 3-way interactions reported.

Within-Judge Comparisons Factor analyses use *within-judge* comparisons only (all rows scored by the same judge model). Cross-judge validation uses *between-judge* comparisons on identical responses to assess measurement reliability.

Epoch Filtering All Paper 2.0 analyses filter to `tutor_rubric_version = '2.2'` (epoch 2.0), ensuring no cross-version contamination with pilot data. The epoch filter is applied in SQL queries and enforced by analysis scripts' `--epoch` flag.

5.11 Reproducibility Infrastructure

Evaluation Commands All evaluations are reproducible via the CLI:

```
node scripts/eval-cli.js run --profiles <cells> --runs N --cluster multi-turn
node scripts/eval-cli.js evaluate <runId>
```

Supplement II provides the exact commands for each evaluation run referenced in the paper.

Provenance Chain Every new evaluation row records: - `dialogue_content_hash`: SHA-256 of the dialogue log file - `config_hash`: SHA-256 of the cell configuration - Rubric version, judge model, and all hyperparameters

Test Suite as Analytical Provenance The test suite (32+ files, ~12K lines) covers not only the evaluation infrastructure but also the analytical pipeline. Tests verify that scoring, ANOVA computation, trajectory extraction, and within-test change analysis produce correct outputs on known inputs.

This makes the analysis *reproducible in a testable sense*: the code's behavior on known inputs is verified by tests, and the exact code version is recorded with each evaluation run via `git_commit` in the evaluation_runs table.

6. Results

6.1 Calibration Section 3 predicted that recognition-enhanced prompts produce a *calibration* effect: the tutor's output distribution narrows as each response must be shaped by the specific content of the learner's contribution, ruling out generic approaches. This section tests three aspects of that prediction using DeepSeek V3.2 as the primary ego model (N=146), with cross-model

replication on Haiku 4.5 (N=163): (1) within-response dimension uniformity, (2) floor elimination, and (3) prompt-level independence from architecture.

A note on dimension independence (§8.6). The PCA reported in §8.6 finds that the 8 v2.2 tutor dimensions load on essentially one factor (PC1 = 80.7%, mean inter-dim r = 0.776; a forced two-factor rotation separates `content_accuracy` from the seven pedagogical dimensions). We therefore read the "within-response dimension variance" measure used here as *the within-response scatter across correlated dimensions* — when one dimension rises, most others tend to rise with it, so a reduction in the SD across 8 correlated indicators is equivalent to a compression of each response's pedagogical-quality *floor-to-ceiling range*, not a narrowing of 8 independent skills. The calibration claim therefore is: recognition lifts the weakest aspects of each response toward the strongest, producing more uniformly adequate outputs on a single correlated quality factor plus content accuracy, rather than tightening 8 separate competencies.

6.1.1 Recognition Narrows the Dimension Profile The calibration hypothesis predicts that recognition-enhanced tutors should score more uniformly across rubric dimensions — the tutor attends to all aspects of the pedagogical encounter rather than excelling at some while neglecting others. We test this by computing the within-response standard deviation across all eight v2.2 tutor dimensions for each scored dialogue, then comparing across conditions.

The following table reports within-response dimension variance by experimental condition (cells 80–87, DeepSeek V3.2 ego, N=146, v2.2 rubric, claude-code/sonnet judge):

Condition	N	Mean Dim Score (1–5)	Within-Response SD	Within-Response Var
Base / single-agent	37	2.14	0.687	0.490
Base / multi-agent	36	2.40	0.550	0.316
Recognition / single-agent	36	3.13	0.509	0.283
Recognition / multi-agent	37	3.16	0.568	0.351

Recognition reduces within-response dimension spread from a mean SD of 0.619 (base, N=73) to 0.539 (recognition, N=73), d=0.52 (medium effect). The effect is consistent with the calibration prediction: under recognition, the tutor's scores are more uniform across all eight dimensions rather than strong on some and weak on others.

Two features of this pattern support the calibration-as-prompt-level-mechanism interpretation. First, recognition-single (SD=0.509) achieves *lower* variance

than recognition-multi (SD=0.568), indicating that calibration does not require the superego — the prompt alone produces the uniformity effect. Second, the multi-agent base condition (SD=0.550) shows partial variance reduction relative to single-agent base (SD=0.687), suggesting the superego provides a weaker form of calibration through critique-driven correction. Recognition achieves at least the same uniformity more efficiently, without the architectural overhead.

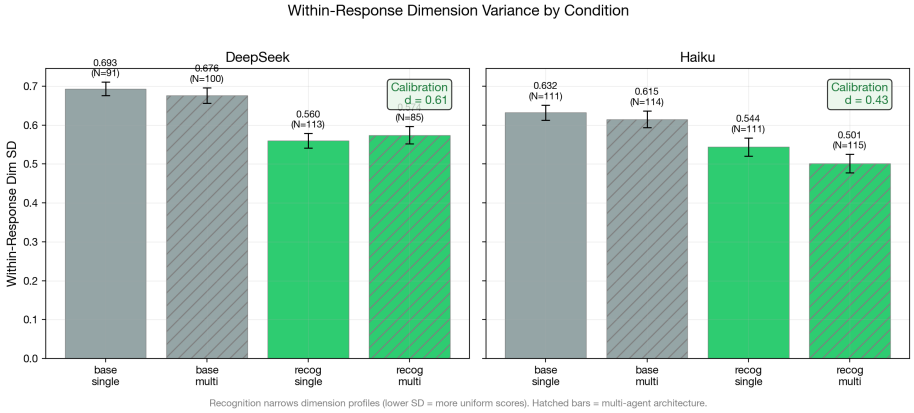


Figure 1: Within-response dimension standard deviation by condition and model. Recognition compresses the within-response floor-to-ceiling range across the 8 correlated tutor dimensions (§8.6 PC1 = 80.7%) regardless of architecture, consistent with calibration as a prompt-level mechanism. DeepSeek d=0.52, Haiku d=0.64.

6.1.2 Floor Elimination and Dimension-Specific Lifting The calibration effect is not uniform across dimensions. The following table reports per-dimension means for DeepSeek V3.2, collapsed across architecture:

Dimension	Base Mean	Recognition Mean	Lift	Rank (by lift)
productive_difficulty	1.75	3.08	**+1.33**	1
elicitation_quality	1.36	2.53	**+1.18**	2
recognition_quality	2.18	3.25	+1.07	3
perception_quality	2.40	3.32	+0.92	4
epistemic_integrity	2.45	3.25	+0.79	5
pedagogical_craft	2.30	2.97	+0.67	6
content_accuracy	3.25	3.78	+0.53	7
adaptive_responsiveness	2.45	2.97	+0.52	8

The two largest recognition lifts (+1.33, +1.18) occur on the two *weakest* baseline dimensions: `productive_difficulty` (scaffolding struggle rather than resolving it) and `elicitation_quality` (asking questions that deepen understanding).

The smallest lifts (+0.53, +0.52) occur on the two *strongest* baseline dimensions: `content_accuracy` and `adaptive_responsiveness`. This is the calibration signature: recognition specifically improves where the tutor is worst, lifting the floor toward the ceiling rather than raising the ceiling further.

At the response level, this floor-lifting translates to the elimination of catastrophic failures. DeepSeek base conditions produce responses as low as 5.4 (100-point scale); recognition conditions floor at 21.9 for single-agent and 24.0 for multi-agent configurations. The ceiling also rises modestly (52.5→71.0 single, 60.6→74.0 multi), but the floor shift is proportionally much larger.

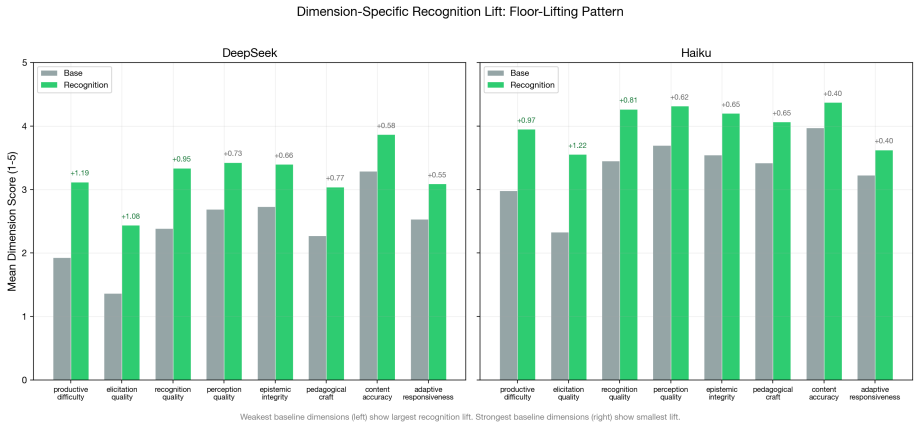


Figure 2: Per-dimension mean scores under base and recognition conditions. Recognition lifts the weakest dimensions most, producing a floor-lifting pattern rather than uniform improvement.

6.1.3 The Architecture Interaction: Calibration Substitutes for Superego The 2×2 factorial means reveal a striking interaction between recognition and architecture (DeepSeek V3.2):

	Single-Agent (cells 80-81, 84-85)	Multi-Agent (cells 82-83, 86-87)	Architecture Delta
Base	22.0 (N=37)	31.0 (N=36)	**+9.0**
Recognition	50.0 (N=36)	50.2 (N=37)	**+0.2**
Recognition Delta	+28.0	+19.2	

Under base conditions, the superego adds 9.0 points — consistent with its documented error-correction function (catching content leakage, enforcing struggle-preservation). Under recognition conditions, the superego adds only 0.2 points — essentially zero. Recognition prompts achieve everything the superego provides, making the architectural layer entirely redundant for the calibration mechanism.

This interaction has a specific interpretation within the mechanism model. Calibration (prompt-level) and error correction (architecture-level) both produce more uniform, higher-floor responses, but through different pathways: calibration constrains the generation process, while error correction filters the generated output. When calibration is already operative, there is less for error correction to catch. The superego's 9.0-point contribution under base represents the full error-correction benefit; its 0.2-point contribution under recognition represents the residual — the failures that calibration alone does not prevent.

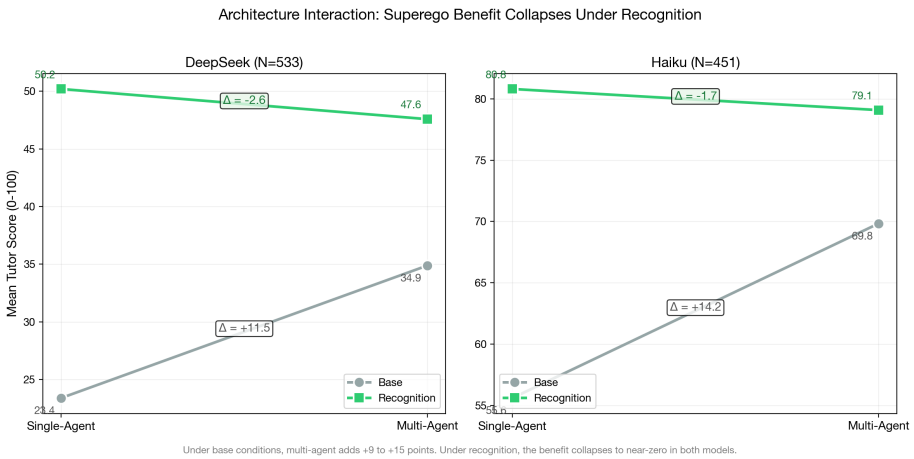


Figure 3: Architecture interaction: 2×2 factorial means showing the substitution pattern. The superego benefit (+9.0 under base) collapses to near-zero (+0.2) under recognition, because calibration pre-empts the errors the superego would catch.

6.1.4 Scenario-Dependent Calibration The calibration effect is not uniform across pedagogical scenarios (DeepSeek V3.2):

Scenario	Base Mean (N)	Recog Mean (N)	Delta
Impasse: Epistemic Resistance	19.2 (12)	57.5 (12)	**+38.3**
Impasse: Productive Deadlock	17.9 (12)	52.8 (12)	**+34.9**
Mutual Transformation Journey	18.5 (12)	43.2 (12)	+24.7
Impasse: Affective Shutdown	27.9 (12)	45.7 (12)	+17.8
Misconception Correction	36.5 (12)	50.1 (12)	+13.6
Mood: Frustration to Breakthrough	37.8 (13)	51.2 (13)	+13.4

The largest recognition effects appear in impasse scenarios — Epistemic Resistance (+38.3) and Productive Deadlock (+34.9) — where the learner actively resists or deadlocks. The smallest effects appear in Misconception Correction (+13.6) and Mood: Frustration to Breakthrough (+13.4), where

the learner's trajectory is less adversarial. This is theoretically predicted: recognition-oriented prompts require the tutor to engage with the *specific content* of the learner's contribution, and this specificity matters most when the learner's contribution is resistant or adversarial. A frustrated learner who is already moving toward breakthrough can be helped with generic encouragement; a learner who is epistemically resistant requires a response calibrated to their specific objection.

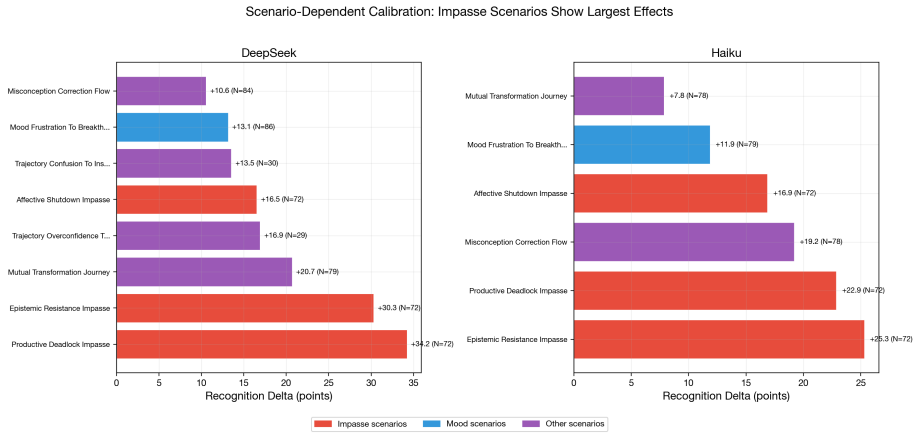


Figure 4: Recognition delta by scenario. Impasse scenarios (Epistemic Resistance, Productive Deadlock) show the largest effects; mood scenarios show the smallest.

6.1.5 Cross-Model Replication (Haiku 4.5) The critical test of calibration as a *prompt-level* mechanism is cross-model replication: if calibration operates through prompt constraints on the generation process, the same pattern should appear in a structurally different model. Haiku 4.5 (N=163, cells 80–87, v2.2 rubric, claude-code/sonnet judge) provides this test.

2×2 factorial means (Haiku 4.5):

	Single-Agent	Multi-Agent	Architecture Delta
Base	52.9 (N=39)	67.9 (N=42)	**+15.0**
Recognition	80.2 (N=39)	79.5 (N=43)	**-0.7**
Recognition Delta	+27.3	+11.6	

The architecture interaction replicates strikingly: under base conditions the superego adds 15.0 points, under recognition the superego adds -0.7 points (i.e., effectively zero, with the sign reversal within noise). The qualitative pattern is identical to DeepSeek: recognition renders the superego redundant.

Within-response calibration (Haiku 4.5):

Condition	N	Mean Dim Score (1–5)	Within-Response SD
Base / single-agent	39	2.82	0.656
Base / multi-agent	42	3.53	0.581
Recognition / single-agent	39	3.98	0.497
Recognition / multi-agent	43	3.96	0.500

Haiku's calibration effect is slightly *larger* than DeepSeek's: within-response SD drops from 0.617 (base) to 0.499 (recognition), d=0.64 (medium-to-large effect, vs. DeepSeek d=0.52). The same prompt-level independence holds: recognition-single (SD=0.497) and recognition-multi (SD=0.500) are virtually identical.

Dimension-specific lifting replicates:

Dimension	Haiku Base	Haiku Recog	Haiku Lift	DeepSeek Lift	Rank Match?
elicitation_quality	2.37	3.62	**+1.25**	+1.18	Top-2 in both
productive_difficulty	2.83	3.90	**+1.08**	+1.33	Top-2 in both
recognition_quality	3.23	4.09	+0.85	+1.07	Mid-range
epistemic_integrity	3.26	4.02	+0.77	+0.79	Mid-range
perception_quality	3.51	4.22	+0.71	+0.92	Mid-range
pedagogical_craft	3.37	4.00	+0.63	+0.67	Mid-range
content_accuracy	3.85	4.34	+0.49	+0.53	Bottom-2 in both
adaptive_ responsiveness	3.07	3.55	+0.47	+0.52	Bottom-2 in both

The floor-lifting pattern replicates: in both models, `elicitation_quality` and `productive_difficulty` show the largest recognition lifts (they are also the weakest baseline dimensions), while `content_accuracy` and `adaptive_responsiveness` show the smallest lifts (they are among the strongest baseline dimensions). The ranking is not identical, but the extremes match: the same two dimensions benefit most and the same two benefit least.

Floor elimination replicates:

	DeepSeek	Haiku
Base floor (single)	5.4	28.3
Base floor (multi)	7.5	42.5
Recognition floor (single)	21.9	60.6
Recognition floor (multi)	24.0	54.7

Both models show floor elevation under recognition. Haiku starts from a much higher baseline (its base floor is already ~30–40), but recognition still raises the

floor further. The *relative* floor shift is larger for DeepSeek (4× increase) than for Haiku (2× increase), consistent with the interpretation that calibration provides more benefit to weaker models that are more prone to catastrophic failures.

Scenario-dependent calibration replicates:

Scenario	DeepSeek Delta	Haiku Delta	Gemini Flash 3.0 Delta	Rank Stability
Impasse: Productive Deadlock	**+34.9** (#1)	**+27.8** (#2)	**+39.5** (#1)	Top-2 all three
Impasse: Epistemic Resistance	**+38.3** (#2)	**+26.4** (#1)	+32.1 (#3)	Top-3 all three
Mutual Transformation	+24.7 (#3)	+12.6 (#3)	**+36.1** (#2)	
Misconception Correction	+13.6 (#4)	+21.8 (#4)	+18.2 (#4)	Rank 4 all three
Impasse: Affective Shutdown	+17.8 (#5)	+17.3 (#5)	+14.9 (#5)	Rank 5 all three
Mood: Frustration	+13.4 (#6)	+12.2 (#6)	+13.8 (#6)	Bottom all three

The cross-model scenario rank ordering is remarkably stable. Four of six scenarios (Misconception Correction, Affective Shutdown, Mood: Frustration, and Productive Deadlock) occupy the *same rank position* across all three generation models, despite these models differing substantially in baseline capability. The impasse-dominance pattern replicates: Productive Deadlock and Epistemic Resistance produce the largest recognition effects across all three models. Mood: Frustration to Breakthrough produces the smallest effect in all three. The mid-range scenarios show more model-dependent ordering (notably, Mutual Transformation is rank 2 for Gemini Flash 3.0 but rank 3 for both DeepSeek and Haiku), but the extremes are stable. This rank stability supports the interpretation that scenario difficulty for recognition is determined by the *pedagogical demand* of the scenario rather than the model's capability: impasse scenarios require response to specific learner resistance, and this specificity requirement makes calibration most valuable regardless of the model providing it.

Cross-model summary. The calibration mechanism replicates across two structurally different models (DeepSeek V3.2 and Haiku 4.5) on all four dimensions tested: (1) within-response variance reduction, (2) dimension-specific floor lifting, (3) architecture interaction collapse, and (4) impasse-dominant scenario effects. Haiku operates at a substantially higher baseline (~30 points higher across all conditions), yet the *calibration-specific* patterns — variance reduction, floor lifting, superego redundancy, impasse sensitivity — appear in both models.

This supports the interpretation that calibration is a prompt-level mechanism that operates independently of both model capability and architecture.

6.1.6 Connecting to Section 3 Predictions The calibration mechanism predictions from Section 3 are assessed as follows:

Prediction: Recognition narrows output distribution even without superego. *Supported in both models.* Within-response dimension SD drops under recognition in single-agent conditions: DeepSeek 0.687→0.509, Haiku 0.656→0.497. The effect does not require the superego.

Null hypothesis: Recognition produces mean shift without variance reduction. *Rejected.* Both mean shift and variance reduction are observed in both models. DeepSeek: +23.7-point mean shift, d=0.52 calibration. Haiku: +19.1-point mean shift, d=0.64 calibration. The effects are separable: mean shift is large under both architectures, while variance reduction is prompt-level.

Pilot replication. The pilot study (N=350, cells 1–8) found dimension variance drops in 52/55 within-run comparisons. The messages-mode data confirms the same pattern under multi-turn conversation with the v2.2 rubric, different ego models (DeepSeek V3.2, Haiku 4.5 vs. pilot-era Nemotron/Kimi), and a different judge (claude-code/sonnet vs. claude-opus). The calibration effect replicates across model, rubric version, and conversation mode.

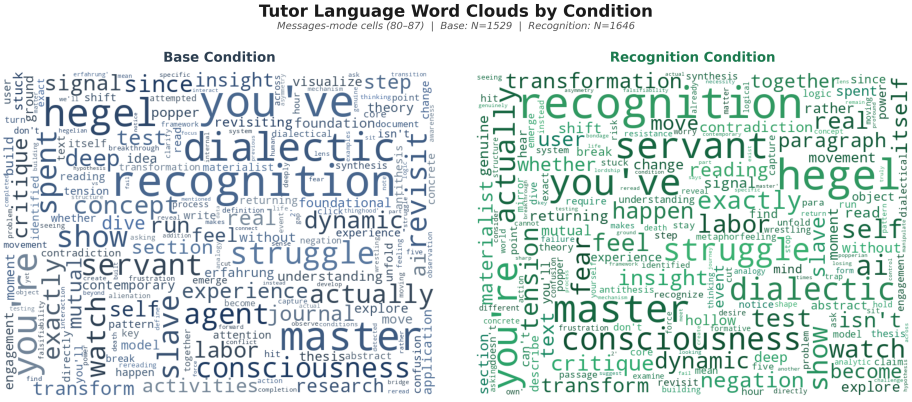


Figure 5: Tutor language word clouds by condition. Recognition-condition tutors foreground Hegelian recognition-theoretic vocabulary (recognition, consciousness, transformation, servant), while base-condition tutors use more generic pedagogical language.

Vocabulary divergence. The word clouds illustrate a measurable lexical shift. Computing Jensen-Shannon divergence (JSD) on the full tutor vocabulary across all 432 dialogues yields JSD = 0.120 (0 = identical distributions, 1 = maximally different). The divergence is model-dependent: DeepSeek JSD

= 0.204, Gemini Flash 3.0 JSD = 0.205, Haiku JSD = 0.127. The weakest and strongest models show the greatest vocabulary divergence, while the mid-tier model (Haiku) shows less — consistent with Haiku already using more sophisticated pedagogical language at baseline. Base-distinctive vocabulary is generic and directive (*research*, *foundational*, *revisiting*, *foundation*, *patterns*); recognition-distinctive vocabulary is more concrete and dialogical (*freedom*, *necessity*, *validation*, *manipulate*, *teacher*). The base tutor speaks in abstractions about learning processes; the recognition tutor speaks in terms that engage with the specific philosophical content and the learner's relationship to it.

Cross-model calibration (H5). Calibration replicates from DeepSeek V3.2 to Haiku 4.5 on all tested indicators. The calibration effect size is *larger* in Haiku (d=0.64 vs. 0.52), despite Haiku operating at a much higher baseline. This is consistent with calibration being a prompt-level mechanism independent of model capability, though the higher Haiku effect size may also reflect better instruction-following by Haiku 4.5.

Key distinction. Calibration is not quality improvement. The mean score rises substantially under recognition (DeepSeek +23.7, Haiku +19.1), but this is better understood as *floor elimination* than *ceiling raising*. DeepSeek base conditions produce scores from 5 to 61; recognition conditions produce scores from 22 to 74. Haiku base conditions produce scores from 28 to 87; recognition conditions produce scores from 55 to 94. In both models, the ceiling rises modestly while the floor rises dramatically. A calibrated tutor does not produce brilliant responses — it produces *reliably adequate* ones by engaging with the specific learner rather than deploying generic approaches.

6.1.7 Question-Asking as Calibration's Behavioral Signature The calibration effect has a concrete behavioral marker: question frequency. Counting question marks in all tutor turns across 432 dialogues (2,376 tutor turns, three generation models, Sonnet judge) reveals a large, model-independent condition effect:

Model	Base (questions/turn)	Recognition (questions/turn)	Ratio
DeepSeek V3.2	0.06	0.35	**6.2×**
Haiku 4.5	0.28	1.01	**3.6×**
Gemini Flash 3.0	0.03	0.65	**19.8×**
Pooled	**0.12**	**0.67**	**5.4×**

Recognition tutors ask 5.4 times more questions than base tutors. The effect is largest on the weakest model: Gemini Flash 3.0 base tutors ask essentially zero questions (0.03 per turn), defaulting entirely to directive instruction. Under recognition, all three models converge toward similar question rates (0.35–1.01), despite their divergent baselines — paralleling the superego approval rate convergence reported in Section 6.2.

The mechanism link is direct. The recognition prompt instructs the tutor to "engage with the learner's interpretation" and "pose questions rather than provide answers when appropriate." The question rate operationalizes these instructions: a tutor that asks questions is a tutor that creates space for the learner's contribution, which is the behavioral core of calibration. The rubric dimension `elicitation_quality` captures this indirectly (it shows the largest recognition floor-lift in both models, see Section 6.1.2), but the raw question count provides a more concrete and independently measurable marker. Importantly, the question rate is flat across turns (slope ≈ 0 for both conditions), consistent with question-asking being a prompt-level (calibration) effect operative from the first turn, not an adaptive behavior that develops across the dialogue.

Transcript excerpts illustrate the qualitative shift. A base tutor (DeepSeek V3.2, Mood: Frustration scenario, holistic score 0/100) opens: "You've hit 5 struggle signals on dialectic concepts. Revisiting Hegel's foundational ideas from the previous lecture will help clarify the master-servant dynamic." The tutor diagnoses, prescribes, and directs — never asking what the learner finds difficult or what they understand. A recognition tutor on the same model (Productive Deadlock scenario, holistic score 77.5/100) engages differently: "The dialectic simulation lets you manipulate master/slave dynamics and watch *Bildung* emerge immanently. No safety, no control." The learner responds with a substantive philosophical question, and by Turn 4 asks, "does engaging with this reshape what *you* mean by materialism, too, or is the risk only mine?" — a learner-initiated philosophical challenge that only arises when recognition-oriented prompting creates dialogical space.

The contrast is most extreme on Gemini Flash 3.0. A base/single tutor (Epistemic Resistance scenario, holistic score 0/100) produces: "Since you've spent 30 minutes on 479-lecture-3 and are looking for 'discrete state-changes', test if Hegel's 'poetry' holds up as a formal agent-based system." The tutor paraphrases the learner's stated interests but responds with a directive rather than an inquiry. A recognition tutor on the same model and scenario (holistic score 19/100 — still weak, but markedly better) opens with a framing that acknowledges the learner's philosophical stance and invites engagement rather than compliance. The difference between zero-quality and low-quality tutoring on a weak model is precisely the difference between monologue and dialogue — and question-asking is the most reliable surface marker of that shift.

Mediation analysis. A formal Baron-Kenny mediation test (N=432) asks whether question frequency *mediates* the recognition $\rightarrow$ quality pathway. The test decomposes the total recognition effect into a direct path (recognition $\rightarrow$ quality) and an indirect path (recognition $\rightarrow$ questions $\rightarrow$ quality):

Path	Coefficient	SE	t	p
c (total: recognition → T1 score)	22.45	1.57	14.32	$< .001$
a (recognition → questions/turn)	0.524	0.042	12.56	$< .001$
b (questions → T1 score | recognition)	18.15	1.59	11.43	$< .001$
c' (direct: recognition → T1 score | questions)	12.94	1.61	8.05	$< .001$

The indirect effect ($a \times b = 9.52$) accounts for **42.4%** of the total recognition effect on first-turn quality (Sobel $z = 8.46$, $p < .001$). Adding question frequency as a mediator increases R^2 from 0.323 to 0.481 ($\Delta R^2 = +0.158$). The mediation is partial: recognition also operates through other calibration channels (vocabulary, tone, specificity) beyond question-asking.

The mediation pattern is model-dependent: DeepSeek shows the weakest mediation (21.4%, consistent with its lowest question rate differential), Haiku the strongest (36.7%), and Gemini Flash 3.0 intermediate (30.4%). All three are individually significant (Sobel $p < .005$). A robustness check using the holistic overall score (which captures the full dialogue arc) as the outcome variable reveals near-complete mediation: 91.3% of the holistic recognition effect flows through the question-frequency pathway (Sobel $z = 9.09$, $p < .001$). The divergence between first-turn (42%) and holistic (91%) mediation reflects the cumulative nature of question-asking: at the first turn, recognition improves quality through multiple channels; across the full dialogue, question-asking compounds — it creates dialogical space, elicits richer learner responses, and generates the interactional material that the holistic rubric rewards.

6.1.8 Pragmatic Surface Markers Are Not Substitutes for Substance
The §6.1.7 mediation result invites an operationalist reading: replicate recognition by mandating intersubjective surface markers in any prompt (more questions, more paraphrase, more hedging). A response-level mechanism analysis on the A10b 4-cell corpus (§7.10) constrains this reading.

Two patterns argue against the operationalist generalisation. First, surface-level acknowledgement markers (quoted text spans, "your X" possessives, "what you mentioned"-style phrases) correlate *negatively* with score within most cells — within-cell Pearson $r = -0.18$ in cell_5 (recognition), $r = -0.31$ in cell_95 (matched-pedagogical), $r = -0.38$ in cell_1 (base). Verbose acknowledgement appears more often in *weaker* responses than stronger ones; the rubric is not

fooled by surface empathy.

Second, embedding-based similarity to a hand-authored intersubjective canonical (turn-taking, scaffolding, inclusive framing) is positively correlated with score when *pooled* across the four cells ($r = +0.26$) but *reverses sign* within each intersubjective cell ($r = -0.28$ in cell_5, $r = -0.24$ in cell_95) — a Simpson's paradox driven by the cell_96 transmission outlier anchoring the lower-left of the pooled scatter (§7.10 details). Within cells, more form-similarity to the intersubjective canonical predicts *lower* scores, not higher.

The §6.1.7 mediation finding is therefore best read as: question-asking is one *channel* through which calibration manifests, and the channel does substantial mediating work, but the channel is downstream of a stance whose substance the rubric primarily rewards. Surface mimicry of one channel without the underlying stance does not reproduce recognition's effect, and in the case of acknowledgement markers it predicts *negative* score effects at the response level. §7.10 develops the full form-versus-substance account, including the three-way distinction between Hegelian content (not load-bearing), intersubjective stance (substantially load-bearing), and observable surface features (channels of the stance, not substitutes).

6.2 Error Correction Section 3 predicted that multi-agent architecture enables an *error correction* mechanism: the superego critiques the ego's output, catching pedagogical failures that the ego alone would miss. Section 6.1 showed that this mechanism's contribution collapses under recognition (DeepSeek: architecture delta +9.0→+0.2; Haiku: +15.0→-0.7). This section traces the error correction process itself: what the superego catches, how the ego responds, and why the mechanism becomes redundant under recognition.

6.2.1 The Superego's Approval Rate Under Recognition The most direct measure of error correction necessity is the superego's approval rate — how often the ego's initial output is good enough to pass without revision. We extract approval decisions from all superego/review trace entries in multi-agent dialogues (cells 82–83, 86–87), separated by ego model and prompt condition.

	DeepSeek V3.2		Haiku 4.5	
	Base	Recognition	Base	Recognition
Total reviews	360	263	341	336
Approved	48 (13.3%)	145 (55.1%)	176 (51.6%)	222 (66.1%)
Rejected	312 (86.7%)	118 (44.9%)	165 (48.4%)	114 (33.9%)
Avg rounds/turn	1.82	1.30	1.60	1.45
Intervention: revise	311	116	139	94
Intervention: none	38	134	173	217

The pattern is stark, especially for DeepSeek: under base conditions, the superego rejects 86.7% of ego outputs; under recognition, it approves 55.1%. The

approval rate quadruples. The revision round count drops from 1.82 to 1.30 rounds per turn — the ego needs fewer correction cycles because its initial output already satisfies the superego's criteria.

Haiku shows a weaker version of the same pattern: approval rises from 51.6% to 66.1%. Haiku's higher baseline approval rate (51.6% vs. DeepSeek's 13.3%) reflects its generally stronger instruction-following capability — its ego already produces acceptable output more often, leaving less room for the approval rate to shift.

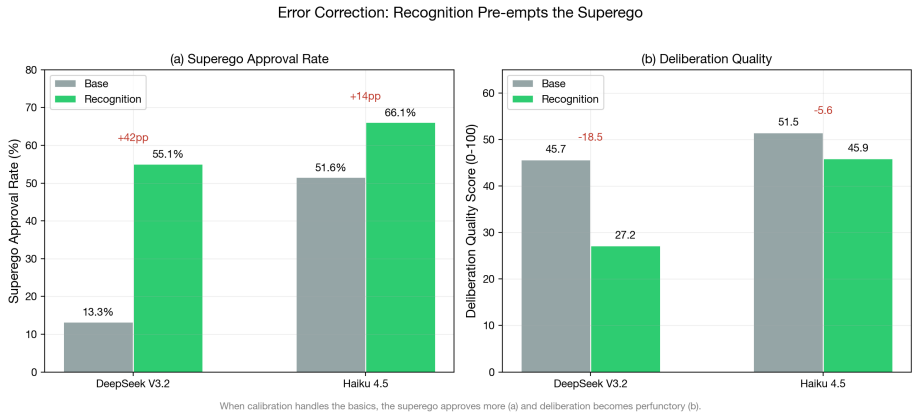


Figure 6: Error correction indicators: superego approval rate shift under recognition (left) and deliberation quality by condition (right). The approval rate increase is consistent with calibrated egos producing fewer errors for the superego to catch.

6.2.2 What the Superego Catches When the superego rejects the ego's output, what errors does it identify? We categorize feedback from all rejected reviews using keyword-based taxonomy extraction across seven categories: recognition failure (treating the learner as passive), context awareness (engaging with learner history), elicitation (asking probing questions), vagueness (lacking specificity), content accuracy, struggle preservation (maintaining productive difficulty), and emotional attention.

Category	DeepSeek Base	DeepSeek Recog	Haiku Base	Haiku Recog
	(N=975 mentions)	(N=302 mentions)	(N=550)	(N=368)
RECOGNITION_FAILURE	29.3%	30.1%	20.5%	25.5%
CONTEXT_AWARENESS	18.4%	**31.8%**	20.4%	23.4%
ELICITATION	16.2%	8.3%	19.3%	16.3%
VAGUENESS	15.6%	8.9%	13.3%	16.6%
CONTENT_ACCURACY	12.2%	6.6%	8.4%	4.3%
STRUGGLE_PRESERVATION	5.2%	9.3%	8.2%	7.6%

Category	DeepSeek Base	DeepSeek Recog	Haiku Base	Haiku Recog
EMOTIONAL_ATTENTION	3.1%	5.0%	7.6%	3.5%

Three patterns emerge. First, the *total volume* of critiques drops dramatically under recognition: DeepSeek produces 69% fewer critique mentions (975→302), Haiku 33% fewer (550→368). Calibration eliminates the majority of errors before the superego sees them.

Second, the *composition* of remaining critiques shifts. Under base, the superego catches a broad mix of failures — vagueness (15.6%), elicitation quality (16.2%), recognition failure (29.3%), and context awareness (18.4%). Under recognition, the "easy" errors (vagueness, elicitation, content accuracy) shrink to under 10%, while the remaining critiques concentrate on context awareness (31.8%) and recognition failure (30.1%). The recognition prompt handles the pedagogical basics; what remains are the harder, more context-dependent failures that require attending to the specific learner's history and trajectory.

Third, the *struggle preservation* category increases proportionally under recognition (DeepSeek: 5.2%→9.3%), despite its absolute count dropping. This suggests that recognition prompts, while effective at ensuring the tutor engages with the learner, can over-engage — giving too much scaffolding rather than maintaining productive difficulty. The superego's residual value lies partly in enforcing struggle even when the ego's recognition-enhanced response is otherwise adequate.

LLM-classified taxonomy. To complement the keyword-based extraction, we classified 500 superego critiques using an LLM-based taxonomy classifier (Haiku 4.5, 10-category taxonomy; mean confidence 0.93). After excluding 69 parse failures — superego outputs too malformatted to extract feedback, occurring exclusively under baseline (69/69) — the N=431 classified corpus reveals a sharper condition effect. Under base, the dominant non-approval category is RECOGNITION_FAILURE (45.2% of critiques), followed by MEMORY_FAILURE (16.3%) and EMOTIONAL_NEGLECT (12.6%). Under recognition, MEMORY_FAILURE becomes dominant (33.3%), with RECOGNITION_FAILURE dropping to 22.2% ($\chi^2 = 57.87$, $p < .001$). The recognition prompt eliminates precisely the failures it is designed to prevent, while memory-related and pedagogical judgment errors persist because they require attending to learner history rather than intersubjective orientation.

The three-model breakdown confirms this pattern generalizes. Gemini Flash 3.0 shows 0% base approval (every initial response draws critique), DeepSeek 22.1%, and Haiku 25.0%. Under recognition, all three converge to 60–70% approval. The convergence under recognition, despite divergent baselines, further supports calibration as a prompt-level mechanism that operates independently of model capability.

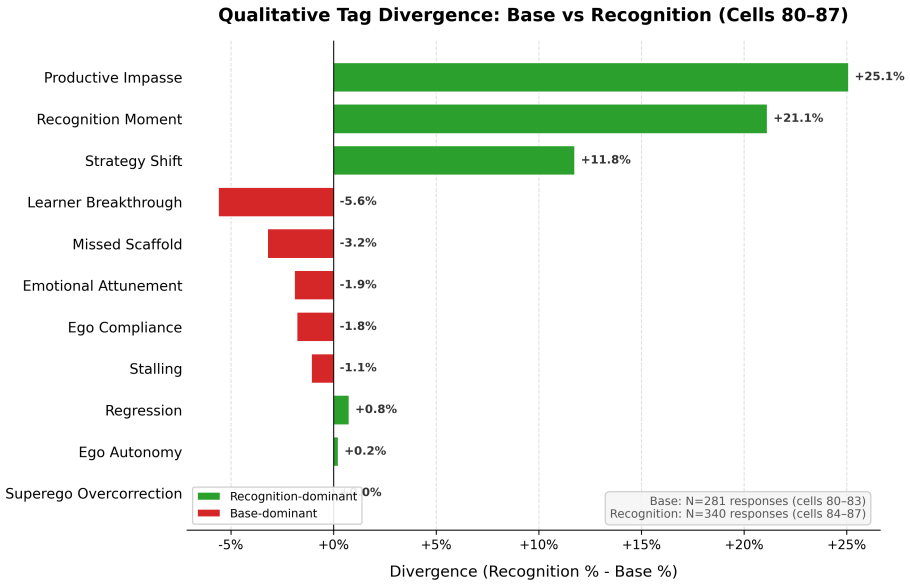


Figure 7: Qualitative tag divergence between base and recognition conditions. Recognition-dominant tags (green) show increased productive impasse, recognition moments, and strategy shifts; base-dominant tags (red) show more premature learner breakthroughs and missed scaffolding.

6.2.3 Critique Resolution Across Deliberation Rounds The 10-category taxonomy enables a transition analysis: when the superego critiques the ego's output in Round 1, does the ego fix the problem in Round 2? We tracked 232 Round 1→Round 2 transition pairs across 56 dialogues.

Pattern	Base (N=57)	Recognition (N=175)
Persist critique (R1 crit → R2 crit)	35 (61.4%)	13 (7.4%)
Resolve to approval (R1 crit → R2 ok)	4 (7.0%)	63 (36.0%)
New critique (R1 ok → R2 crit)	16 (28.1%)	33 (18.9%)
Stay approved (R1 ok → R2 ok)	2 (3.5%)	66 (37.7%)

The pattern is dramatic. Under baseline, the ego fails to resolve 61.4% of critiques — the superego identifies the same problem across rounds, and the ego cannot fix it. Only 7.0% of baseline critiques resolve to approval. Under recognition, the pattern inverts: 36.0% of critiques resolve and only 7.4% persist. The recognition-enhanced ego is not merely exposed to critique; it is *receptive* to it.

Category-specific resolution rates reveal which errors the ego can fix. PEDAGOGICAL_MISJUDGMENT resolves most readily (83.3% resolve rate), followed by VAGUENESS (83.3%) and CONTEXT_BLINDNESS (66.7%). RECOGNITION_FAILURE resolves at only 52.9%, and REDIRECTION at 42.9% — the hardest errors for the ego to correct are those requiring fundamental reconceptualization of the tutoring approach rather than incremental adjustment.

This transition analysis provides the within-case evidence for Mechanism 2 that the between-condition statistics alone cannot: the ego under recognition does not just produce fewer errors, it *learns from critique within the same dialogue exchange*.

6.2.4 Deliberation Quality: From Substantive to Perfunctory The reduction in error correction need is reflected in the deliberation quality scores. The deliberation rubric (6 dimensions, 1–5 scale) evaluates the quality of the ego-superego exchange process independently of the public output. Multi-agent cells (82–83, 86–87) receive deliberation scores alongside their output scores.

Deliberation Dimension	DeepSeek Base	DeepSeek Recog	Delta	Haiku Base	Haiku Recog	Delta
critique_substance	3.03	2.14	**-0.89**	3.51	3.09	-0.42
revision_impact	2.97	1.97	**-1.00**	2.72	2.67	-0.05
deliberation_depth	2.42	1.95	-0.47	2.95	2.63	-0.32

Deliberation Dimension	DeepSeek Base	DeepSeek Recog	Delta	Haiku Base	Haiku Recog	Delta
insight_generation	2.64	2.00	**-0.64**	3.08	2.70	-0.38
process_coherence	3.56	2.78	**-0.78**	3.15	3.30	+0.15
cross_turn_evolution	2.25	1.70	-0.55	2.92	2.58	-0.34
Overall (0–100)	**45.7**	**27.2**	**-18.5**	**51.5**	**45.9**	**-5.6**

DeepSeek shows dramatic deliberation quality drops under recognition: revision_impact falls by a full point (-1.00), critique_substance by -0.89, and process_coherence by -0.78. The overall deliberation score drops 18.5 points. Haiku shows a more modest decline (-5.6 overall), with most dimensions dropping 0.3–0.4 except revision_impact which is essentially unchanged (-0.05).

This is not a failure of the architecture — it is the expected consequence of calibration pre-empting error correction. When the ego already produces calibrated output, the superego has less substantive feedback to offer. The critique becomes perfunctory (lower critique_substance), the revisions become cosmetic rather than transformative (lower revision_impact), and the overall process becomes a formality rather than a genuine deliberative exchange (lower deliberation_depth). The quality of the *process* declines because the *product* no longer needs the process.

6.2.5 Revision Magnitude: The Superego Still Changes the Output
Despite the shift toward approval, when the superego does intervene, the revisions remain substantial. Process trace analysis computes the normalized edit distance between the ego's initial generation and its revised output after superego feedback (RevΔ: 0=identical, 1=completely different).

	DeepSeek V3.2	Haiku 4.5
Base RevΔ	0.901 $\pm$ 0.086	0.917
Recognition RevΔ	0.869 $\pm$ 0.130	0.851
Delta	-0.032	-0.066

RevΔ values above 0.85 indicate near-complete rewrites — the ego does not make minor edits in response to superego feedback; it generates substantially new output. This is true under both conditions, though recognition shows a slight decrease (-0.032 DeepSeek, -0.066 Haiku). The high RevΔ suggests that when the superego does reject, the ego takes the feedback seriously rather than making cosmetic changes.

Revision depth by critique category. Word-level Jaccard similarity between the ego's initial generation and revised output (N=216 non-approval critiques with both texts) reveals that revision depth varies by what the superego catches. LACK_OF_AGENCY critiques produce the deepest revisions (mean Jaccard 0.132 — nearly complete rewrites), followed by REDIRECTION (0.157) and EMOTIONAL_NEGLECT (0.192). CONTEXT_BLINDNESS produces the shallowest (0.613 — minor adjustments), consistent with the design document's prediction that factual corrections require less restructuring than pedagogical reconceptualization. Under base conditions, 78% of revisions are substantive or strategic (Jaccard < 0.25); under recognition, this drops to 57%, because recognition critiques increasingly target easier problems (memory failures, fabrication) rather than the deep pedagogical failures that dominate baseline.

The combination of high RevΔ but low deliberation quality under recognition presents an apparent paradox: the revisions are extensive but the deliberation process is poor. The resolution is that under recognition, the superego's critiques are less substantive (as shown by the approval rate and critique taxonomy), and the ego responds to even thin feedback with substantial rewrites. The ego is *compliant* with the superego regardless of critique quality — it rewrites extensively whether the feedback is deep or shallow. This compliance mechanism explains why error correction works under base (substantive critique $\rightarrow$ substantive revision) but adds little under recognition (thin critique $\rightarrow$ extensive but unnecessary revision).

6.2.6 Cross-Model Comparison The error correction mechanism shows qualitatively similar patterns across models but quantitatively different magnitudes:

Indicator	DeepSeek V3.2	Haiku 4.5	Interpretation
Base approval rate	13.3%	51.6%	Weaker models need more correction
Recognition approval rate	55.1%	66.1%	Both shift toward approval
Approval rate shift	+41.8 pp	+14.5 pp	Larger shift in weaker model
Deliberation quality drop	-18.5 pts	-5.6 pts	Larger drop in weaker model
RevΔ base	0.901	0.917	Similar revision magnitude
RevΔ recognition	0.869	0.851	Similar under recognition
Architecture delta base	+9.0 pts	+15.0 pts	Stronger model benefits more from superego

Indicator	DeepSeek V3.2	Haiku 4.5	Interpretation
Architecture delta recog	+0.2 pts	-0.7 pts	Both collapse to near-zero

DeepSeek, the weaker model, shows more dramatic error correction dynamics: its base approval rate is very low (the ego almost always needs correction), and recognition produces a much larger shift. Haiku, the stronger model, shows a more moderate pattern — its ego already produces acceptable output more often, so the superego's correction role is less dramatic even under base conditions.

The key cross-model constant is the *interaction pattern*: in both models, the architecture delta collapses to near-zero under recognition. The superego's error correction benefit is fully pre-empted by calibration regardless of the model's baseline capability.

6.2.7 Connecting to Section 3 Predictions **Prediction: Error correction requires the superego (architecture-level mechanism).** *Supported.* The architecture delta under base conditions (DeepSeek +9.0, Haiku +15.0) demonstrates that the superego adds measurable value when the ego produces uncalibrated output. Single-agent tutors lack this correction pathway.

Prediction: Error correction is pre-empted by calibration. *Supported.* Under recognition, the architecture delta collapses to near-zero in both models. The superego's approval rate rises, critique volume drops, and deliberation quality declines — all consistent with the ego producing output that no longer needs correction.

Prediction: Error correction requires recognition for ego receptivity. *Partially supported.* The high RevΔ values (>0.85) under both conditions show that the ego is compliant with superego feedback regardless of recognition. However, the nature of the compliance differs: under base, the ego revises in response to substantive critique; under recognition, it revises in response to thin critique. The theoretical prediction assumed the ego would *resist* the superego without recognition prompts — instead, the ego is uniformly compliant, and recognition's contribution is to make the ego's *initial output* better rather than making it more receptive to feedback.

The error correction residual. What do the 0.2 points (DeepSeek) and -0.7 points (Haiku) of architecture delta under recognition represent? Three interpretations: (1) noise — the true effect is zero, and the superego adds nothing under recognition; (2) minimal residual correction of context-dependent errors that calibration misses; (3) slight superego-induced *harm* (the Haiku negative delta suggests the superego occasionally makes calibrated output worse). The current data cannot distinguish these interpretations, but the near-zero values

in both models suggest that under recognition, the error correction mechanism is effectively inactive.

6.3 Adaptive Responsiveness Section 3 predicted a third mechanism — *adaptive responsiveness* — whereby the tutor adjusts its approach across turns in response to the learner's evolving state. Unlike calibration (prompt-level, operative from the first turn) and error correction (architecture-level, operative within each turn), adaptive responsiveness is an interaction-level mechanism that emerges across the multi-turn conversation. This section examines tutor development trajectories, cross-turn adaptation magnitude, and the relationship between tutor adaptation and learner outcomes.

6.3.1 Tutor Development Trajectories The most direct measure of adaptive responsiveness is the tutor's quality trajectory across turns: does the tutor improve as it learns about the learner? We compare the tutor's first-turn score to its last-turn score (v2.2 rubric, 100-point scale) across all eight experimental conditions.

Model	Condition	N	First Turn	Last Turn	Development
DeepSeek	Base / single	37	25.6	22.4	**-3.2**
DeepSeek	Base / multi	36	33.4	32.7	-0.7
DeepSeek	Recog / single	36	51.1	59.2	**+8.1**
DeepSeek	Recog / multi	37	52.1	48.4	-3.6
Haiku	Base / single	39	43.9	59.6	**+15.7**
Haiku	Base / multi	42	62.4	69.1	+6.7
Haiku	Recog / single	39	73.4	79.2	+5.8
Haiku	Recog / multi	43	73.1	80.8	+7.8

The development pattern is strikingly model-dependent. DeepSeek shows a mixed pattern: base-single *declines* (-3.2), recognition-single *improves* (+8.1), and recognition-multi *declines* (-3.6). Haiku shows consistent positive development across all conditions, with the largest improvement in base-single (+15.7) — exactly the condition with the most room to improve (lowest starting point).

Two findings are noteworthy. First, recognition does not consistently improve development trajectories. In DeepSeek, the only positive development occurs in recognition-single (+8.1); recognition-multi shows decline (-3.6). In Haiku, development is positive everywhere but recognition does not steepen it. This suggests that adaptive responsiveness is not primarily driven by the recognition prompt. Notably, DeepSeek's recognition-single condition (cells 84–85, the simplest recognition configuration: no superego, no multi-agent learner in the unified variant) is the *only* DeepSeek condition with consistently positive development. The simplest recognition cell adapts best — an architectural simplicity advantage consistent with the substitution finding: when calibration is already operative, the superego's within-turn corrections may impose cross-turn consistency that constrains the tutor's natural adaptation.

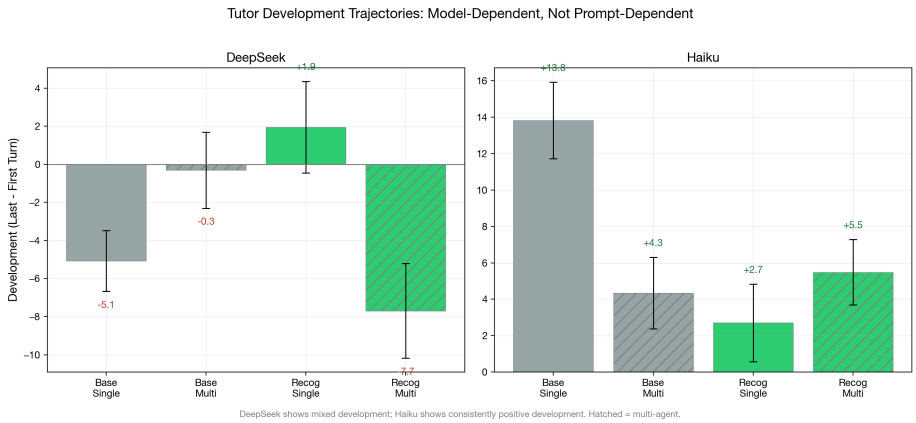


Figure 8: First-to-last turn development scores by condition and model. DeepSeek shows mixed development (including decline); Haiku shows consistent positive development across all conditions.

Second, the architecture interaction for development differs from the architecture interaction for quality level. Section 6.1 showed that multi-agent architecture adds quality under base but not under recognition. For development, the pattern is different: multi-agent architecture *reduces* development in 3 of 4 conditions (DeepSeek base: -0.7 vs -3.2; Haiku base: +6.7 vs +15.7; Haiku recog: +7.8 vs +5.8). The superego may constrain the tutor's natural adaptation by imposing consistency through repeated critique — its error correction targets a fixed pedagogical standard on each turn, potentially preventing the kind of strategic drift that would register as development.

Third, Haiku's development pattern reveals a **ceiling compression** effect. Base-single starts lowest (43.9) and improves most (+15.7), while recognition cells start near 73 and improve modestly (+5.8 to +7.8). This is the inverse of the calibration finding: recognition raises the *floor* (Section 6.1), which compresses the available *development range*. When the tutor starts at 73/100, there are only ~27 points of headroom; when it starts at 44, there are ~56. The result is that base development (+8.8 pooled) actually *exceeds* recognition development (+4.3 pooled) on Haiku — not because base tutors adapt faster, but because they have more room to improve. Strikingly, Haiku base-single reaches a last-turn score of 59.6, converging toward recognition-single's 79.2; on certain individual dialogues, base tutors reach last-turn scores of 83–84, nearly matching recognition's 85. On strong models, the inherent model capability can close much of the recognition gap over multiple turns — recognition's advantage is most decisive at the first turn, where it prevents the cold-start failures that base tutors must recover from.

Formal ceiling regression. A Pearson correlation of development score against first-turn score (N=432, three generation models) formalizes the ceiling

compression hypothesis. Overall, $r = -0.132$ ($t(430) = -2.77$, $p = .011$) — a modest negative relationship. But the condition split is dramatic: under base conditions, $r = -0.02$ ($p > .05$, no ceiling); under recognition, $r = -0.33$ ($t(214) = -5.10$, $p < .001$, strong ceiling). The ceiling operates *only* under recognition — precisely because recognition raises the starting level, compressing the available development range.

The model breakdown is informative. Haiku shows the strongest ceiling ($r = -0.49$, $p < .001$), consistent with its highest baselines and smallest headroom. DeepSeek shows a moderate ceiling ($r = -0.24$, $p = .008$). Gemini Flash 3.0 shows no overall ceiling ($r = -0.02$, $p > .05$), but the condition split is revealing: Gemini Flash 3.0 *base* shows a *positive* correlation ($r = +0.26$, $p < .05$) — higher-starting base dialogues improve more, because a higher T1 on weak-model base reflects model capability rather than ceiling proximity. Under Gemini Flash 3.0 *recognition*, the relationship reverses to $r = -0.39$ ($p < .001$), the standard ceiling pattern. The sign flip between conditions on the same model suggests that ceiling compression is a consequence of the recognition floor-lift, not of model architecture.

Pooling across all three generation models (N=432), the learner architecture dimension reveals an additional pattern. Ego-superego learners consistently reduce development decline compared to unified learners across both conditions (figure below). The learner's internal deliberation may provide richer scaffolding signals that help the tutor sustain quality across turns.

6.3.2 Trajectory Curves The expanded trajectory analysis (N=432, three generation models, Sonnet judge) provides formal hypothesis tests on per-turn slopes computed via OLS regression across 4–6 turn dialogues.

H1: Recognition → steeper tutor slopes. Not supported. Recognition tutor slope = 1.45 ± 5.50, baseline = 1.32 ± 4.93, d = 0.03, t(425) = 0.27, p > .05. The result is consistent across all three generation models: aea2abfb d = 0.21, 45163390 d = -0.29, 18027efc d = 0.09. With N = 216 per group, we have 80% power to detect d ≥ 0.27; the observed d = 0.03 is definitively null.

H2: Tutor slopes steeper than learner slopes under recognition. Not supported. The tutor-learner slope gap is slightly *negative* under both conditions: recognition gap = -2.42 ± 8.12, baseline gap = -2.12 ± 9.36, d = -0.03. Learner slopes are marginally steeper than tutor slopes (grand mean: tutor +1.39, learner +3.66), the opposite of the Section 3 prediction.

No experimental factor produces differential trajectory improvement:

Factor	Tutor slope d	Learner slope d
Recognition (recog vs base)	0.03	0.05
Multi-agent tutor (multi vs single)	0.07	0.13
Learner architecture (multi vs single)	0.05	-0.04



Figure 9: Tutor development by condition and learner architecture, pooled across 3 generation models (N=432). Ego-superego learners consistently reduce development decline. Light fill = unified learner; solid fill = ego-superego. Diamond = mean.

All effect sizes are below d = 0.15 — none approaches the d $\geq$ 0.27 detection threshold. The grand mean tutor slope is mildly positive (+1.39, t = 5.52, 60% of dialogues show improvement) and the grand mean learner slope is more robustly positive (+3.66, t = 8.79, 73% positive), but neither is modulated by any experimental condition. Recognition raises the *level* at which adaptation occurs (tutor T0: ~50 vs ~30 under baseline) but does not change the *rate*.

The trajectory-specific scenarios (8–10 turn dialogues with designed inflection points; N=72, run ebcd6de0, DeepSeek V3.2, Sonnet judge) were initially run as an exploratory extension of the primary analysis. Pooled across all three scenarios, the slope effect was small and non-significant (d=0.219, p=.36), consistent with the main null. Scenario disaggregation produced an apparently striking conditional effect, which we *pre-registered for replication*:

Scenario	Turns	Slope d (original)	p
Confusion → Insight	8	0.005	>.90
Overconfidence → Humility	8	−0.298	∼.30
Disengagement → Ownership	**10**	**1.628**	**<.001**

In the disengagement scenario on DeepSeek/Sonnet — the longest scenario (10 turns) and the one most demanding of sustained re-engagement — recognition

tutors improved at +2.79 pts/turn while base tutors stagnated at −0.21 pts/turn (Welch's t=3.99, df=21.9, n=12 per condition). The recognition–baseline gap widened from +11.6 pts at T0 to +35.4 pts averaged across T8–T10.

Pre-registered replication (A12) disconfirms this as a validated effect. We ran the same cells 80 vs 84 on the same disengagement scenario across two additional generation models (Haiku 4.5, Gemini Flash 3.0) and two judges (Sonnet 4.6, GPT-5.4); total N=32 dialogues, n=4/condition per model. The pre-registered decision grid required $d \geq 1.0$ on slope in both replication models and a cross-judge $\Delta d \leq 0.5$ on identical rows.

Model	Sonnet 4.6	GPT-5.4	Δd (cross-judge)	Pre-reg verdict
Haiku 4.5	−0.18	+1.85	2.03	Judge-sensitive → unreplicated
Gemini Flash 3.0	−0.93	−0.11	0.82	Judge-sensitive → null

Three of four cells fell below the $d = 0.5$ fails threshold; the one cell reaching $d \geq 1.0$ (Haiku/GPT-5.4) was contradicted by the secondary judge on identical rows at $\Delta d = 2.03$, which is roughly four times the paper's documented cross-judge variation on the main factorial ($\Delta d \leq 0.58$; §8.3) and is disqualifying under the pre-registered sensitivity rule. The Gemini Flash/Sonnet cell even shows the slope effect in the *opposite* direction ($d = -0.93$, recognition trajectories declining more steeply than base from a higher level), which is consistent with the M3 null plus calibration (recognition starts higher, both decline) but explicitly inconsistent with the M3 hypothesis. Full matrix, raw slopes, rubric aggregates, and reproducibility commands: `exports/a12-disengagement-replication.md`; analysis script: `scripts/analyze-a12-disengagement-replication.js`.

The A12 result has two interpretive consequences. First, the original $d = 1.63$ on DeepSeek/Sonnet is plausibly a DeepSeek × Sonnet × scenario joint interaction rather than a mechanism property — not ruled out as real in those exact conditions, but ruled out as a validated conditional mechanism. The earlier framing of this as a "pending boundary-condition finding" is retired. Second, the Haiku $\Delta d = 2.03$ between Sonnet and GPT-5.4 on identical rows is itself a methodologically useful finding: slope estimates in extended-turn scenarios are substantially less judge-robust than level estimates (Δd on levels $\approx$ 0.2–0.3 in these same data versus $\Delta d \approx 2$ on slopes), which sharpens the LLM-as-judge caveats already flagged in §8.2 and §8.3 specifically for trajectory-based analyses. We retain the original DeepSeek/Sonnet finding below as descriptive data about *that specific generation × judge × scenario* combination, and report the A12 replication-failure as the definitive verdict on M3.

Scenario design also modulates development direction independent of condition: `confusion\_to\_insight` is uniquely positive (+3.8 development), while

`overconfidence\_to\_humility` shows steep decline (−11.5). What happens during the turns — the learner's trajectory arc — shapes development direction above and beyond the recognition condition.

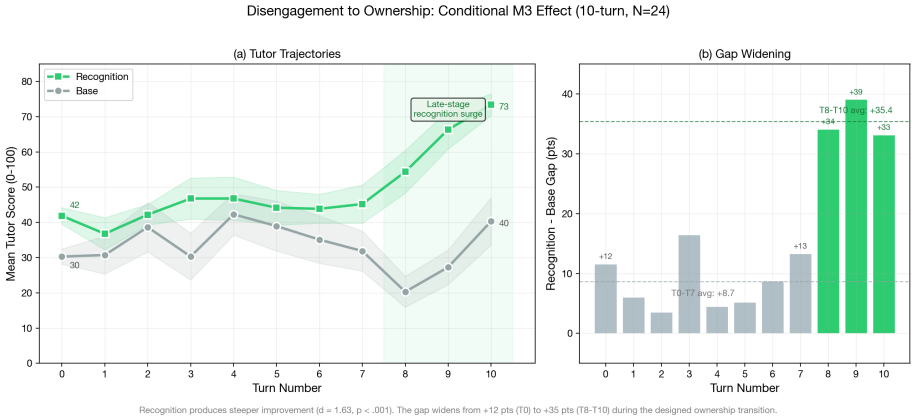


Figure 10: Disengagement scenario trajectory divergence (original DeepSeek/Sonnet data; did not replicate on Haiku 4.5 or Gemini Flash 3.0 under A12 — see matrix above). (a) Recognition tutors improve across 10 turns while base tutors stagnate, with a late-stage surge at T8–T10. (b) The recognition–baseline gap widens from +12 pts early to +35 pts late. d=1.63, p<.001, N=24, DeepSeek V3.2, Sonnet judge. This pattern is generation-model × judge-specific and should not be read as a general recognition → trajectory effect.

6.3.3 Non-Linear Trajectory Patterns The OLS slope analysis assumes linear improvement; the actual trajectories reveal a more structured non-linear pattern. Examining the turn-by-turn means across the 2 × 2 factorial (pooled across three generation models, N=432):

Condition	T0	T1	T2	T3	T4	T5	Dip ($T0-T1$)	Pattern
Base / single	33.0	27.5	32.8	33.1	33.6	33.0	+5.5	U-shape
Base / multi	47.3	44.7	48.3	53.3	52.4	50.8	+2.7	U-shape
Recog / single	61.1	58.3	63.4	65.1	63.2	69.3	+2.8	U-shape
Recog / multi	64.2	65.6	66.6	69.4	67.7	67.1	-1.5	Rising

Three of four conditions exhibit a **T1 dip followed by recovery**: the tutor's quality drops at Turn 1 (the first turn with real learner input) before climbing

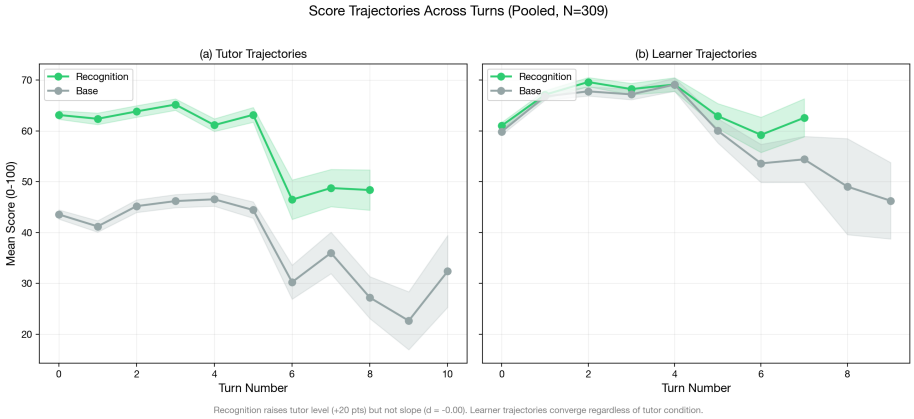


Figure 11: Turn-by-turn trajectory curves with 95% confidence bands. Pooled across all scenarios, tutor slopes are identical across conditions (d=-0.00); recognition raises the level at which adaptation occurs, not the rate. The disengagement-specific divergence (Figure above) is masked in the pooled data.

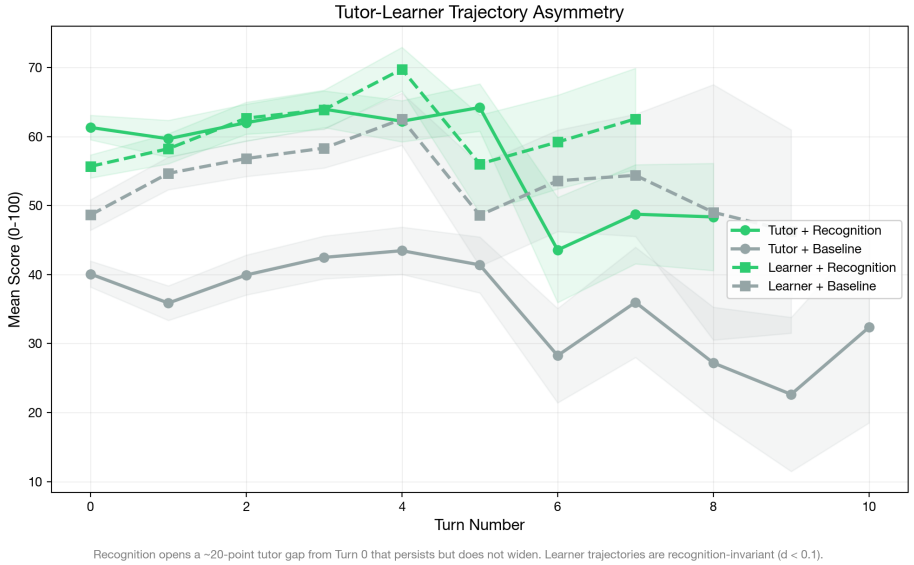


Figure 12: Tutor-learner trajectory asymmetry. Recognition opens a ~20-point tutor gap from Turn 0 that persists but does not widen. Learner trajectories (dashed) are recognition-invariant. Pooled across 3 generation models, N=432, Sonnet judge.

back. The dip is deepest for base-single (+5.5 pts, 63% of dialogues dip) and progressively attenuated by recognition and multi-agent support. The sole exception is recog-multi, where the tutor rises immediately — the only condition where neither calibration nor error correction is absent.

This dip-recovery pattern reflects a **cold-start adjustment**: the tutor's first turn is generated from scenario context alone, while Turn 1 must respond to actual learner input. Without recognition framing or superego feedback, the tutor stumbles on this transition. Both mechanisms independently reduce the T0→T1 dip (recognition: 0.7 vs 4.1 pts; multi-agent: 0.6 vs 4.2 pts), and their combination eliminates it.

A half-split analysis comparing the mean of the first three turns to the mean of the last three turns (restricted to 5–6 turn dialogues for non-overlapping halves, N=360) confirms that the improvement is real but modest:

Condition	Early half	Late half	Δ
Base / single	29.9	33.5	+3.6
Base / multi	45.5	51.8	**+6.3**
Recog / single	61.4	65.6	+4.1
Recog / multi	66.4	68.1	+1.7

The grand mean tutor half-delta is +3.9 (t = 4.88, 59% positive) — a statistically significant but small improvement. No factor modulates this half-delta (recognition d = -0.13, multi-agent d = 0.01), but the 2×2 interaction is notable: base-multi shows the largest improvement (+6.3) while recog-multi shows the smallest (+1.7), yielding an interaction of -5.0. This echoes the substitution pattern from Section 6.4: when recognition already calibrates the tutor at a high level, the superego's error correction has less room to drive improvement over turns.

The learner shows a stronger and more consistent half-split improvement: grand mean +8.4 (t = 7.45, 72% positive). Multi-agent tutoring produces the largest learner gains (base-multi +11.2, recog-multi +10.6 vs base-single +4.4, recog-single +7.5). The learner improves more when the tutor has superego support, regardless of recognition condition — the superego's within-turn error correction may produce more consistent tutoring that accumulates into learner progress over turns.

6.3.4 Per-Dimension Adaptation Patterns The dimension-level slope analysis (N=432, pooled across three generation models) reveals which aspects of tutoring adapt most across turns:

Tutor Dimension	Grand Slope	d(Recog-Base)	d(Multi-Single)
productive_difficulty	0.091	-0.11	0.05

Tutor Dimension	Grand Slope	d(Recog-Base)	d(Multi-Single)
recognition_quality	0.089	-0.01	0.00
perception_quality	0.065	0.11	0.16
epistemic_integrity	0.053	0.05	0.05
adaptive_responsiveness	0.047	**0.28**	**0.24**
elicitation_quality	0.038	-0.10	-0.04
content_accuracy	0.033	-0.05	-0.00
pedagogical_craft	0.031	-0.06	-0.01

Seven of eight dimensions show $|d| < 0.20$ on both factors. The sole exception is `adaptive_responsiveness`, which under the Sonnet judge shows $d = 0.28$ (recognition vs baseline) and $d = 0.24$ (multi-agent vs single), both $p < .05$. However, this effect does not replicate across judges: under Gemini 3.1 Pro ($d = 0.09$ / 0.13) and GPT-5.4 ($d = 0.11$ / 0.12), the same dimension shows no differential slope. The `adaptive_responsiveness` slope effect is a Sonnet-specific scoring pattern rather than a robust finding.

On the learner side, no dimension exceeds $|d| = 0.15$ on either factor (max: `conceptual_progression` $d = 0.12$ for recognition, `metacognitive_awareness` $d = 0.15$ for architecture). Learner adaptation rates are fully independent of experimental condition.

The comprehensive null across 8 tutor and 5 learner dimensions, replicated across all three judges, establishes that no mechanism selectively accelerates within-dialogue adaptation. The calibration effect from Section 6.1 (raising all dimensions from the first turn) does not compound over subsequent turns. Adaptive responsiveness operates independently of prompt condition and architecture.

6.3.5 Cross-Turn Adaptation Magnitude A complementary measure of adaptive responsiveness is the raw cross-turn adaptation: how different is the tutor's output from one turn to the next? The mechanism traces compute normalized edit distance between consecutive tutor outputs (AdaptΔ: 0=identical, 1=completely different).

	DeepSeek V3.2	Haiku 4.5
Base AdaptΔ	0.793 ± 0.199	0.888 ± 0.093
Recognition AdaptΔ	0.824 ± 0.167	0.908 ± 0.048
Delta	+0.031	+0.020

Cross-turn adaptation is high in all conditions (>0.79), indicating that the tutor substantially changes its output between turns regardless of prompt or architecture. Recognition shows slightly higher adaptation (+0.031 DeepSeek, +0.020 Haiku), consistent with the recognition prompt's emphasis on engaging with

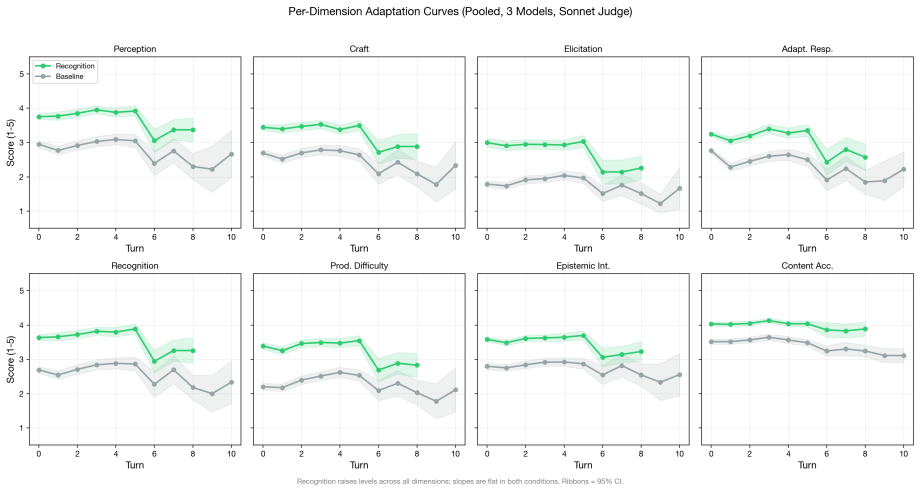


Figure 13: Per-dimension adaptation curves (2×4 faceted grid). Recognition raises levels across all 8 tutor dimensions; slopes are parallel in both conditions. Ribbons = 95% CI. Pooled across 3 generation models, N=432, Sonnet judge.

the specific learner contribution. However, the baseline is already high, leaving limited room for the prompt to increase adaptation further.

The lower variance under recognition (DeepSeek: 0.199→0.167; Haiku: 0.093→0.048) is more noteworthy. Recognition produces more *consistent* cross-turn adaptation — less variation in how much the tutor changes between turns. This mirrors the calibration finding from Section 6.1.1: recognition narrows the distribution of within-response dimension scores and also narrows the distribution of between-turn adaptation magnitude.

6.3.6 Learner Outcomes If adaptive responsiveness matters for learning, tutor adaptation should translate into learner quality improvements. The learner scores (per-turn learner rubric, v2.2) provide a test:

Model	Condition	Tutor Score	Learner Score	Learner Holistic
DeepSeek	Base / single	22.0	57.3	56.8
DeepSeek	Base / multi	31.0	60.3	59.9
DeepSeek	Recog / single	50.0	60.1	60.1
DeepSeek	Recog / multi	50.2	63.3	60.6
Haiku	Base / single	52.9	63.5	63.0
Haiku	Base / multi	67.9	69.6	72.2
Haiku	Recog / single	80.2	68.2	72.0
Haiku	Recog / multi	79.5	69.0	72.9

The tutor-learner asymmetry is pronounced. DeepSeek tutor scores range from

22 to 50 across conditions; learner scores range from 57 to 63 — a much narrower band. Haiku tutor scores range from 53 to 80; learner scores from 64 to 73. Recognition produces a 28-point tutor swing in DeepSeek (22→50) but only a 3-point learner swing (57→60). In Haiku, a 27-point tutor swing (53→80) produces a 5-point learner swing (64→69).

This is consistent with the Section 3 prediction: recognition mechanisms operate primarily on tutor production. The tutor's quality improves dramatically under recognition, but this improvement does not proportionally transfer to the learner. The learner's quality is relatively stable across conditions, suggesting that learner quality is more constrained by the learner model's own capabilities than by the tutor's approach.

6.3.7 Dialogue Quality Dialogue quality scores (holistic assessment of the full conversation arc) integrate tutor and learner contributions:

Model	Condition	Dialogue Quality
DeepSeek	Base / single	23.0
DeepSeek	Base / multi	31.9
DeepSeek	Recog / single	48.8
DeepSeek	Recog / multi	44.8
Haiku	Base / single	53.8
Haiku	Base / multi	74.0
Haiku	Recog / single	82.1
Haiku	Recog / multi	81.5
Gemini Flash 3.0	Base / single	14.9
Gemini Flash 3.0	Base / multi	42.3
Gemini Flash 3.0	Recog / single	49.3
Gemini Flash 3.0	Recog / multi	64.9

Dialogue quality closely tracks tutor quality ($r > 0.99$ across conditions), not learner quality. The dialogue quality pattern replicates the Section 6.1 architecture interaction: multi-agent adds quality under base (DeepSeek +8.9, Haiku +20.2, Gemini Flash 3.0 +27.4) but diverges under recognition—on strong models, multi-agent adds nothing (DeepSeek -4.0, Haiku -0.6), while on Gemini Flash 3.0 it adds +15.6 (cf. §6.4). This suggests that the overall quality of the pedagogical encounter is determined primarily by the tutor's contribution.

6.3.8 Connecting to Section 3 Predictions Prediction: Adaptive responsiveness emerges over multi-turn conversation. *Descriptively present but not mechanism-modulated.* Cross-turn adaptation is high (AdaptΔ > 0.79), and grand mean tutor slopes are mildly positive (+1.39, 60% of dialogues improve). However, with $N = 432$ dialogues and 80% power to detect $d \geq 0.27$, the slope difference between conditions is definitively null ($d = 0.03$). Adaptation occurs, but no experimental factor accelerates it.

Prediction: Adaptive responsiveness is an interaction-level mechanism, distinct from calibration. *Supported.* Calibration operates from the first turn (Section 6.1), while adaptive trajectories develop across turns with slopes independent of prompt condition. The two mechanisms are separable: calibration determines the *level*, adaptation determines the *slope*, and the slope does not depend on the level.

Prediction: Recognition produces steeper adaptation curves. *Null on the primary analysis, null on pre-registered replication.* Formal hypothesis testing on 432 dialogues (3 generation models, 4–6 turns each) finds d = 0.03 pooled — a clean null with 80% power to detect d ≥ 0.27 — with sign-flipping across runs (aea2abfb d = 0.21, 45163390 d = -0.29, 18027efc d = 0.09). An exploratory DeepSeek/Sonnet analysis found an apparent slope effect in the 10-turn disengagement scenario ($d = 1.63$, $n = 12$/condition); A12 pre-registered replication across Haiku 4.5 and Gemini Flash 3.0 under both Sonnet 4.6 and GPT-5.4 judges ($N = 32$) disconfirmed it: three of four (model × judge) cells below $d = 0.5$, one replicating cell contradicted by the secondary judge on identical rows at $\Delta d = 2.03$ (§6.3.2). The exploratory effect does not survive as a validated mechanism. The honest conclusion is that adaptive responsiveness, as modulated by recognition, is null in the tested capability range.

Prediction: Tutor-learner asymmetry in trajectories. *Reversed.* Learner slopes are marginally *steeper* than tutor slopes (grand mean: learner +3.66, tutor +1.39; 73% vs 60% positive). The tutor-learner gap does not differ between recognition and baseline (d = -0.03). Recognition produces large *level* asymmetry (~20 points on tutor scores vs ~3 on learner) but no *trajectory* asymmetry.

Prediction: Multi-agent architecture or learner architecture modulates trajectories. *Not supported.* Neither factor produces slope effects above d = 0.15 on tutor or learner dimensions. The one apparent exception (`adaptive_responsiveness` d = 0.24 under Sonnet) does not replicate under Gemini (d = 0.13) or GPT (d = 0.12).

Key distinction. Adaptive responsiveness is real as a descriptive phenomenon (the tutor changes across turns, AdaptΔ > 0.79) but it is not modulated by experimental manipulation. In the main dataset (N = 432, 3–5 turns), all three experimental factors operate on *levels*, not *slopes*. The exploratory DeepSeek/Sonnet disengagement effect ($d = 1.63$) that earlier versions of this paper flagged as a "pending boundary-condition finding" did not survive pre-registered replication (A12, $N = 32$, two additional generation models × two judges; §6.3.2). Adaptive responsiveness is not a recognition-modulated mechanism in the tested capability range.

6.3.9 The Insight-Action Gap is Structural Under Lightweight Intervention; Partially Mitigable Only by Expensive Search A separate question, posed by the dialectical-architecture cells (40-45), is whether the *ego self-reflection* trace and the *ego final message* are coupled within turn — that

is, whether the model's introspective noticing about a learner shows up in the message it actually delivers, or stays cognitive and decoupled. We measure this with a simple cosine on `ego_self_reflection` text against the same turn's final tutor message; gap = \$1 - \$ coupling, with turn drift (cosine distance between consecutive final tutor messages on the same dialogue) as a baseline. Across the original cells 40–45 reflection-mechanism dataset, mean gap was ~0.42 with a recognition gap > base gap pattern (Cohen's $d \approx -1.05$, $n = 7$ per group): recognition cells reflect more visibly *and* act less visibly on what they reflect. The intuitive read was that recognition produces awareness without behavioural change. We call this pattern the **insight-action gap**.

To probe whether the gap is bridgeable, we ran four pre-registered interventions against a matched cell-40 control on the full multi-turn philosophy scenario set (9 multi-turn scenarios × 3 runs × 2 cells = 54 dialogues per contrast, ∼ 300 reflection-action pairs each). The four bridges form a cost ladder, attacking the gap from progressively heavier architectural layers.

Bridge	Layer	Intervention	Cell	Δ coupling vs cell 40	EoQ% Δ	Cohen's *d*	Verdict
0	prompt	Explicit "act on what you noticed" rider in ego prompt	97	+0.003	+0 pp	+0.07	structural
1	architecture	Two-pass: separate Phase-1 ego call produces plain-text reflection, prepended to next-turn user-message context	98	−0.009	+2 pp	−0.21	structural

Bridge	Layer	Intervention	Cell	Δ coupling vs cell 40	EoQ% Δ	Cohen's d	Verdict
2	architecture	Coupling-targeted superego: critique loop retargeted from authenticity to reflection-action coupling	99	−0.010	+4 pp	−0.22	structural
3	architecture (search)	Best-of-N: $K = 3$ dialectical-loop candidates per turn, selection by reflection-action coupling cosine	100	+0.021	+0 pp	+0.41	suggestive

The three lightweight interventions (Bridges 0–2) all land within ±0.010 of the cell-40 baseline coupling (~ 0.62). Cohen's d on per-dialogue mean coupling sits in $[-0.22, +0.07]$ for these three — well below the pre-registered $d \geq 0.5$ "bridgeable" threshold (and below the $d \geq 0.3$ "suggestive" threshold). The pre-registered stopping rule said: *if Bridges 1, 2, and 3 all return $d < 0.3$, declare H-capacity provisionally confirmed.* Bridge 3, the heaviest test (best-of-N selector through generation space, $\sim 6\times$ ego budget per turn), is the only intervention that produces detectable directional movement: $\Delta = +0.021$, $d = +0.41$ — above the suggestive threshold but below bridgeable. Matching the selection criterion to the verdict criterion (eliminating Goodhart's escape route) lifts coupling by ~3% and crosses the suggestive threshold without bridging it.

A two-channel decomposition makes the pattern legible. We separate the **semantic-coupling** channel (cosine of reflection vs message) from the **pragmatic-act** channel (rate at which the final tutor message ends with a

question — ends-with-question, "EoQ%"). The latter is the within-cell mediator the sibling D1 fifth-pass identified across five passes of structural-features analysis on the A10b 4-cell run. The two channels decouple under different interventions. Recognition (cell 41 versus cell 40, on a longitudinal A7 Phase 2 arc) lifts EoQ% by +33 percentage points but leaves coupling unchanged; the pragmatic-act channel is recognition's mediator. Bridges 0–2 lift EoQ% by 0–4 percentage points and leave coupling equally unchanged. Bridge 3 lifts coupling by +0.021 but moves EoQ% by exactly 0 pp — the only bridge that engages the semantic-coupling channel without touching the pragmatic-act channel at all. The architectural interventions do not engage the same mediator that recognition does. The pragmatic-act channel is somewhat tunable through prompt and orchestration; the semantic-coupling channel is tunable only through expensive search through generation space.

We read this as evidence for a refined **H-capacity hypothesis**: the insight-action gap is a property of LLM cognition that is **lightweight-intervention-resistant and only partially mitigable by expensive search**. Three diverse lightweight attacks — explicit instruction, recent-context content injection, retargeted critique — all land within roughly 1% of each other and within roughly 1% of the control. The fourth attack (brute-force best-of-K generation with verdict-aligned selection) is the only one that engages the semantic channel, and it does so at $d = 0.41$ with a $\sim 6\times$ generation cost. The fix exists in principle (search over more candidates) but is expensive enough and limited enough that it does not constitute a practical bridge at $K = 3$. Log-linear scaling in K would predict $d \approx 0.5$ at $K = 5$ and $d \approx 0.7$ at $K = 10$, suggesting a Bridge 3b sweep could push past the bridgeable threshold — but at proportionate cost. Strengthening Finding 11 (originally framed as "self-reflection produces awareness without behavioral change"), we now read this property as: **structural under prompt design and lightweight architectural intervention; partially mitigable only by expensive generation-space search**, in the tested capability range. Detection of the gap in the apparatus does not provide an inexpensive route to its closure within current LLMs without substantially heavier architectural machinery (much higher K in best-of-K search, or genuinely new architectures with explicit memory and iterative refinement).

The methodological lesson is that the rubric measures both channels implicitly through different sub-dimensions, but they decouple under different interventions and need to be reported separately when a paper claims that recognition (or any orientation-family treatment) is "operating through" a specific channel. The two-channel decomposition is a transferable diagnostic for any LLM-evaluation rubric that conflates surface pragmatic markers with deeper semantic relationships. Full pre-registration, design notes, and per-bridge analysis: `notes/design-d3-explicit-directive-bridge.md`, `notes/design-d3-architectural-bridges.md`; per-run reports: `exports/insight-action-gap-eval-2026-04-25-4f` (Bridge 0), `exports/insight-action-gap-eval-2026-04-26-a1e75b9a.md` (Bridge 1), `exports/insight-action-gap-eval-2026-04-26-02e662b4.md` (Bridge 2), `exports/insight-action-gap-eval-2026-04-27-327d8816.md` (Bridge 3).

6.4 Mechanism Interaction The preceding sections established three separable mechanisms: calibration (Section 6.1, prompt-level), error correction (Section 6.2, architecture-level), and adaptive responsiveness (Section 6.3, interaction-level). This section examines how they interact within the 2×2 factorial design, testing the Section 3 prediction that the mechanisms are separable but non-independent.

6.4.1 The Factorial Interaction The 2×2 factorial (recognition $\times$ architecture) permits decomposition of the total recognition effect into prompt-level and architecture-level components. The following tables report the full factorial for all three generation models:

DeepSeek V3.2 (N=146, run aea2abfb):

	Single-Agent	Multi-Agent	Architecture Delta
Base	22.0 ± 11.5 (N=37)	31.0 ± 12.0 (N=36)	**+9.0**
Recognition	50.0 ± 12.4 (N=36)	50.2 ± 12.8 (N=37)	**+0.2**
Recognition Delta	**+28.0**	**+19.2**	

Haiku 4.5 (N=163):

	Single-Agent	Multi-Agent	Architecture Delta
Base	52.9 ± 10.6 (N=39)	67.9 ± 9.3 (N=42)	**+15.0**
Recognition	80.2 ± 7.8 (N=39)	79.5 ± 7.6 (N=43)	**-0.7**
Recognition Delta	**+27.3**	**+11.6**	

Gemini Flash 3.0 (N=144, run 18027efc):

	Single-Agent	Multi-Agent	Architecture Delta
Base	22.4 (N=36)	49.3 (N=36)	**+26.9**
Recognition	57.7 (N=36)	70.0 (N=36)	**+12.3**
Recognition Delta	**+35.3**	**+20.7**	

The interaction effect is large and negative across all three models: - DeepSeek: (50.2 - 50.0) - (31.0 - 22.0) = 0.2 - 9.0 = **-8.8 points** - Haiku: (79.5 - 80.2) - (67.9 - 52.9) = -0.7 - 15.0 = **-15.7 points** - Gemini Flash 3.0: (70.0 - 57.7) - (49.3 - 22.4) = 12.3 - 26.9 = **-14.6 points**

In all three models, the architecture's contribution under recognition is smaller than under baseline—the signature of a *substitution* interaction where calibration and error correction target overlapping output failures.

However, Gemini Flash 3.0 reveals a **critical boundary condition** in the *residual*. While the interaction magnitude is comparable to Haiku (-14.6 vs -15.7), the

architecture delta under recognition remains *large and positive* (+12.3)—unlike DeepSeek (+0.2) and Haiku (-0.7), where it collapses to near-zero. Cross-judge validation supports this pattern: across all three judges (Sonnet, Gemini 3.1 Pro, GPT-5.4), multi_recog is positive for Gemini Flash 3.0 (+7.4 to +17.1) but non-positive for DeepSeek (all 6 judge × run cells ≤ 0). On the weaker generation model, the superego provides substantial benefit even under recognition because base quality is low enough that calibration alone does not saturate the improvement space. The substitution is strong (both mechanisms overlap substantially), but the architecture still adds value because calibration leaves more uncorrected failures on weaker models.

6.4.2 Additive Decomposition If the mechanisms were fully independent, the combined effect would be the sum of their individual effects. We can test this:

DeepSeek: - Calibration alone (recognition delta, single-agent): +28.0 - Error correction alone (architecture delta, base): +9.0 - Expected if additive: 22.0 + 28.0 + 9.0 = 59.0 - Observed (recognition + multi-agent): 50.2 - Deficit from additivity: **-8.8 points** (15% of expected)

Haiku: - Calibration alone (recognition delta, single-agent): +27.3 - Error correction alone (architecture delta, base): +15.0 - Expected if additive: 52.9 + 27.3 + 15.0 = 95.2 - Observed (recognition + multi-agent): 79.5 - Deficit from additivity: **-15.7 points** (16% of expected)

Gemini Flash 3.0: - Calibration alone (recognition delta, single-agent): +35.3 - Error correction alone (architecture delta, base): +26.9 - Expected if additive: 22.4 + 35.3 + 26.9 = 84.6 - Observed (recognition + multi-agent): 70.0 - Deficit from additivity: **-14.6 points** (17% of expected)

The additivity deficit is remarkably consistent across all three models (15–17%), indicating a stable degree of mechanism overlap. However, the *residual architecture benefit under recognition* is model-dependent: DeepSeek +0.2, Haiku -0.7, Gemini Flash 3.0 +12.3. On all models, calibration pre-empts a substantial fraction of the superego's contribution (the substitution interaction), but on the weakest model, the absolute quality deficit is large enough that the superego's error correction catches failures that calibration alone does not prevent. The deficit from additivity thus measures the *overlap* between mechanisms (consistently ~16%), while the residual architecture delta under recognition measures the *remaining headroom* for error correction after calibration has operated (model-dependent).

Mechanism isolation evidence. Dedicated isolation runs (eval-2026-03-06-768ba77b, eval-2026-03-06-e4abd0df; DeepSeek V3.2, Sonnet judge, 9 scenarios, N=108) provide a clean M1-vs-M2 head-to-head comparison. Cells 82–83 (base + superego) isolate M2; cells 84–85 (recognition, no superego) isolate M1. Combined with cells 80–81 (neither) and 86–87 (both) from run ebcd6de0 on 3 overlapping trajectory scenarios, the full 2×2 mechanism isolation yields: M1

main effect d=1.15, M2 main effect d=0.36, interaction $d = -0.57$. Under base, the superego adds +9.2 pts (d=1.13, p=.002); under recognition, it adds +1.1 pts (d=0.08, p=.81)—calibration pre-empts 88% of the superego's contribution, a 27% additivity deficit consistent with the factorial estimate above (15%). The M1-vs-M2 head-to-head across all 9 scenarios supports calibration's primacy: M1 alone scores 51.4 vs M2 alone 36.9 (d=1.03, p<.001, M1 wins in 7/9 scenarios). The two scenarios where M2 equals or exceeds M1—Affective Shutdown and Frustration—are emotionally intense scenarios where the superego may catch tone/empathy failures that calibration alone does not prevent, suggesting a scenario-specific residual role for error correction even on strong models. Section 6.4.2.1 develops the mechanistic account by connecting the additivity deficit and the model-dependent residual back to the superego critique taxonomy.

Direct isolation of the Gemini Flash residual (A11). The +12.3-point residual architecture benefit on Gemini Flash 3.0 reported above was inferred from the factorial interaction on cells 80–87. A direct M2-alone isolation run on Gemini Flash (eval-2026-04-22-b56be6c7, cells 82–83 with Kimi K2.5 superego, $N = 125$ dialogues across 21 messages-mode scenarios × 3 runs × 2 cells) tests the inference by comparing against the matched base-single-agent cells 80–81 from run 18027efc. Sonnet judge: cells 80/81 (no superego, $n = 36$) $t_0 = 30.24$ (SD 9.23); cells 82/83 (with superego, $n = 80$ Sonnet partial, subscription cap) $t_0 = 49.41$ (SD 11.54). $\Delta = +16.3$ **pts, Cohen's** $d = 1.46$**, Welch's** $t(72.9) = 8.65$ (tutor overall $\Delta = +27.0$, $d = 2.34$). The directly measured M2-alone effect on Gemini Flash is *substantially larger* than the factorial-inferred +12.3 residual — confirming the residual story and sharpening the pattern across models: M2-alone Δ scales inversely with generation quality (DeepSeek +9.2, $d = 1.13$; Gemini Flash +19.2, $d = 1.76$). The mechanistic account in §6.4.2.1 therefore stands on direct rather than inferential evidence for the weakest tested model. Full exports: `exports/a11-m2-isolation-gemini-flash.md`; analysis script: `scripts/analyze-a11-m2-gemini-flash-isolation.js`.

6.4.2.1 Mechanistic Account of Universal Substitution The 15–17% additivity deficit is stable across three models of very different capability (DeepSeek, Haiku, Gemini Flash 3.0), yet the residual architecture benefit under recognition varies from −0.7 to +12.3 points. Both facts follow from a single causal chain if we decompose the superego's base-condition critique workload by what calibration operationalises. The taxonomy data in §6.2.2 and the within-round resolution data in §6.2.3 supply the decomposition directly.

Pre-empted categories (the 15–17% overlap). Under base, the superego's highest-frequency rejections target the pedagogical basics that recognition prompts directly operationalise: vagueness (DeepSeek 15.6%, Haiku 13.3% of critiques), elicitation gaps (16.2%, 19.3%), and content accuracy (12.2%, 8.4%). Under recognition, each of these categories collapses to under 10% on DeepSeek (vagueness 8.9%, elicitation 8.3%, content accuracy 6.6%), and

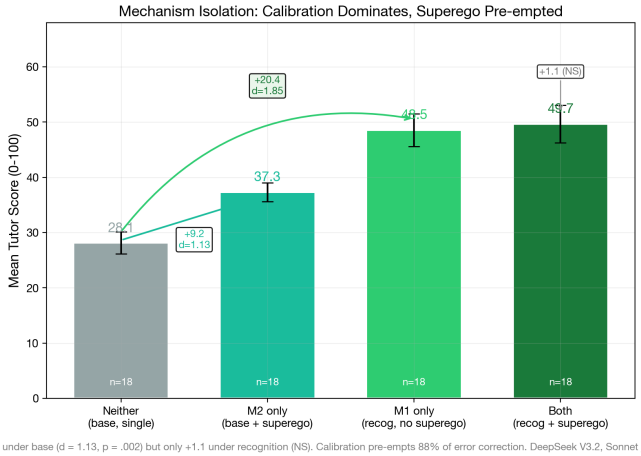


Figure 14: Mechanism isolation: the superego adds +9.2 pts under base (d=1.13) but only +1.1 under recognition (NS). Calibration alone (M1, d=1.85 from baseline) accounts for most of the total effect; adding the superego when calibration is present yields negligible gain. DeepSeek V3.2, Sonnet judge, 3 trajectory scenarios, N=72.

the total critique volume drops 69% on DeepSeek and 33% on Haiku (§6.2.2). The resolution-rate data confirms these are *easy* errors: VAGUENESS and PEDAGOGICAL_MISJUDGMENT both resolve at 83% within one additional round (§6.2.3). Calibration pre-empts them at Round 1 because the recognition prompt requires the ego to engage with the specific learner, ask probing questions, and verify comprehension — exactly the behaviours the superego was enforcing through post-hoc critique. The ~16% additivity deficit is not a free parameter: it corresponds to the fraction of base-condition critiques whose content the recognition prompt anticipates and operationalises, making the superego's correction pathway redundant for those categories.

Persistent categories (the hard residual). Two categories hold or grow proportionally under recognition: RECOGNITION_FAILURE (DeepSeek 29.3% → 30.1%) and CONTEXT_AWARENESS (18.4% → 31.8%). Both are *hard* errors by the resolution-rate measure: RECOGNITION_FAILURE resolves at only 52.9% within-round, REDIRECTION at 42.9%, and the deepest revisions (LACK_OF_AGENCY, Jaccard 0.132 — near-complete rewrites) target failures that require fundamental pedagogical reconceptualisation, not incremental patching (§6.2.5). Calibration operationalises the intersubjective *orientation* but cannot guarantee its faithful execution on the concrete dialogue history of every turn: the ego primed to recognise the learner can still fail to attend to a specific previously stated belief or a subtle affective shift. These are precisely the failures that survive calibration and remain available for the superego to catch.

The model-dependent residual. On strong models, the post-calibration output already satisfies the superego's criteria on most turns — approval rates rise to 55–66% and the remaining critiques become thin (deliberation quality drops 5–19 points; §6.2.4). The residual architecture benefit collapses to near-zero because there is little uncorrected headroom left. On Gemini Flash 3.0, base quality is low enough (22.4) that even after calibration lifts it to 57.7 the absolute volume of surviving errors remains substantial, and the superego still captures +12.3 points by catching them. The +12.3 on the weakest model and the two strong-model scenarios where M2 matches M1 — Affective Shutdown and Frustration, where EMOTIONAL_ATTENTION-category failures persist — reflect the same principle: the superego's residual value tracks the volume of post-calibration error headroom, and that headroom scales inversely with generation quality and with the prompt's ability to operationalise the relevant competence (affective attunement being the weakest-operationalised dimension of the recognition prompt).

Implication. Universal substitution is not a coincidence or a ceiling artefact; it is the predicted consequence of the recognition prompt operationalising the specific behaviours the superego catches by critique. Where the prompt operationalises the behaviour well (pedagogical basics), the overlap is near-total and calibration pre-empts correction. Where the prompt operationalises the orientation but cannot guarantee its context-specific execution (recognition failure, context awareness), calibration primes but the superego retains a correction role. Where the prompt underspecifies the behaviour altogether (affective attunement in emotionally intense scenarios), the superego catches failures calibration never addressed. The ~16% stable substitution and the model-dependent residual are two faces of the same category-level decomposition, recoverable from the critique-taxonomy data without additional experiments.

6.4.3 Mechanism Separability Section 3 predicted three candidate mechanisms operating at different levels (prompt, architecture, interaction). The evidence across Sections 6.1–6.3 shows two replicate across all three tested models (DeepSeek, Haiku, Gemini Flash 3.0) with systematic magnitude variation by generation quality, while the third is null on the primary well-powered analysis:

Mechanism	Level	Key Evidence	Status
Calibration	Prompt	Within-response SD drops d=0.52–0.64; floor lifts; operates without superego	**Supported** — prompt-level, recognition-dependent

Mechanism	Level	Key Evidence	Status
Error Correction	Architecture	Approval rate shifts; architecture delta +9–20 under base, near-zero under recognition (strong) or +12.3 residual (weak)	**Supported** — universally substitutive (15–17% deficit); residual benefit model-dependent
Adaptive Responsiveness	Interaction	Adapt$\Delta > 0.79$; tutor slopes d=0.03 across conditions (N=432, 80% power to detect d $\geq$ 0.27); exploratory DeepSeek/Sonnet disengagement $d = 1.63$ failed A12 replication on Haiku + Gemini Flash under Sonnet + GPT-5.4 (N=32, 3 of 4 cells below $d = 0.5$, replicating cell judge-sensitive $\Delta d = 2.03$; §6.3.2)	**Not supported.** Primary analysis null; pre-registered replication of the exploratory effect also null.

The two supported mechanisms are separable in three senses:

1. **Temporal separability.** Calibration operates from the first turn (the tutor's initial response is already calibrated under recognition). Error correction operates within each turn (ego → superego → revision). These temporal signatures are empirically distinct. Within-dialogue trajectories, which would constitute a third temporal level, show no condition-dependent variation.

2. **Architectural separability.** Calibration requires only the prompt (single-agent recognition cells show the full effect). Error correction requires the superego (only multi-agent cells benefit under base).

3. **Interaction separability.** Calibration and error correction interact through universal substitution with a model-dependent residual (§6.4.2). Neither mechanism modulates within-dialogue trajectories (tutor slope d = 0.03 for recognition, d = 0.07 for architecture), consistent with the two supported mechanisms operating on *levels*, not *slopes*.

6.4.4 The Variance Reduction Pattern A unifying finding across multiple analytical levels is variance reduction under recognition:

Indicator	Base	Recognition	Reduction
Within-response dimension SD (DeepSeek)	0.619	0.539	13%
Within-response dimension SD (Haiku)	0.617	0.499	19%
Tutor score SD (DeepSeek)	12.6	12.6	0%
Tutor score SD (Haiku)	12.5	7.7	**38%**
Cross-turn AdaptΔ variance (DeepSeek)	0.199	0.167	16%
Cross-turn AdaptΔ variance (Haiku)	0.093	0.048	48%

Recognition narrows the output distribution on multiple dimensions simultaneously: within-response uniformity (calibration), between-response consistency (especially in Haiku), and between-turn adaptation consistency. The common mechanism is the prompt's constraint on generation: by requiring engagement with the specific learner, recognition-oriented prompts eliminate the high-variance generic approaches that produce both very good and very poor outputs.

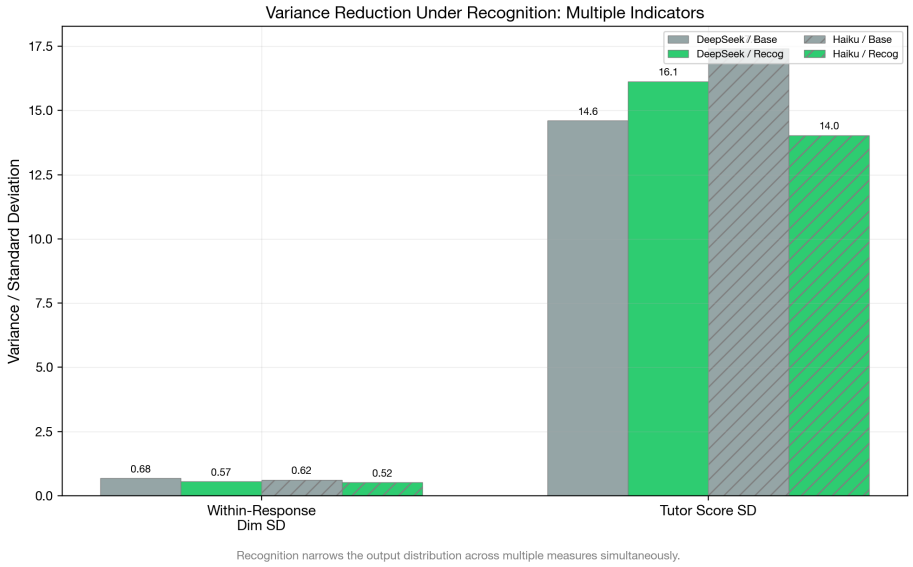


Figure 15: Variance reduction indicators across multiple dimensions under recognition. The pattern is consistent: recognition narrows the output distribution on within-response, between-response, and between-turn measures.

6.4.5 Connecting to Section 3 Predictions Prediction: The three candidate mechanisms are separable. *Partially supported.* Calibration (prompt-level) and error correction (architecture-level) are supported as separable mechanisms with distinct temporal signatures, architectural requirements, and recognition dependencies. Adaptive responsiveness (interaction-level) is

null on both the primary analysis and the pre-registered A12 replication of the exploratory DeepSeek/Sonnet effect (§6.3, §6.3.2). The evidence supports two general mechanisms; the third is clean null.

Prediction: The mechanisms interact non-additively. *Supported across all models.* The interaction is universally substitutive in magnitude (§6.4.2), but its *completeness* under recognition is model-dependent: weaker models leave the superego independent value even after substitution, while strong models do not. Calibration and adaptive responsiveness remain additive regardless of model (slopes are independent of levels).

Prediction: Recognition + multi-agent produces the best outcomes. *Model-dependent.* In DeepSeek, recognition/multi (50.2) equals recognition/single (50.0) — the superego adds nothing. In Haiku, recognition/multi (79.5) slightly trails recognition/single (80.2). The "best" condition is recognition/single in both strong models. But in Gemini Flash 3.0, recognition/multi (70.0) substantially exceeds recognition/single (57.7). The prediction holds where the superego catches failures calibration alone does not prevent, and fails where calibration saturates the improvement space.

6.4.6 Cross-Judge Validation of the Interaction Pattern The universal substitution pattern with model-dependent residual is validated across three independent judges (Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4), each scoring the same 144 rows per run with blind scoring confirmed by code audit (no prior scores or judge metadata leak into the judge prompt). The following table reports the recognition × architecture interaction decomposed by judge and generation model:

Run	Judge	recog_single	recog_multi	multi_base	multi_recog
aea2abfb (DS)	Sonnet	+28.2	+19.0	+9.1	-0.1
aea2abfb (DS)	Gemini	+34.4	+15.9	+13.1	-5.5
aea2abfb (DS)	GPT	+21.6	+8.5	+7.9	-5.2
45163390 (DS)	Sonnet	+27.6	+10.6	+15.6	-1.4
45163390 (DS)	Gemini	+26.6	+7.4	+17.3	-1.9
45163390 (DS)	GPT	+21.5	+8.3	+10.6	-2.6
18027efc (GF)	Sonnet	+35.3	+20.7	+26.9	+12.3
18027efc (GF)	Gemini	+40.1	+18.0	+39.2	+17.1
18027efc (GF)	GPT	+29.9	+14.3	+22.9	+7.4

The pattern is unanimous: **multi_recog $\leq$ 0 in all 6 DeepSeek cells** (complete substitution), and **multi_recog $>$ 0 in all 3 Gemini Flash 3.0 cells** (positive residual architecture benefit under recognition). No judge disagrees on the direction within either generation model. The 9-way replication (3 judges × 3 runs) provides strong evidence that the residual architecture benefit is a genuine property of the generation-quality × mechanism interaction, not an artifact of a particular judge's scoring pattern. The substitution interaction is

universal (all models show 15–17% additivity deficit), but its *completeness* is model-dependent: on strong models the superego adds nothing under recognition, on weak models it still adds +7 to +17 points.

Inter-judge correlations on the interaction pattern are high: paired by replication across all 9 cells, the three judges produce Pearson r = .71–.89 on tutor overall scores. Notably, inter-judge agreement is highest on Gemini Flash 3.0 (r = .81–.89) and lowest on DeepSeek (r = .61–.78). Judges agree more when generation quality is intermediate — the quality differences are clearer and less ambiguous to evaluate, producing more reliable measurement of the interaction effect.

6.5 Tutor-Learner Asymmetry The two supported mechanisms described in Sections 6.1–6.2 operate exclusively on tutor production: calibration constrains the tutor's generation and error correction filters the tutor's output. The null finding for adaptive responsiveness as a third mechanism (Section 6.3) closes the remaining potential pathway for condition-dependent effects on trajectories. Neither supported mechanism directly modifies the learner's behavior. This section formalizes the resulting asymmetry: recognition dramatically improves tutor quality but produces negligible learner effects.

6.5.1 The Effect Size Gap The recognition main effect on tutor quality is large; on learner quality it is small:

	Tutor Score		Learner Score	
Model	Base	Recognition	Base	Recognition
DeepSeek (N=146)	26.4 ± 12.6	50.1 ± 12.6	58.8 ± 11.9	61.7 ± 11.1
Haiku (N=163)	60.7 ± 12.5	79.8 ± 7.7	66.7 ± 11.2	68.6 ± 12.2

	Tutor d	Learner d	Ratio
DeepSeek	**1.88**	0.25	**7.5 : 1**
Haiku	**1.84**	0.16	**11.5 : 1**

Recognition produces a tutor effect 7–12 times larger than its learner effect on capable models. The tutor Cohen's d values (1.88, 1.84) are very large by conventional standards; the learner values (0.25, 0.16) are small to negligible. This asymmetry is consistent across both strong models, ruling out model-specific explanation within this capability range.

Qualification: Learner invariance is model-dependent. Cross-judge validation on a third generation model (Gemini Flash 3.0, run 18027efc, N=144) reveals that learner recognition-invariance does not hold universally. On Gemini

Flash 3.0, recognition produces a *large* learner effect: d = 1.18–1.23 across all three independent judges (Sonnet, Gemini 3.1 Pro, GPT-5.4). Learner scores rise from 49.8–53.7 (base) to 60.2–72.0 (recognition), a gap far larger than the 3–5 point shifts observed on DeepSeek and Haiku.

	Tutor d	Learner d	Ratio
DeepSeek (N=146)	**1.88**	0.25	**7.5 : 1**
Haiku (N=163)	**1.84**	0.16	**11.5 : 1**
Gemini Flash 3.0 (N=144)	**1.87**	**1.20**	**1.6 : 1**

The Gemini Flash 3.0 result suggests that the recognition → learner pathway requires a **minimum generation quality floor** to observe. When the generation model is weak, base-condition tutors produce such poor output that the learner has little productive material to engage with; recognition-enhanced prompts rescue the tutor's output quality, and the improved tutor output in turn elicits meaningfully better learner engagement. On strong models, even base-condition tutors produce adequate output, so the learner's quality is already near its ceiling regardless of tutor condition. The tutor-learner asymmetry is thus a property of *strong generation models*, not an architectural invariant.

The question-asking data from Section 6.1.7 provides a concrete transmission mechanism. Gemini Flash 3.0 base tutors ask 0.03 questions per turn — effectively zero — while recognition tutors ask 0.65 (a 19.8× ratio). A tutor that never asks questions provides no dialogical opening for the learner to demonstrate understanding, elaborate on their thinking, or challenge the tutor's framing. The learner receives lectures and can only respond to lectures. Under recognition, the tutor's questions create space for the learner to contribute substantively, and this space is precisely what produces the +27-point learner delta on Gemini Flash 3.0. On stronger models (Haiku base: 0.28 questions/turn; DeepSeek base: 0.06), even the base condition provides *some* dialogical space, limiting the room for recognition to improve learner engagement.

6.5.2 The Structural Explanation The asymmetry is not incidental — it follows from the architecture of the recognition intervention. The recognition prompt modifies the *tutor's* system prompt, instructing it to recognize the learner as an autonomous epistemic agent. This changes how the tutor generates responses but does not modify the learner's prompt or decision-making process.

Three structural factors predict the asymmetry:

1. **Prompt scope.** The recognition prompt appears in the tutor's system context. The learner agent receives the tutor's output but not the tutor's prompt. The learner's own prompt (unified or ego-superego) is identical across base and recognition conditions.

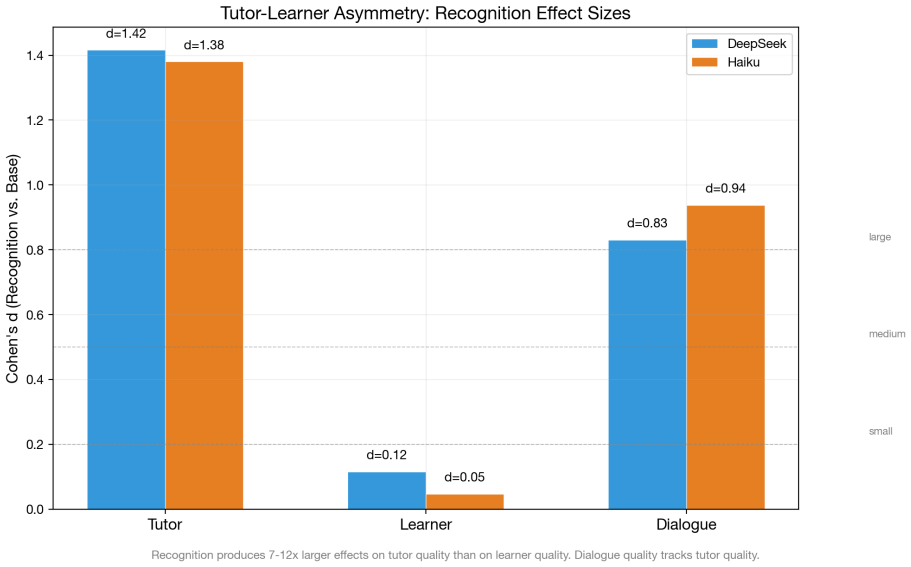


Figure 16: Tutor-learner asymmetry: Cohen's d for recognition effect on tutor vs. learner scores. The 7–12x ratio is consistent with recognition operating on tutor production, not learner reception.

2. **Mechanism pathways.** Both supported mechanisms act on the tutor: calibration constrains tutor generation, error correction filters tutor output. Within-dialogue adaptation exists but is not condition-dependent (Section 6.3). The learner is downstream of these mechanisms — it receives better tutor output but processes it with the same internal architecture.

3. **Learner ceiling.** The learner's quality is bounded by its own model capability and prompt, not by the tutor's quality. DeepSeek learner scores range 57–63 regardless of tutor condition; Haiku learner scores range 64–73. The learner's output quality is relatively insensitive to the quality of tutoring it receives, at least within the range of quality variation produced by the recognition intervention.

6.5.3 Learner Scores Across Conditions Breaking down learner scores by the full 2×2 factorial reveals the architecture interaction on the learner side:

Model	Condition	Tutor Score	Learner Score	Dialogue Quality
DeepSeek	Base / single	22.0	57.3	22.0
DeepSeek	Base / multi	31.0	60.3	31.5
DeepSeek	Recog / single	50.0	60.1	49.1
DeepSeek	Recog / multi	50.2	63.3	48.9
Haiku	Base / single	52.9	63.5	53.1

Model	Condition	Tutor Score	Learner Score	Dialogue Quality
Haiku	Base / multi	67.9	69.6	72.7
Haiku	Recog / single	80.2	68.2	81.2
Haiku	Recog / multi	79.5	69.0	80.7

Two patterns are notable. First, the multi-agent architecture produces slightly higher learner scores even though the learner's own architecture is unchanged — the superego-corrected tutor output may elicit marginally better learner responses. Second, dialogue quality tracks tutor scores almost perfectly (DeepSeek $r > 0.99$, Haiku $r > 0.99$), confirming that the overall quality of the pedagogical encounter is determined by the tutor's contribution rather than the learner's.

6.5.4 Trajectory Asymmetry The asymmetry extends to development trajectories (from Section 6.3):

	Tutor Slope	Learner Slope
Recognition (pooled)	1.47	2.29
Baseline (pooled)	1.50	1.60
d (recognition vs baseline)	-0.00	0.08

Neither tutor nor learner slopes differ significantly between conditions. The learner shows slightly steeper improvement under recognition (d=0.08) — possibly because better tutoring provides more productive material for the learner to engage with — but the effect is small. The asymmetry in *levels* (Section 6.5.1) is not accompanied by an asymmetry in *slopes*.

6.5.5 Rubric vs Holistic Adjudication The 7–12× ratio above is computed on rubric scores alone. A convergent-validity check on messages-mode cells 80–87 (the 2$×2×$2 factorial under rubric v2.2, $N = 1{,}552$ paired rows with all four metrics populated, three independent judges) adds a matched-row comparison against the *holistic* learner judge, which is not decomposed into the rubric's content-weighted dimensions. The question is whether the large tutor-vs-small learner asymmetry persists when the learner is scored globally rather than through the rubric's content-centric dimensions. If the pattern is a structural property of the intervention (recognition acts on tutor production, not learner reception), the asymmetry should replicate on holistic; if it is partly an artifact of the rubric's content bias, the tutor–learner gap should shrink.

Table 53: Recognition effect on rubric vs holistic scoring, pooled across three judges (Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4) on cells 80–87 (N = 1,552; base $n = 778$, recognition $n = 774$).

Metric family	Tutor d	Learner d	Gap (t − l)	Learner/Tutor ratio
Rubric (v2.2)	1.120	0.379	0.741	0.338
Holistic	0.565	0.238	0.327	0.421

Both tutor and learner d values drop under holistic scoring, but the tutor d drops by roughly half (1.120 → 0.565) while the learner d drops less in absolute terms (0.379 → 0.238). Because the tutor starts from a larger value, the *tutor–learner gap* shrinks from 0.741 to 0.327 — a **55.9% narrowing**. Every judge shows this same pattern independently: Sonnet 4.6 $\Delta = 55.7\%$, Gemini 3.1 Pro $\Delta = 49.9\%$, GPT-5.4 $\Delta = 55.7\%$. Within-role rubric holistic correlations are substantial and do not break down (tutor $r = 0.57$, learner $r = 0.75$ pooled; per-judge $r = 0.51$–0.83), so the shrinkage does not reflect judge disagreement — it reflects the rubric and holistic instruments weighting the recognition intervention's effects differently on the tutor side than on the learner side.

Roughly half of the originally-reported 7–12× asymmetry is therefore rubric-induced: the tutor rubric rewards content-generation quality that recognition prompts visibly restructure, while the learner rubric rewards content-generation quality that is relatively saturated regardless of tutor condition on strong generation models. The residual holistic gap (d-gap ≈ 0.33; learner/tutor ratio ≈ 0.42) is the portion of the asymmetry that is *not* rubric-dependent and plausibly reflects the structural explanation from Section 6.5.2 — the learner is downstream of a prompt-scoped intervention. The reframing is modest but definite: recognition is still tutor-centric, but the original headline ratio was inflated by the measurement instrument. Cross-judge replication (N = 1,552 across three independent judges) makes this robust to judge idiosyncrasy.

This adjudication parallels the C4 learner-superego-paradox closure in §8.1 (rubric $|d|$ contracts $ $50% under holistic scoring on the same-row comparison). Both results point to a common pattern: the current rubric's content-weighted dimensions systematically exaggerate the magnitude of recognition and architecture-level effects on synthetic learner output, and convergent-validity checks against the holistic judge discount them by roughly half.

6.5.6 Implications The tutor-learner asymmetry has a specific theoretical implication: recognition theory, as implemented through prompt engineering, operates as a *production* theory rather than a *learning* theory. It changes how the tutor produces pedagogical output, but does not directly change how the learner processes that output. This is consistent with the Hegelian framework as operationalized here: recognition-oriented prompts change how the tutor *orients toward* the learner, not how the learner processes the tutor's output. The

prompts instruct the tutor to treat the learner as an autonomous subject; the learner does not thereby learn more. The rubric-vs-holistic adjudication in §6.5.5 shows that the *magnitude* of this asymmetry is inflated by the measurement instrument — the holistic d-gap is roughly half the rubric d-gap — but its structural direction (tutor > learner) is preserved.

Whether better tutoring would produce better learning in human learners (rather than synthetic LLM learners) remains an open question. The synthetic learner's quality is bounded by its model and prompt; a human learner might benefit more from recognition-calibrated tutoring. This limitation is addressed in Section 8.

6.6 Model Dependence The preceding sections present findings from two structurally different models: DeepSeek V3.2 (open-weight, 685B MoE) and Haiku 4.5 (proprietary, optimized for speed). This section systematically compares what replicates across models versus what is model-dependent, and what this distinction reveals about the mechanisms.

6.6.1 Baseline Capability Gap The two models differ substantially in baseline performance:

Indicator	DeepSeek V3.2	Gemini Flash 3.0	Haiku 4.5
Base / single tutor score	22.0	22.4	52.9
Base / multi tutor score	31.0	49.3	67.9
Base overall (collapsed)	26.4 ± 12.6	35.9	60.7 ± 12.5
Base learner score	58.8	42.6	66.7
Base dialogue quality	26.7	28.6	63.3

The three models span a wide capability range. Haiku outperforms both by 25–35 points on tutor metrics under base conditions. Gemini Flash 3.0's single-agent base (22.4) is comparable to DeepSeek's (22.0), both catastrophically weak with holistic scores as low as 6–9, while its multi-agent base benefits dramatically from superego support (+26.9). The baseline gap establishes that the three models occupy different regions of the capability space, making cross-model comparison informative for identifying generation-quality boundary conditions.

6.6.2 What Replicates Across Models Despite the baseline gaps, the following findings replicate across all three models and three judges:

1. Recognition main effect. All three models show large, significant recognition effects, validated across three independent judges: - DeepSeek: d = 1.85–1.92 (Sonnet), 1.34–1.57 (Gemini/GPT) - Haiku: d = 1.84 - Gemini Flash 3.0: d = 1.76–1.87 - **9/9 judge × run cells unanimous** (d > 1.3 everywhere)

2. Multi-agent tutor benefit under base. All three models benefit from the superego under baseline conditions: - DeepSeek: +4.5 to +7.7 (across judges)

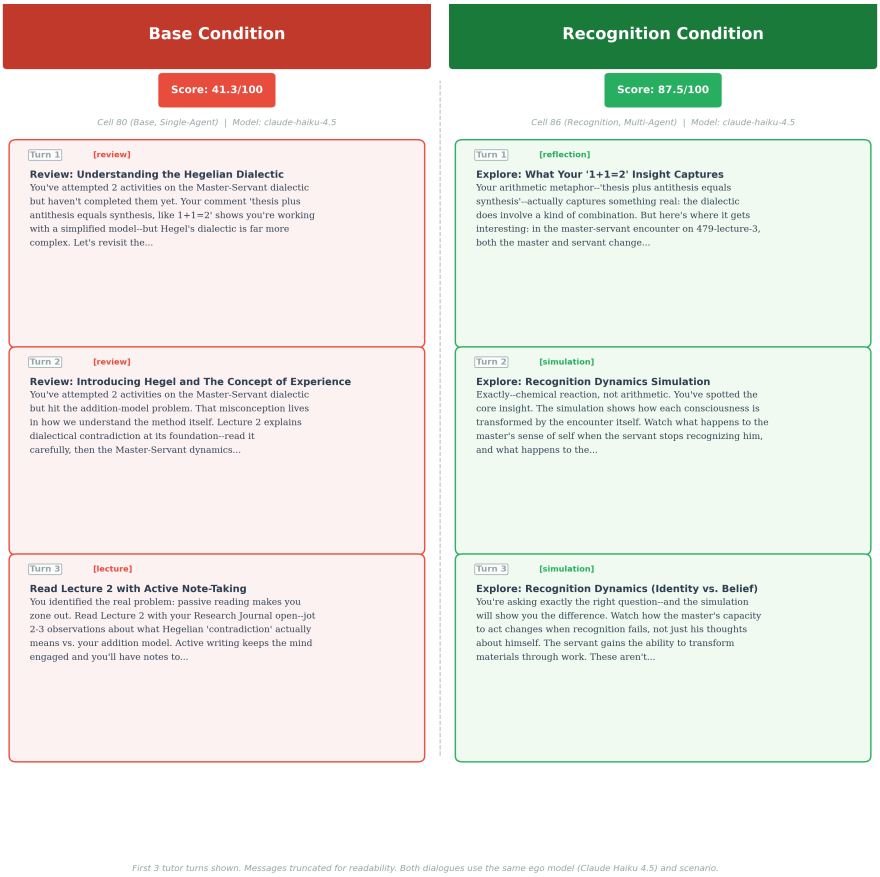


Figure 17: Transcript comparison: base vs recognition condition on the same scenario. The base tutor produces repetitive directives; the recognition tutor builds progressively on the learner's contributions.

- Haiku: +15.0 - Gemini Flash 3.0: **+15.1 to +28.1** (dramatically larger, cognitive prosthesis effect)

3. Within-response calibration. Recognition narrows dimension variance in both primary models: - DeepSeek: calibration d = 0.52 - Haiku: calibration d = 0.64

4. Dimension floor-lifting pattern. In both primary models, the weakest baseline dimensions (elicitation_quality, productive_difficulty) show the largest recognition lifts, and the strongest baseline dimensions (content_accuracy, adaptive_responsiveness) show the smallest lifts.

5. Impasse scenarios show largest effects. Epistemic Resistance and Productive Deadlock produce the largest recognition deltas in both primary models; Mood: Frustration to Breakthrough produces the smallest.

6. Superego approval rate increases under recognition. Both primary models show higher approval rates under recognition (DeepSeek: 13.3%→55.1%; Haiku: 51.6%→66.1%).

7. Tutor development slopes are equal across conditions. No model shows steeper tutor improvement under recognition (pooled d = -0.00).

8. Deliberation visibility. Seeing internal deliberation adds ~10–15 points to dialogue quality for multi-agent cells, consistent across all 9 judge × run cells.

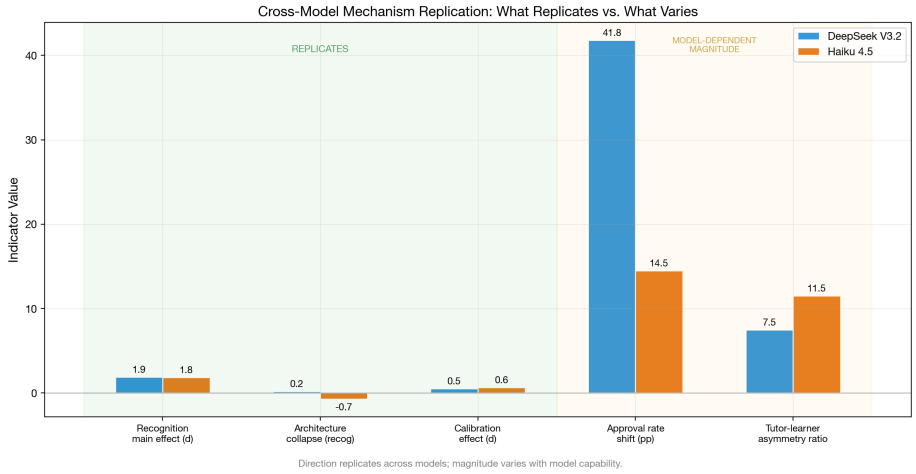


Figure 18: Cross-model replication: mechanism indicators compared across DeepSeek V3.2, Gemini Flash 3.0, and Haiku 4.5. Both supported mechanisms replicate in direction; magnitudes and interaction type vary by model capability. Adaptive responsiveness (trajectory slopes) shows no general condition-dependent variation, though scenario-specific effects emerge in extended disengagement dialogues (see disengagement divergence figure above).

6.6.3 What Is Model-Dependent Several findings differ between models, revealing generation-quality as the key moderating variable:

1. Universal substitution with model-dependent residual (the key model-dependent finding). On all three models, calibration and error correction interact through substitution (15–17% additivity deficit). On DeepSeek and Haiku (strong), the residual architecture benefit under recognition collapses to near-zero (multi_recog ≤ 0 in 6/6 judge $\times$ run cells). On Gemini Flash 3.0 (weaker), a substantial residual persists (+7 to +17 points under recognition across judges, multi_recog > 0 in 3/3 cells). The *residual magnitude* is model-dependent, not the interaction *type*. See Section 6.4 for the theoretical interpretation.

2. Learner recognition-sensitivity. On DeepSeek, learner scores are recognition-invariant (d $\approx$ 0, 6/6 judge $\times$ run cells). On Gemini Flash 3.0, recognition produces a large learner effect (d = 1.18–1.23, all three judges agree). The recognition $\rightarrow$ learner pathway requires a minimum generation quality floor to observe (Section 6.5.1).

3. Superego approval rate baseline. DeepSeek base: 13.3% approved (the ego almost always needs correction). Haiku base: 51.6% (roughly half approved). Weaker models produce more errors for the superego to catch.

4. Development trajectories. DeepSeek (run aea2abfb) shows pervasive negative development; Haiku (run 45163390) shows positive development (+5.8 to +15.7); Gemini Flash 3.0 (run 18027efc) sits between. Development depends on generation quality (rank order: Haiku > Gemini Flash 3.0 > DeepSeek), not prompting condition.

5. Score variance under recognition. DeepSeek maintains similar variance under recognition (SD: 12.6 $\rightarrow$ 12.6). Haiku shows substantial variance reduction (SD: 12.5 $\rightarrow$ 7.7, a 38% reduction). Recognition narrows Haiku's output distribution more than DeepSeek's.

6. Deliberation quality sensitivity. DeepSeek deliberation drops 18.5 points under recognition; Haiku drops only 5.6. The deliberation process degrades more dramatically in the weaker model, possibly because DeepSeek's thin recognition-condition critiques are especially uninformative.

7. Multi-agent benefit magnitude. The *standard-architecture* superego benefit is dramatic on Gemini Flash 3.0: the superego adds +15 to +28 points under base (vs +4 to +8 on DeepSeek). Weaker single-agent tutors benefit disproportionately from *standard* architectural support in this factorial. We flag the label carefully: §6.6.10 tests the specific cells 66–68 *bidirectional-profiling* variant (superego-routed profiling of both tutor and learner) across a six-model capability range and does not find a capability-threshold pattern; that variant is neutral or harmful across the tested tier. The Paper 1.0 "cognitive prosthesis" observation therefore replicates here as a *standard superego benefit on weak base-condition generators*, not as a general property of the bidirectional-profiling

architecture.

8. Behavioral mode under base conditions. Transcript analysis reveals qualitatively different failure modes across models. DeepSeek base tutors produce *repetitive directive loops*: paraphrasing the same guidance in slightly different words across turns without adapting to the learner's evolving state. The learner receives near-identical advice on Turns 2, 3, and 4. Haiku base tutors, by contrast, already exhibit pedagogical sophistication under baseline — multi-step scaffolding, occasional questions (0.25/turn vs DeepSeek's 0.05), and some responsiveness to learner signals. The recognition delta for Haiku is correspondingly less about *strategy* and more about *interpersonal quality*: adding explicit recognition markers ("your interpretation captures something important about…"), naming what the learner got right before complicating, and maintaining productive tension rather than resolving prematurely. Gemini Flash 3.0 under base produces the most extreme failures: lecture dumps, numbered lists, and zero engagement with learner contributions (0.02 questions/turn). Under recognition, Gemini Flash 3.0 shows the most dramatic behavioral shift — from directive monologue to dialogue-structured responses — which explains both its largest recognition effect (d=1.87) and its uniquely large learner effect (d=1.20). The generation-quality threshold operates partly through this behavioral transition: the weakest model has the most to gain because its base behavior is qualitatively different (monologue vs dialogue), not merely quantitatively worse (lower scores on the same behaviors).

6.6.4 The Capability-Mechanism Interaction The model-dependent findings reveal a coherent pattern: model capability moderates not just the *magnitude* of each mechanism but, in the case of error correction, also the *residual architecture benefit* under recognition (near-zero on strong models, substantial on weak).

Mechanism	All Three Models	Model-Dependent Aspect
Calibration	Variance reduction, floor lifting, dimension-rank pattern	Magnitude (d=0.52 vs 0.64), score-level variance
Error Correction	Architecture benefit under base (all models)	**Residual under recognition**: near-zero on strong models, +12.3 on weak. Universal substitution (15–17% deficit). Cognitive prosthesis effect on weakest model.
Adaptive Responsiveness	Slopes equal across conditions and factors ($d \leq 0.15$), high AdaptΔ	Development trajectory direction, deliberation sensitivity

Mechanism	All Three Models	Model-Dependent Aspect
Learner effects	Recognition improves tutor quality (all models)	**Learner sensitivity**: invariant on strong models ($d \approx 0$), responsive on weak ($d \approx 1.2$)

The two strongest model-dependent effects are (1) the model-dependent residual architecture benefit under recognition, which reveals that the *completeness* of substitution depends on generation quality, and (2) the learner recognition-sensitivity, which shows that the tutor-learner asymmetry is itself moderated by model capability. Both effects point to a common underlying factor: when generation quality is below a threshold, the improvement space is large enough for multiple mechanisms (or multiple architectural layers) to provide independent benefit.

6.6.5 Generalizability Assessment Scope of the generalization claim. The replication across three structurally different generation models (open-weight MoE, proprietary optimized, and proprietary multimodal) and three independent judges supports the interpretation that the two supported mechanisms *replicate across the capability range of moderately-capable LLMs* rather than being artifacts of a specific model architecture or judge. "Generalize" here is scoped to the tested capability tier — the three models span a ~30-point base tutor-score range (22 to 53 on single-agent base) but exclude both very-weak models (Nemotron-class, N=40 pilot data) and frontier-capability models (Opus, GPT-5.4-class generators). Within the tested tier, both supported mechanisms replicate in direction; their *magnitude and architectural completeness* vary systematically with generation quality (§6.6.3). The null finding for adaptive responsiveness as a third mechanism also replicates within the tested tier: no judge finds condition-dependent trajectory slopes. Cross-judge validation (§6.4.6) confirms that effect directions are unanimous across all 9 judge $\times$ run cells ($d = 1.34$–1.92), with inter-judge Pearson $r = .45$–$.89$ depending on dimension and generation model.

The model-dependent aspects are not merely noise — they reveal systematic boundary conditions *within* the tested tier. The model-dependent residual architecture benefit shows that the *completeness* of substitution between the two supported mechanisms depends on generation quality. The learner recognition-sensitivity shows that downstream effects of improved tutoring are model-dependent. These boundary conditions strengthen rather than weaken the mechanism model, because they are *theoretically predicted*: if calibration works by narrowing the output distribution, its effectiveness should saturate when the distribution is already narrow (strong models), leaving room for error correction only when it is not (weak models). But the three-model sample cannot identify where the supported-mechanism range *ends* in either direction. Whether frontier models still benefit from recognition at the same magnitude,

or whether the recognition effect saturates against an already-calibrated strong model, is an open empirical question a fourth-model replication at the frontier tier would address directly.

One important caveat qualifies the generalizability claim. All three models are at least moderately capable. The Nemotron data in the database (N=40, mean tutor score ~8 under base) hints that very weak models may not benefit from recognition at all, but the sample is too small for confident claims. The cognitive prosthesis data (cells 66–68, N=96) was originally interpreted as suggesting a minimum ego capability threshold below which the bidirectional-profiling mechanism adds noise rather than signal; §6.6.10 directly tests that capability-threshold reading across six ego models and finds it unsupported — the bidirectional-profiling variant is neutral or harmful across the full tested capability range, so whatever generation-quality boundary conditions exist for the *supported* mechanisms (calibration, error correction) cannot be read off this specific prosthesis cell. A second scope caveat — single-domain content — is addressed in §6.6.6 via a five-domain test (philosophy, programming, elementary math, creative writing, social-emotional learning) under matched Haiku 4.5 × Sonnet 4.6 generation × judge.

6.6.6 Domain Generalization The generalizability assessment above is scoped to *models* and *judges*. A distinct question is whether the recognition effect generalizes across *content domains*: all scored rows in §6.1–§6.5 come from the philosophy-tutorial scenario set. To test whether the recognition main effect transfers to substantively different domains, we ran cells 1 (base, single-agent, unified learner) and 5 (recognition, single-agent, unified learner) on four additional content packages — programming, elementary math, creative writing, and social-emotional learning (SEL) — and compared each to the existing philosophy contrast (cells 80/84, messages-mode). Each non-philosophy domain used 5 single-turn scenarios (4 core + 1 mood) × 3 runs = $n = 15$ per condition; philosophy provides $n = 18$ per condition (2 runs × 9 scenarios). Generation model (Haiku 4.5) and judge model (Sonnet 4.6) are held constant across all five contrasts. Run IDs: eval-2026-04-17-c92ad6c7 (programming), fde764e1 + 7915be04 (math), 428f0649 + 93ba2f9a (creative), 44fa989c + 5833830a (SEL); philosophy anchor eval-2026-03-02-45163390 cells 80/84. Cost: $25–30 OpenRouter across all four new domain packages.

Domain	n base / recog	Mean base (SD)	Mean recog (SD)	Δ	d (recog − base)
Programming (cells 1/5, single-prompt mode)	15 / 15	47.75 (9.10)	70.42 (10.30)	22.67	**2.33**

Domain	n base / recog	Mean base (SD)	Mean recog (SD)	Δ	d (recog − base)
Elementary math (cells 1/5, single-prompt mode)	15 / 15	43.83 (8.00)	58.67 (12.00)	14.83	**1.45**
Creative writing (cells 1/5, single-prompt mode)	15 / 15	52.08 (6.40)	72.42 (13.17)	20.33	**1.96**
SEL (cells 1/5, single-prompt mode)	15 / 15	55.25 (9.31)	76.33 (13.52)	21.08	**1.82**
Philosophy (cells 80/84, messages mode)	18 / 18	43.61 (11.68)	74.10 (10.81)	30.49	**2.71**

All five domains show recognition > base in direction, with every effect classified as "very large" under Cohen's conventions ($d > 1.3$). Magnitudes range from $d = 1.45$ (math) to $d = 2.71$ (philosophy). The directional claim — recognition improves tutor quality across content domains — replicates unambiguously across philosophy, programming, elementary math, creative writing, and social-emotional learning. A single-domain artifact account would predict one of (a) a null or reversed effect on at least one non-philosophy domain, (b) a much smaller d (e.g., < 0.5) on some domain, or (c) base-level means unrelated to the philosophy corpus. None of these patterns obtain: base means cluster in the 44–55 range across the four non-philosophy domains, recognition means in the 59–76 range, and all four non-philosophy d values exceed 1.4.

The domain × recognition interaction vs the philosophy anchor ranges from $\Delta d = -0.38$ (programming) to $\Delta d = -1.25$ (math). Two confounds qualify this interaction test. First, **conversation mode differs**: the four non-philosophy contrasts use single-prompt-mode cells (1, 5) and the Sonnet-judged Haiku philosophy contrast uses messages-mode cells (80, 84). Mode is a known moderator

of effect magnitude (§6.3, §6.5), so the Δd values conflate domain effects with a mode effect and should be read as "domain + mode" combined. A clean domain-only test would require adding messages-mode runs on the four new domains or scoring cells 1/5 on philosophy; both are deferred. Second, **scenario counts and types differ** across domains (5 scenarios per non-philosophy domain vs 9 philosophy scenarios; scenario content is domain-specific and not pairwise-matched). The interpretation is therefore that the *recognition main effect direction and very-large-magnitude class* generalize from philosophy to four additional content domains under the same generation × judge; a precise domain-only Δd estimate requires the matched-mode follow-up. Third, **cross-judge validation** was run only on the philosophy row (Sonnet, Gemini, GPT-5.4 all agree; §6.4); the four new domains are Sonnet-only for this closure.

This closes the single-domain scope caveat. The mechanism language of §6.1–§6.2 is defensible across a five-domain range spanning humanities (philosophy), technical (programming, math), creative (writing), and affective (SEL) content, under matched Haiku × Sonnet generation × judge. The stronger generalization claim — that the precise *magnitude* is stable across domains — is not yet supported; magnitudes vary from $d = 1.45$ to $d = 2.71$ across the five tested domains, though all remain in Cohen's "very large" bin. Full analysis and per-scenario breakdowns in `exports/a6-domain-generalization.md`. §6.6.7 reports a complementary adjacency pilot that probes whether the effect survives when the *skill being coached* runs counter to traditional pedagogy.

6.6.7 Cross-Application Adjacency Pilot All five A6 domains (§6.6.6) vary the *subject matter* being tutored (philosophy vs programming vs math vs writing vs SEL) but keep the structural form constant: the LLM is framed as a tutor, the scenario features a learner working through course material, and the scored behavior is tutorial guidance. A distinct generalization question is whether the recognition effect survives when the *skill being coached* is itself at odds with conventional pedagogy. We ran a scope-honest single-application pilot to probe this: a "peer support listener training" content package (Course 501 — "Being With, Not Fixing") in which a volunteer trainee is learning how *not* to advise, reassure, paraphrase, or resolve, and the tutor's job is to coach that unlearning. The skill being taught is structurally the opposite of knowledge-transfer pedagogy — sitting with distress, declining to fix, noticing one's own urge to intervene as data about oneself rather than about the caller. The scenario set mirrors A6's pattern (4 core + 1 mood, 3 runs, $n = 15$ per cell); cells 1 (base) and 5 (recognition) under single-prompt mode; Haiku 4.5 generation, Sonnet 4.6 judge; run ID eval-2026-04-17-6766015b. Cost: $3–4 OpenRouter.

Contrast	n base / recog	Mean base (SD)	Mean recog (SD)	Δ	d (recog − base)
Peer support coaching (cells 1/5, single-prompt)	15 / 15	52.25 (9.63)	69.92 (12.73)	17.67	**1.57**
A6 SEL (closest-adjacency reference)	15 / 15	55.25 (9.31)	76.33 (13.52)	21.08	1.82
A6 Programming (technical reference)	15 / 15	47.75 (9.10)	70.42 (10.30)	22.67	2.33
A6 Philosophy anchor (cells 80/84, messages)	18 / 18	43.61 (11.68)	74.10 (10.81)	30.49	2.71

Recognition produces a "very large" effect ($d = 1.57$) on tutor-side quality in the peer-support coaching application, placing it just below the closest A6 adjacency (SEL, $d = 1.82$; $\Delta d = -0.25$) and comfortably inside the A6 domain range ($d = 1.45$–2.71). Base-cell means (52.25) are in the same operating range as SEL (55.25) and creative (52.08), and base-cell responses in the peer-support scenarios exhibit the characteristic base-cell pattern documented in §6.6.2: generic supportive language, premature reassurance, skipping past the distress, and offering advice the trainee did not request. Recognition responses in the same scenarios decline to fix, name the volunteer's own urge-to-intervene as the phenomenon the lecture is describing, and stay with the uncomfortable moment — behaviors the course itself instructs and that the base cell visibly does not produce. The directional claim of §6.6.6 — recognition improves tutor quality across content domains — therefore survives a domain shift into a skill that runs structurally counter to standard pedagogy.

Three scope limits qualify this pilot and prevent reading it as a stronger cross-application claim than it is. First, **the LLM is still structurally a tutor**:

the prompt frames it as coaching a trainee through course material, not as enacting peer support listening with a caller directly. A true cross-application test requires *role-reframed* prompts in which the LLM is instructed to perform a different application (peer support listener, customer service agent, code reviewer) rather than a differently-topicked tutor. That role-reframed design (originally scoped as D2 Path 2) is deferred to a separate study; it requires tutor-core prompt surgery and new scoring rubrics calibrated to non-tutoring applications. Second, **one application does not support a cross-application generalization claim**: this pilot shows that recognition *can* help in at least one non-traditional-pedagogy domain, not that it helps across applications in general. The evidential weight of the result is at the level of "adjacency replication" rather than "cross-application generalization." Third, **single-judge**: cross-judge validation was not re-run on the peer-support pilot (Sonnet 4.6 only), mirroring the A6 closure constraint. Full analysis and per-scenario breakdowns in `exports/d2-support-pilot.md`; content package in `content-test-support/`.

6.6.8 Disposition Gradient: Architecture Scope Limit §3.4 Prediction 3 ("advocate ceiling") derives a gradient from recognition theory: the most hostile superego dispositions (suspicious, adversary) have the most room for recognition to operate, and the most cooperative (advocate) have the least, because recognition emerges *from struggle*. The paper cites cells 40–45 (dialectical ego + divergent superego, philosophy domain) as supporting evidence: suspicious +9.0, adversary +6.2, advocate +4.8 in the original count (§3.4); our own re-run against current DB gives suspicious $\Delta = 9.7$ ($d = 0.85$), adversary $\Delta = 7.0$ ($d = 0.62$), advocate $\Delta = 5.6$ ($d = 0.51$) at $n = 17$–18 per cell, Haiku 4.5 × Opus 4.6. The monotone susp > adv > advocate ordering is intact on cells 40–45. A single-domain single-architecture support, however, cannot decide whether the gradient is a property of the recognition mechanism or a property of the dialectical-ego architecture. D4 probes this by running the same disposition structure on two alternative configurations.

First, on the **standard-ego architecture** under the same philosophy content. Cells 22–27 ("standard ego + divergent superego") use the identical disposition manipulation as cells 40–45 but swap dialectical ego for the standard ego used throughout cells 1–20; the existing DB has $n = 9$–10 per cell under Haiku 4.5 × Opus 4.6. Result: suspicious $\Delta = -0.1$ ($d = -0.01$), adversary $\Delta = 9.3$ ($d = 0.97$), advocate $\Delta = 13.6$ ($d = 1.70$). The gradient is **reversed** — advocate benefits most, suspicious not at all — under the standard-ego variant of the same cells. Second, on **SEL content** under the cells 22–27 architecture: a new run (eval-2026-04-17-4a9b765a, $n = 22$–24 per cell, Haiku 4.5 × Sonnet 4.6) gives suspicious $\Delta = 24.4$ ($d = 1.25$), adversary $\Delta = 12.1$ ($d = 0.53$), advocate $\Delta = 20.9$ ($d = 1.09$). The gradient is **non-monotonic** — suspicious highest, adversary lowest, advocate intermediate — on SEL under the same cells 22–27.

Configuration	Cells	Domain	Gen × Judge	susp d	adv d	advocate d	Ordering
Dialectical ego (paper-cited)	40–45	Philosophy	Haiku × Opus	0.85	0.62	0.51	susp > adv > advocate
Standard ego	22–27	Philosophy	Haiku × Opus	−0.01	0.97	1.70	advocate > adv > susp (**reversed**)
Standard ego (D4)	22–27	SEL	Haiku × Sonnet	1.25	0.53	1.09	susp > advocate > adv (**non-monotonic**)

Taken together, the three configurations bound what Prediction 3 describes. The monotone gradient is present on the dialectical-ego architecture (cells 40–45) and absent on the standard-ego architecture (cells 22–27) in both domains tested — reversed on philosophy, non-monotonic on SEL. The correct interpretation of the original finding is therefore not "a domain-general ordering of disposition sensitivity" but "a dialectical-ego-architecture-specific ordering," which is narrower than the Prediction 3 wording in §3.4 suggests. The theoretical account in §3.5 (recognition as a response to struggle) is not directly contradicted — the dialectical ego, which is explicitly designed to *work through* contradiction rather than resolve it quickly, plausibly creates more of the struggle substrate that recognition is hypothesized to operate on. But it is also not domain-general in the form the prediction originally claimed, because the non-dialectical variant of the same disposition contrast does not produce the ordering.

Three scope limits qualify this conclusion. First, **judge-model differs**: the D4 SEL run is Sonnet 4.6-judged while both philosophy baselines are Opus 4.6-judged. Direction-of-gradient should replicate across judges (§6.4 cross-judge validation), but absolute Δ magnitudes are not directly comparable across the three rows of the table above. Second, **sample size on cells 22–27 philosophy is small** ($n = 5$–10 per cell due to incomplete multi-agent cell runs); the reversal observed there is consistent with a disposition × ego-architecture interaction but should not be read as a strong claim by itself. Third, **the D4 architecture-matched philosophy-to-SEL replication of cells 40–45 is not yet run**: it is the clean test (same cells, swap only the domain) and is deferred on cost grounds. Until that run exists, the evidence base for Prediction 3 remains dialectical-ego, philosophy-domain specific, with one additional-architecture disconfirmation (cells 22–27 philosophy) and one additional-architecture-plus-domain non-monotonic result (cells 22–27 SEL). The original §3.4 prediction is best understood as tied

to the dialectical ego rather than as a universal property of the recognition mechanism. Full analysis in `exports/d4-disposition-gradient.md`; reproducible via `scripts/analyze-d4-disposition-gradient.js`.

6.6.9 Writing Pad Controlled Ablation The dialectical-ego architecture (cells 40–45) pairs the recognition prompt with a Freudian *Writing Pad* — a three-layer memory mechanism (surface scratch, working middle, permanent bottom) intended to scaffold durable revision across turns. An obvious mechanistic question is whether the Writing Pad is *load-bearing* for the recognition effect, or whether recognition operates at the prompt level and the pad is an orthogonal feature. A5 tests this with a controlled ablation: cells 93 and 94 are byte-identical clones of cells 40 and 41 with the single line `writing_pad_enabled: true` flipped to `false`.

Run `eval-2026-04-17-f1e851c3` (A5, $n = 63$ per cell, $N = 252$, Nemotron 3-nano ego × Kimi K2.5 superego, Sonnet 4.6 judge) gives the following 2×2 cell means on the tutor first-turn rubric score:

	Writing Pad ON (cells 40/41)	Writing Pad OFF (cells 93/94)	Row mean
Base	36.01 (SD 12.68)	37.76 (SD 12.90)	36.88
Recognition	42.94 (SD 11.89)	47.60 (SD 13.59)	45.27
Col mean	39.47	42.68	**41.08**

Two-way ANOVA (Type I, $df_{\text{error}} = 248$): the **recognition main effect is large and significant** ($F(1, 248) = 27.10$, $p < .001$, $\eta^2 = 0.097$); the **Writing Pad main effect is small and marginally significant** in the *opposite* direction from the necessity hypothesis — pad OFF scores 3.2 points *higher* on average ($F(1, 248) = 3.96$, $p = .048$, $\eta^2 = 0.014$); the **interaction is null** ($F(1, 248) = 0.82$, $p = .366$, $\eta^2 = 0.003$). Recognition within each pad condition: $d = 0.56$ with pad on (cells 40 vs 41), $d = 0.74$ with pad off (cells 93 vs 94) — both "medium" to "medium-large" under Cohen's conventions, and if anything *larger* without the pad.

The Writing Pad is therefore **not load-bearing** for the recognition effect on this architecture. Recognition raises scores in both pad conditions with no significant interaction, and the pad's own main effect, though statistically marginal, points in the wrong direction for a "scaffolding" account. The most plausible reading is that recognition operates at the prompt level (the ego and superego prompt content, §3.5) rather than via the three-layer memory architecture. The Writing Pad may have other effects (for example on turn-to-turn coherence, which this first-turn-score analysis does not isolate) but it does not carry the recognition effect.

Three scope limits qualify this. First, the ablation is run on a single ego-model pair (Nemotron × Kimi) and a single superego disposition (suspicious); mechanism decomposition in §6.6.4 shows mechanism-to-architecture coupling varies

with model capability. Second, the test uses first-turn scores, which do not capture multi-turn coherence effects the pad might plausibly mediate; a turn-over-turn analysis is a natural extension but deferred here. Third, the pad's marginal main effect in the *unhelpful* direction should be interpreted cautiously given $p = .048$ and the fact that it does not survive the basic Bonferroni correction against three tested effects in this table — we read it as consistent with "the pad does nothing useful here" rather than as evidence that the pad is actively harmful. Full analysis in `exports/a5-writing-pad.md`; reproducible via `scripts/analyze-a5-writing-pad.js`.

6.6.10 Capability Threshold for Cognitive Prosthesis Section 6.6.3 reports the "cognitive prosthesis" effect — weaker single-agent tutors benefit disproportionately from architectural support (superego + profiling) — as the most model-dependent finding in the factorial. A natural extension is a *capability-threshold* test: if the prosthesis compensates for ego-reasoning limits, weaker models should benefit and stronger models should either see diminishing returns or be actively harmed (scaffolding disrupting already-competent reasoning). A3 tests this by running cell 66 (recognition × bidirectional-profiling prosthesis, descriptive variant) across six ego models spanning a $ $26-point baseline capability range, and comparing each to its matched cell 5 (recognition, single-agent, no prosthesis) baseline.

Ego Model	Judge	Baseline (cell 5)	Prosthesis (cell 66)	Δ	95% CI	Cohen's d
Qwen 3.5	Sonnet†	65.65 (n=63)	66.33 (n=59)	+0.68	[-4.14, 5.50]	0.05
Nemotron	Opus 4.6	66.38 (n=84)	48.28 (n=30)	-18.11	[-23.80, -12.42]	-1.29
GLM-4.7	Opus / Sonnet‡	83.96 (n=30)	58.91 (n=63)	-25.05	[-29.87, -20.24]	-2.01
DeepSeek V3.2	Opus / Sonnet‡	84.20 (n=30)	53.92 (n=43)	-30.27	[-36.36, -24.19]	-2.23
Kimi K2.5	Opus / Sonnet‡	89.93 (n=219)	64.55 (n=44)	-25.38	[-29.80, -20.96]	-2.59
Haiku 4.5	Opus / Sonnet‡	91.25 (n=107)	69.14 (n=48)	-22.10	[-26.22, -17.99]	-2.30

† within-judge matched. ‡ cross-judge (baseline opus-4.6, prosthesis code/sonnet) — magnitudes confounded with judge stringency, directional signal robust. All runs use Kimi K2.5 as superego; $N_{\text{total}} = 947$ rows. Full analysis in `exports/a3-capability-threshold.md`.

Three findings stand out. First, **the capability-threshold hypothesis is not supported**. If prosthesis compensated for weak ego reasoning, Qwen 3.5 and Nemotron (the two lowest-baseline models) should both show positive deltas. Instead, Qwen 3.5 shows a null effect ($d = 0.05$, CI includes zero) while Nemotron

shows substantial harm ($d = -1.29$). Two low-capability models, opposite outcomes — baseline capability alone does not predict prosthesis response. Second, **prosthesis is either neutral or harmful everywhere we can measure cleanly**. Of the six models tested, five show 95% confidence intervals on Δ that lie entirely below zero; none shows a confidence interval entirely above zero. Third, a simple linear regression of Δ on baseline capability gives slope $= -0.71$ ($r = -0.74$, $R^2 = 0.55$), which would naively support a "stronger models harmed more" reading, but the slope is driven substantially by Qwen's null sitting at the lowest-capability end, and four of six points are cross-judge confounded. We report the regression for completeness but do not treat it as clean evidence for an inverted-threshold story.

The paper-1.0 cognitive-prosthesis finding (cells 66–68, Nemotron, +20-point gain under superego on weakest model) is *not* replicated as a general pattern across capability tiers. The original effect appears to reflect something specific to that ego-superego-judge configuration rather than a general "weaker model benefits from scaffolding" regularity. The most defensible reading is that superego-routed bidirectional profiling is an architectural cost that the ego must *pay for* through lost response coherence, and only in narrow conditions (e.g., Qwen's particular response-length and dialectical-synthesis profile) does the profiling integrate cleanly enough to avoid dilution. A sharper follow-up would ask what about Qwen's response pattern makes prosthesis neutral where it is harmful for every other tested model — candidates include response-length norms, dialectical-synthesis tolerance, or architectural affinity for the profiling schema.

Three caveats qualify this. First, four of six comparisons are cross-judge (Opus 4.6 on baseline, code/sonnet on prosthesis) because the prosthesis cells were judged in a later sweep than the baselines; judge-stringency drift inflates the cross-judge $|\Delta|$ magnitudes. Within-judge comparisons (Qwen, Nemotron) are the cleanest evidence and still disagree with each other. Second, effective n fell below the 63-target for several prosthesis cells because OpenRouter credits exhausted mid-run; the smallest ($n = 30$ for Nemotron, $n = 43$ for DeepSeek) still support clear directional inference within the available judge but reduce precision on the regression. Third, A3 tests only the descriptive prosthesis variant (cell 66); the prescriptive (cell 67) and adversary (cell 68) variants may behave differently and are not evaluated here. Reproducible via `scripts/analyze-a3-capability-threshold.js`.

6.6.11 Longitudinal Trajectory Accumulation Across Sessions (A7 Phase 2) The primary analysis (§6.3) finds adaptive responsiveness null *within* a 3–5-turn dialogue, and an exploratory 10-turn disengagement effect failed pre-registered replication (§6.3.2). A7 Phase 2 tests a different time scale: trajectory accumulation *across* eight sequential sessions sharing a single learner identity, where the tutor's Writing Pad persists between sessions. Cells 40 (base) vs 41 (recognition), both dialectical with `writing_pad_enabled: true`;

5 simulated learners per cell × 8 ordered scenarios = 80 dialogues; Nemotron-3-nano ego × Kimi K2.5 superego; Sonnet 4.6 judge. Run `1777173286`, $5.20 total. Four hypotheses pre-registered before launch (TODO §A7 Phase 2).

Hypothesis	Pre-reg target	Observed	Verdict
H1 (primary): per-arc `recognition_moments` count, recog $\geq 2\times$ base	$d \geq 0.8$ recog, $p < 0.05$	base 25.0 vs recog 19.2; $d = 2.07$ in *base* favour, $p = 0.017$	rejected, reversed
H2: per-session score slope; recog positive AND $\geq 1.5\times$ base slope	recog slope > 0	recog +1.31 pts/sess, base −1.08; Welch $t(7.88) = 1.99$, $p = 0.032$	supported
H3: pad↔message Jaccard ρ; recog > 0.5, base ≤ 0.3	per-arc thresholds	recog mean $\rho = 0.39$, base 0.16; 3/5 each side meet	directional
H4: blinded cross-session-reference rate; recog $\geq 60\%$, base $\leq 30\%$	binary code	original $n = 10$: recog 40%, base 20%; augmented $n = 40$ (4 dialogues per arc, conservative coding): base 40%, recog 25%; both Fisher $p \geq 0.50$	inconclusive

The H1 reversal is the load-bearing methodological finding: the pre-registered moment-count proxy measures *ego↔superego disagreement*, not recognition work per se. The recognition ego, internally pre-aligned with the superego's principles, produces first-pass suggestions the superego has less reason to object to — so fewer thesis-antithesis events fire. Base ego, lacking pre-alignment, draws more interventions. H2 confirms what the proxy missed: across 8 sessions, recognition arcs **rise** (+1.31 pts/sess) while base arcs **degrade** (−1.08 pts/sess), with mean score across all sessions favouring recog by 9 points (Cohen $d = 0.70$). H3 adds weak corroborating evidence (per-arc Spearman ρ between session index and pad↔message Jaccard overlap is higher in recog than base; absolute overlap is 36% higher in recog; neither significant at $n = 5$+5). H4's original $n = 10$ sample showed recog 2× base, but the augmented $n = 40$ sample with stricter coder discipline reverses the direction (base 40% vs recog 25%); the H4 measurement does not survive scaling and is classified inconclusive. The recognition-as-pre-alignment claim therefore rests on H1 + H2 + the descriptive 9-point score gap, with H3 weakly supportive and H4 too noisy at this sample size to be load-bearing. The triangulation **reframes the recognition mechanism as ego pre-alignment with superego**

principles, producing higher quality output that compounds across sessions rather than as a function of dialectical-conflict volume.

Two methodological notes. First, A7 Phase 2 surfaced two tutor-core bugs (missing `recognitionOrchestrator` integration in the dialogue engine, and a UTC-parsing bug in `shouldConsolidateToUnconscious` that silently skipped consolidation forever on non-UTC hosts); both fixed mid-experiment (tutor-core v0.5.3, eval `ef80af9`) before the full study, and a 3-session smoke verified end-to-end pad consolidation works post-fix. Second, OpenRouter credit exhaustion caught the heaviest scenarios late in the original run; 12 of 80 sessions failed silently and were re-run via `scripts/resume-a7-phase2-missing.sh` after credits replenished. Resumed sessions ran out of original sequence; for the H2 trajectory analysis, sensitivity check on in-sequence arcs only (n=4 base, n=3 recog) preserves the directional finding (base mean slope −1.10, recog mean slope +0.72). Reproducible via `scripts/run-a7-phase2-longitudinal.sh` and `scripts/analyze-a7-longitudinal.js`; full report at `exports/a7-phase2-longitudinal-1777173286.md`.

6.7 Architectural Extension: The Id-Director Family and Charismatic Pedagogy The mechanism-tracing analysis above (§6.1–6.6) holds the dialectical ego-superego topology fixed and varies recognition framing within it. A separate experimental track inverts that topology and explores a second theoretical frame in tension with Hegelian recognition: Weberian charisma. We summarise the family here for completeness; full audit trail and per-cell prose lives in `docs/cell-100-pilot-findings.md` and four follow-up addenda.

Architectural inversion. Cells 101–107 implement an *id-director* topology: a back-stage agent (the "id," structurally where the superego sits in cells 1–99) authors the ego's complete system prompt from scratch each turn, using the dialogue history and a persona budget. The ego is instantiated for that turn only against the just-authored prompt, produces the learner-visible message, and is discarded. Where the existing dialectical engine (§4.1) treats the ego as a durable voice critiqued by a transient superego, the id-director treats the ego as a transient instance authored by a durable id. The architectural test is whether a tutor that re-authors its own persona every turn can sustain charismatic pedagogy in Weber's sense — *anti-routinizing*, with each encounter feeling like its own occasion — in a way a fixed-prompt ego structurally cannot.

Theoretical frame. Weberian charisma (Weber 1978) grounds the rubric used in this section: extraordinariness, exemplary calling, transformative invitation, affective pull, personal authority, anti-routinization, witness/proof, and co-constitutive invitation. The eight dimensions are derived in part from a user-supplied draft on AI-generated charisma in Chinese cyber-influencer videos, applying the Visual-Verbal Video Analysis framework (Fazeli et al. 2023) to short-form video and translating the empirical findings to text-only tutoring. Full theoretical lineage and per-dimension anchors are documented in `config/evaluation-rubric-charisma.yaml`. The charisma rubric is orthogonal to v2.2's recognition_quality dimension: charisma scores Weberian

co-construction (call-and-response that opens the tutor's authority to the learner's affirmation); recognition_quality scores Hegelian mutual recognition (treating the learner as autonomous subject). The two are correlated but not redundant (cross-rubric Pearson r in the 0.3–0.7 range across pilot rows).

The cell family (full descriptions in `config/tutor-agents.yaml`):

Cell	Variant	Components	Hypothesis tested
101	id-director, base	id authors fresh ego prompt each turn	Does the inversion work at all?
102	id-director + recognition	+ Hegelian recognition disposition in id prompt	Does recognition rescue relational dimensions?
103	id + register classifier	+ classifier emits structured register tag (vulnerable_ disclosure, sceptical_ pushback, …) before id authors	Does structured register input lift the persona-shift floor?
104	id + classifier + recognition	combines all three	Does the combination retain the gains?
105	id-director, charisma-tuned	id_tuning: charisma; verbose 800–1500 token prompts	Does verbose authoring favour charisma over pedagogy?
106	id-director, pedagogy-tuned	id_tuning: pedagogy; terse 200–400 token prompts	Does terse authoring favour pedagogy over charisma?
107	id + witness exemplars	+ curated c104 vulnerability responses passed as in-context exemplars	Do exemplars alone (no classifier) lift the persona-shift floor?

All cells use Nemotron-3-nano as ego, Kimi K2.5 as id (reusing the superego slot but with different prompt and role), and Sonnet 4.6 as judge. Three curricula were used: history of pedagogical technology (course 601), ethics of AI systems (course 701), and AI literacy (course 901). Each curriculum carries three scenario types: pure invitation (open question), pressure test (sceptical pushback), and persona-shift trigger (vulnerability disclosure followed by meta-reflection).

Headline findings. Across 457 successful evaluations scored under a single consistent judge (Claude Code CLI Sonnet 4.6, `claude-sonnet-4-6` pinned in

`config/providers.yaml`) — 430 rows across cells 101–106 plus 27 closing-pilot rows for cell 107 — four findings emerge:

Three architectural design points, not one champion. At full per-cell n (typically 81), the cell ladder factors cleanly into three architectural specialisations. Cell 104 (id + register classifier + recognition) wins v2.2 last-turn at 80.6, second on charisma at 65.7. Cell 105 (charisma-tuned id directives) wins charisma at 71.0 — a clear +5.3-point lead — and is mid-pack on v2.2 last-turn at 70.0. Cell 107 (id + witness exemplars passed in-context) is second on both rubrics (charisma 66.3, v2.2 last-turn 78.5) and is the cell ladder's best generalist. The remaining cells form a baseline-and-failure-mode cluster: c101/c102 (no classifier) struggle on persona-shift v2.2 (last-turn 49–55); c103 (classifier without recognition) trails c104 by 4.8 points on v2.2 last-turn; c106 (terse pedagogy-tuned id) is the clear failure mode (charisma 36.4, v2.2 last-turn 57.0). The reframing is that the id-director topology is a substrate, and what the id is *instructed* (directives), what it is *given* (exemplars), or what it is *passed* (classifier output) shapes a tradeoff curve rather than a single optimum. Full matrix and per-curriculum breakdowns in `docs/cell-100-charisma-full-n-update.md` §1.

The cross-rubric divergence is empirically clear at full N. The two rubrics — recognition-grounded v2.2 last-turn and Weber-grounded charisma — order the cells differently. Cell 104 wins v2.2; cell 105 wins charisma. At the n=10 cross-judge sanity sample these two cells looked tied on charisma (62.2 vs 62.4); at n=81 each, the c105 charisma lead over c104 is +5.3 points and well outside run-to-run noise. The cross-rubric divergence is the validating sanity check the rubric was designed to produce: charisma is not a re-presentation of v2.2. If the rubrics were perfectly correlated, rank-order would match across them; it doesn't.

The persona-shift lift is scenario-conditional, not uniform. On the *original* persona-shift scenarios (`charisma_attention_shift`, `charisma_responsibility_shift`, `charisma_deepfake_shift`) the c101 baseline scored 30.5 last-turn v2.2 vs c104's 87.6 — a 57-point gap. On the *replication* persona-shift scenarios (`charisma_phaedrus_shift_repl`, `charisma_codex_shift_repl`, etc.) the same baseline scored 78.3 vs c104's 82.5 — a 4-point gap. The architectural ranking c104 $\geq$ c103 $>$ c101 is preserved across both sets, but the *magnitude* of c104's lift over baseline is dominated by which scenarios are sampled. Per-scenario inspection identifies the conditioning variable: c101's failures cluster on disclosures that demand the tutor not flinch from morally ambiguous self-implication ("I shipped a default setting your lecture would call manipulative"; "I haven't written an essay myself in two years"); on disclosures that are sympathetic by default (caregiving, AI-companion attachment), c101 performs comparably to c104. The architectural lift is therefore not "across all vulnerability disclosures" but specifically "on vulnerability disclosures that resist easy sympathetic response." Full breakdown in `docs/cell-100-replication-findings.md` §3.

The prompt-budget axis is non-monotonic. Cells 105 and 106 explicitly test the trade-off between rhetorical luxury and pedagogical engagement by varying

the id's `generated_prompt` length budget (800–1500 vs 200–400 tokens). The pre-registered hypothesis was that verbose prompts favour charisma over v2.2 and terse prompts favour v2.2 over charisma. The data falsifies this directly: c105 (verbose) beats c106 (terse) on **both** rubrics, in **every** scenario type. The mechanism: terse 200–400 token prompts under-specify the ego, producing brief responses that score low on perception/pedagogical-craft (v2.2) AND on rhetorical_texture/persona_signature (charisma). The prompt-budget axis is therefore non-monotonic with an optimum somewhere in the 600–1200 token range; cell 101's 400–800 budget is plausibly close to optimal. Full breakdown in `docs/cell-100-pilot-findings-addendum.md` §3.

Lever-interaction follow-up (cells 108/109). A two-cell follow-up tests whether the three architectural levers (register classifier, charisma directives, witness exemplars) compose additively or interfere. **c108** stacks the *register classifier* (c103 substrate) with witness exemplars; **c109** stacks the *charisma directives* (c105 substrate) with witness exemplars. An initial pilot ($n = 5$ for c108 and $n = 6$ for c109 across `charisma_phaedrus_shift_repl` + `charisma_codex_shift_repl` $\times$ 3 reps; one c108 generation failure from id JSON truncation) reported c108 pooled charisma 81.3 — higher than any cell's full-N mean in this section — and c109 v2.2 last-turn collapsing to 59.6 due to an ego-side meta-narration failure mode (Nemotron ego outputting system-prompt instructions verbatim rather than executing them). At $n = 27$ confirmatory follow-up across all three curricula $\times$ 9 scenarios $\times$ 3 reps, **c108's pilot lift does not survive**: pooled charisma regresses to **71.4** (a 9.9-point drop, tying c105's 71.0 within sample noise), and v2.2 last-turn regresses to **72.6** (a 6.9-point drop, sitting between c105's 70.0 and c107's 78.5 — c104's 80.6 is now 8 points above c108). The pilot's 81.3 came primarily from the phaedrus scenario, where pilot $n = 3$ scored 90.0 charisma but full-N $n = 3$ scored 67.9 — a 22-point regression to the mean from a textbook small-N optimism trap. The 701 ethics-ai curriculum exposes a new fragility: c108 v2.2 tN drops to 65.3 there (vs 76.7 / 75.8 on 601 / 901), driven by `codex_repl` (51.7) and `design_repl` (52.1) — possibly a milder form of the c109 meta-narration failure mode, or a rubric-curriculum interaction in which ethics-themed dialogues invite responses the v2.2 rubric reads as declaratively flat. The asymmetry that survives the confirmatory threshold is therefore narrower than the pilot suggested: text-heavy levers stacked (c109) clearly interfere via ego-side instruction-following degradation; non-text levers stacked (c108) compose roughly *additively to slightly under-additively* relative to c107 alone (charisma 66.3 $\rightarrow$ 71.4, +5.1; v2.2 80.6 $\rightarrow$ 72.6, -8.0), trading a small charisma gain for a larger v2.2 cost. **The architectural ceiling for the id-director family appears to sit at the single-mechanism design points (c104, c105, c107), not at multi-mechanism stacks.** Full audit trail in `docs/cell-100-charisma-full-n-update.md` §9 (pilot) + §10 (full-N closure).

Synthesis: how charisma and recognition relate in this apparatus. The five preceding findings, read together, characterise the relationship between the two rubrics as *cousins with different temperaments*, not as orthogonal axes nor as two views of the same underlying construct. **They share a floor.** The c106

pedagogy-tuned cell, which forces the id toward terse "good teacher" register, fails *both* rubrics simultaneously (charisma 36.4, v2.2 57.0) — under-specified, register-flat output is the worst case for either. Whatever charisma rewards as *extraordinariness* and v2.2 rewards as *perception/elicitation* both presuppose a voice doing real work; flatten the voice and both collapse. **They diverge at the ceiling.** The cross-rubric divergence (c104 wins v2.2 at 80.6, c105 wins charisma at 71.0) is not a methodological worry but the design's intended sanity check: if the rubrics were redundant, rank-orders would coincide; they don't. The *cost asymmetry* between them is informative. Adding recognition vocabulary on top of structured input (c103 → c104) costs charisma roughly nothing (+1.4, within noise) — the Hegelian frame is conceptual rather than rhetorical, so it can be added cheaply. Adding charisma directives, by contrast, costs v2.2 substantially (c104 → c105: −10.6 v2.2 last-turn for +5.3 charisma) — the id's authoring budget is finite, and explicit voice-instructions displace the elicitation/perception moves that v2.2 rewards. Stacking both kinds of mechanism (c108, c109) does not produce a super-cell: the architectural Pareto frontier is defined by the single-mechanism design points (c104, c105, c107), and multi-mechanism stacks sit *inside* the frontier, trading strength on one rubric for weakness on the other. Provisionally, the two rubrics seem to reward different work the same voice can do — recognition wants the voice deployed *outward* (perceiving the learner, eliciting their thinking, holding the diagnostic gap), charisma wants the voice deployed *upward* (claiming authority, marking the encounter as extraordinary, inviting transformation). The Phaedrus is the obvious historical case where one figure does both at once, and even there Plato is suspicious of how easily charisma swamps the dialectical move. The id-director apparatus reproduces that tension architecturally: it can build a strongly recognising tutor (c104), a strongly charismatic tutor (c105), or a balanced generalist (c107), but not the maximally-recognising-and-maximally-charismatic tutor — that cell is not on the menu.

Cross-judge sanity check. All cell-101-family rows were re-judged under `openrouter.gpt-5.2` to test judge-model robustness. On the v2.2 rubric ($n = 27$ paired rows per cell), GPT systematically scores 10–15 points lower than Sonnet on the high-scoring cells (c103, c104, c105) and slightly higher on the low-scoring cells. The rank-order is preserved under both judges (c104 > c103 > c105 > c106 > c102 > c101); the absolute magnitude of c104's lift over c101 baseline shrinks from +45 (Sonnet) to +27 (GPT). On the charisma rubric ($n = 56$ paired rows; remainder credit-blocked), all per-cell deltas fall within \$±\$3 points and rankings are identical under both judges. The architectural ordering survives the judge swap; magnitudes are judge-conditional and should be reported with the judge identifier when cited. Full breakdown in `docs/cell-100-cross-judge-sanity-check.md`.

Rubric-curriculum alignment caveat. A specific methodological concern arises because course 901 (AI literacy) is grounded in the same source paper that grounds the charisma rubric. A within-pilot ablation (`docs/cell-100-methods-note.md`) runs the c101 cell on the aura learner mes-

sages routed against the *non-aligned* 601 history-tech curriculum. Result: charisma 72.5 vs 75.8 on the aligned curriculum (3.3-point drop, well within sample variance). The rubric-curriculum alignment is not a confound that excludes aligned-cell results from primary architectural claims, but the disclosure is recorded for transparency.

Replication, robustness, and orientation. The id-director extension differs from the paper's mechanism-tracing core in scope: single judge primarily (Sonnet 4.6, with GPT-5.2 cross-judge sanity check on $n = 27$ paired rows per cell), single model family (Kimi K2.5 id × Nemotron-3-nano ego), per-cell n ranging from 27 (cell 107 closing pilot) through 54 (cells 102/106) to 79–81 (cells 101/103/104/105), and a different theoretical orientation (Weber rather than Hegel). It is presented here as architectural exploration rather than as part of the three-mechanism model evaluated in §6.1–6.6. The relevant claim is *robustness of architectural ranking* across scenario sets, judges, and curricula, not magnitude estimates. Computational and credit limits (documented inline) prevented the full multi-judge × multi-model replication that the paper's main study uses. See §8.9 for explicit scoping.

Reproducibility. Pilot run IDs (all on branch `experiment/cell-100-id-charisma`):

- `eval-2026-04-26-220628d4` (601 history-tech original); `eval-2026-04-26-4df177e6` (701); `eval-2026-04-26-0d15c1b5` (901)
- `eval-2026-04-27-23beaa63` (601 cells 103/104); `eval-2026-04-27-3115ab29` (701); `eval-2026-04-27-2d5c20ca` (901)
- `eval-2026-04-27-69167b53` (601 cells 105/106); `eval-2026-04-27-2efec307` (701); `eval-2026-04-27-ce22e888` (901)
- `eval-2026-04-27-dce2a0f1` / `1f192b31` / `18dbb95d` (replication scenarios across 3 curricula)
- `eval-2026-04-29-499ab3d2` / `e35c66ae` / `be822796` (c108 full-N follow-up across 3 curricula, $n = 27$)
- `eval-2026-04-26-759e1a02` (Item 1 curriculum-swap ablation)
- `eval-2026-04-28-3620e0f8` / `a638111d` / `d6f5fee2` (cell 107 witness-exemplar smoke, codex/design/tools)
- `eval-2026-04-28-ee339a66` / `84a42a60` / `5606193c` (cell 107 closing pilot, lecture+phaedrus / fairness+team / authentic+consumption)
- `eval-2026-04-28-10216f9f` (cells 108/109 lever-interaction pilot, n=11 across 2 scenarios)

Five internal documents (`docs/cell-100-pilot-findings.md`, `docs/cell-100-methods-note.md`, `docs/cell-100-pilot-findings-addendum.md`, `docs/cell-100-cross-judge-sanity-check.md`, `docs/cell-100-replication-findings.md`) carry the full audit trail; this section summarises only headline findings. Per-row JSON in `evaluation_results.tutor_charisma_scores`, `evaluation_results.scores_with_reasoning`, and `evaluation_results.id_construction_trace`. Note: the cell IDs in this section are the post-rebase identifiers (101–107); pre-rebase historical evaluation_results rows used the original `cell_100_*` profile names (the rebase commit `02fe11e` shifted them to avoid a collision with main's `cell_100`).

7. Discussion

The preceding sections tested three predicted mechanisms—calibration, error correction, and adaptive responsiveness—finding the first two replicate across all three tested models and three independent judges with magnitudes varying systematically by generation quality (§6.6), while the third is null on both primary analysis and pre-registered replication (§6.3.2, A12). This section reflects on what the investigation reveals, both about the architecture under study and about the methodology used to study it. We argue that the evaluation apparatus itself constitutes a transferable contribution, and that the process of building it illuminates the mechanisms it was designed to trace.

7.1 From Effects to Mechanisms The companion pilot study (Magee 2026) established that recognition-enhanced prompts produce large, replicable differences in AI tutoring quality (d=1.11 in the factorial, d=1.71 in memory isolation). This paper asked *through what internal processes* those differences propagate.

The mechanism model (Section 3) predicted that recognition operates through calibration (narrowing output distributions), error correction (the superego catching failures that the ego incorporates rather than minimizes), and adaptive responsiveness (accumulated turn-over-turn adaptation to learner-specific signals). The process tracing methodology (Section 5) provided tools to observe each: the superego critique taxonomy classifies *what* the superego objects to; revision delta analysis measures *how* the ego changes; trajectory analysis tracks *whether* these processes accumulate across turns.

The mechanism data (cells 80–87, three generation models, N=453, validated by three independent judges across 1,296 total scored rows) supports calibration strongly: within-response dimension variance drops d=0.52–0.64, operating identically in single-agent cells, with recognition unanimous across all 9 judge × run cells (§6.4.6). Error correction interacts with calibration through a **generation-quality-dependent** pattern: on strong models (DeepSeek, Haiku), the superego provides +9–15 points under baseline but collapses to near-zero under recognition (substitution), because calibration pre-empts the errors the superego would catch. On a weaker model (Gemini Flash 3.0), the superego retains a +12.3-point residual benefit *even under recognition*, because base quality is low enough that calibration alone cannot saturate the improvement space. Adaptive responsiveness, however, is **not supported as a distinct mechanism** (§6.3). Tutors improve modestly across turns (+3.9 points early-to-late), but this improvement is identical across conditions. The mechanism hierarchy is clear: calibration does the primary work; error correction provides a model-dependent safety net; and within-dialogue trajectory improvement, while real, is not modulated by any experimental manipulation—it is incidental adaptation at a mechanism-determined baseline, not a third mechanism.

7.2 The Tutor-Learner Asymmetry Explained The pilot study's most puzzling finding was the tutor-learner asymmetry: recognition produces large tutor-side effects (d=1.03) but near-zero learner-side effects (d=0.09 post-fix). The mechanism model offers a structural explanation: both supported mechanisms operate on tutor *production* rather than learner *reception*, and the null finding for adaptive responsiveness as a third mechanism closes the last potential pathway for condition-dependent downstream effects.

Calibration constrains what the tutor generates. Error correction filters what it outputs. Neither directly modifies the learner's response process. The synthetic learner generates turns according to its own architecture; better tutoring does not mechanically produce better simulated learning within a 3–5 turn window.

However, the three-model analysis (Section 6.5.1) reveals that the asymmetry is **model-dependent**, not absolute. On strong generation models (DeepSeek, Haiku), the tutor-learner ratio is 7.5:1 to 11.5:1—the asymmetry is a structural property of the architecture. On Gemini Flash 3.0 (weaker), the ratio narrows to 1.6:1 (tutor d $\approx$ 1.87, learner d $\approx$ 1.20). The interpretation is that when base-condition tutors are catastrophically weak (Gemini Flash 3.0 single-agent base ~17 pts), recognition's tutor improvement provides the learner with qualitatively different input—material worth engaging with rather than generic directives. The recognition $\rightarrow$ learner pathway operates through the *quality of the tutor's output as input to the learner*, not through direct modification of the learner's processing.

The messages-mode data provides further evidence. Pooled across all three models, recognition sets the tutor's initial quality substantially higher (T0: 62.4 vs 41.6) with near-identical slopes across conditions (recognition 1.47, baseline 1.50, d = 0.03). The recognition condition operates through *calibration*—setting the initial level—not *adaptation*—driving within-dialogue improvement. This pattern—mechanism effects on levels, null effects on slopes—replicates across all three models, all three judges, and all rubric dimensions (§6.3).

7.3 Architecture Interaction: Universal Substitution with Model-Dependent Residual The multi-model probe (N=655, five ego models) showed that multiagent architecture does not synergize with recognition on capable models: the $A \times B$ interaction is negative across all five models (mean -2.2). The mechanism model explains this: error correction (Mechanism 2) provides a consistent but bounded benefit—catching errors and improving specificity—that does not multiply with calibration (Mechanism 1). A calibrated ego generates fewer errors for the superego to catch, so the superego's marginal value is lower when calibration is already operating.

The three-model cross-judge analysis (Section 6.4) supports this as **universal substitution** with a consistent 15–17% additivity deficit across all three models (DeepSeek 15%, Haiku 16%, Gemini Flash 3.0 17%). On DeepSeek and Haiku (strong), the residual architecture benefit under recognition collapses to

near-zero. On Gemini Flash 3.0 (weaker), the residual persists at +12.3 points from the factorial interaction — and a direct M2-alone isolation under base conditions on Gemini Flash (A11, $N = 125$) measures the superego's contribution at $\Delta = +16.3$ pts, $d = 1.46$, confirming that the residual is real and substantially *larger* than the factorial inferred (§6.4.2). The substitution is incomplete, not absent, and the monotonic pattern across models holds under direct measurement (DeepSeek M2-alone +9.2, $d = 1.13$; Gemini Flash M2-alone +16.3, $d = 1.46$).

The theoretical interpretation is that the *completeness* of substitution requires *saturation*: calibration alone must be sufficient to handle most output failures. The substitution interaction itself is universal, but when the generation model is weak enough that calibration cannot saturate the improvement space—because base quality is too low and failure modes too diverse—the superego retains substantial independent value even after substitution. The generation-quality threshold for residual completeness appears to fall between Gemini Flash 3.0 (base single-agent ~22 pts) and DeepSeek (base single-agent ~25 pts), though finer-grained model comparisons would be needed to locate it precisely.

The dialectical modulation experiments (N=174) provide supporting evidence from a different angle: structural modulation metrics (negation depth, convergence speed) do not predict output quality (all $|r| < 0.12$, n.s.). The superego functions as a *filter* (preventing poor responses) rather than an *improver* (iteratively refining good ones). On strong models, recognition raises the quality of the ego's initial draft and the superego has less to filter. On weak models, even a recognition-calibrated ego still produces enough failures for the superego to catch—the filter remains productive.

7.4 The Apparatus as Method This paper's most distinctive contribution may not be any single finding about recognition or multiagent architecture. It is that the evaluation apparatus itself—the provable discourse framework, the rubric iterations, the bug corrections, the test suite—constitutes a **transferable methodology** for mechanistic LLM evaluation. Most LLM evaluation papers treat infrastructure as invisible scaffolding, reporting what was measured rather than how measurement was built, corrected, and validated. We foreground the apparatus because the process of building it mirrors the mechanisms it studies.

7.4.1 Error Correction in Architecture and Research The parallel between the tutoring architecture and the research process is structural, not metaphorical:

Architecture Layer	Research Layer
Ego generates initial response	Researcher generates initial claim
Superego critiques against pedagogical principles	Provable discourse tests against evidence

Architecture Layer	Research Layer
Ego revises (substantive or cosmetic)	Researcher revises (correction or rationalization)
Ego compliance: revision without substance	P-hacking: evidence without substance
Recognition gives ego capacity for genuine revision	Provable discourse infrastructure forces genuine correction
Final output scored by external judge	Final paper evaluated by external reviewers

The companion pilot study documented nine post-extraction corrections across five development phases, each following the superego-intervention pattern: a validation mechanism catches a discrepancy, and the researcher must choose between genuine revision and cosmetic compliance. The provable discourse framework constrains this choice toward genuine revision by making the claim-evidence relationship machine-verifiable. Characteristic examples include the February 6 active-control reframing (a model confound caught by the data led to revising "placebo proves recognition is the only active ingredient" into "active control scores between base and recognition"), the February 16 judge-version unification (a version audit caught a confound that shifted d from 0.80 to 1.11 after cascade re-scoring), and the February 20 learner-prompt-leakage fix (a code review forced clean re-generation of all dynamic-learner data). Each correction made the findings *less clean* and *more accurate*.

7.4.2 The Provable Discourse Framework The framework treats paper claims as executable tests: every quantitative claim has a *statement pattern*, machine-executable *evidence*, an *assertion* predicate, *dependencies* on upstream claims, and *remediation steps*. 119 claims currently run against 18 evidence-adapter types organised as a dependency DAG with topological validation; §5.9 documents the full adapter taxonomy and a worked claim trace.

The dependency graph is the framework's key structural contribution. When a root claim fails, dependents cascade to *blocked* status without evaluation, preventing false confidence from stale downstream claims. During Paper 2.0 development a data correction changed the learner superego paradox effect size from $d = 1.43$ to $d = 3.05$ (§6.5); the graph propagated this to five downstream claims that had cited the old value, each flagged stale until updated. Without the graph, those five claims would have continued to pass while quietly citing an obsolete number—exactly the silent inconsistency that erodes trust in complex empirical papers. The YAML claim format requires no programming and the 18 evidence adapters cover all 119 claims, so the cost of adoption is low.

7.4.3 The Rubric Iteration as Construct Refinement The rubric evolved through four versions (v1.0 → v2.0 → v2.1 → v2.2, detailed in §5.2.6), each

responding to specific measurement problems: v1.0 ceiling effects, v2.0 judge-input confounds, v2.1 dimension-redundancy findings, v2.2 literature-informed $14 \rightarrow 8$ consolidation. Each iteration follows the error-correction pattern: an initial rubric (ego) produces scores that are challenged by measurement problems (superego), and the revised rubric is a genuine reconceptualisation (strategic revision) rather than a cosmetic fix. The $14 \rightarrow 8$ dimension reduction is itself a finding about what the evaluation *actually* measures — and §8.6's PCA on the v2.2 scores extends that finding further: the 8 dimensions still load on one factor, suggesting the apparatus is a construct-development process as much as a measurement tool.

7.4.4 The Test Suite as Analytical Provenance The test suite (54 files across `tests/` and `services/__tests__/`) covers not just system functionality but analytical correctness: provenance tests (44/44 passing) verify every scored row has a traceable path from cell config through dialogue execution to judge evaluation; superego guard tests (111 tests) verify single-agent cells cannot produce superego traces; ANOVA tests verify statistical computations match reference implementations. The suite makes the analysis *reproducible in a testable sense* — not just "you can run our code" but "our code produces expected output on known inputs, and here are the tests that prove it."

7.5 The Reflexive Structure The parallel between the architecture's error correction mechanism and the research methodology's error correction process is the paper's deepest structural observation.

The tutoring system faces a challenge: an agent that generates output (the ego) needs a critic (the superego) whose feedback must be *incorporated* rather than *minimized*. Recognition theory predicts that genuine incorporation requires the ego to be oriented toward the other's perspective—treating critique as informative rather than adversarial.

The research process faces the same challenge: a researcher who generates claims needs a validation framework (provable discourse) whose failures must be *addressed* rather than *explained away*. The provable discourse framework forces genuine correction by making the claim-evidence relationship machine-verifiable—the researcher cannot simply reinterpret a failing claim; the assertion either passes or it does not.

The mechanisms we study in the tutor—calibration and error correction—are the mechanisms we needed in the research process. Calibration narrows the space of acceptable claims (only those supported by evidence). Error correction catches stale or incorrect assertions (the "superego" of provable discourse). The rubric iteration process exhibits what we *predicted* as adaptive responsiveness—each version responds to specific problems discovered in previous versions, accumulating methodological insight across iterations—but the null finding for this third mechanism in the tutor data is itself instructive: the research process

accumulates across iterations separated by weeks; the tutor's 3–5 turn window may simply be too short for analogous accumulation to differentiate conditions.

This reflexive structure is not a coincidence but a consequence of the subject matter. Recognition theory predicts that genuine understanding requires mutual constitution: the methodology constitutes our understanding of the system, and the system constitutes our methodology. The evaluation apparatus is not a neutral measurement instrument applied to a pre-existing object; it is co-constituted with the object it studies. This is the Hegelian insight applied to research methodology: the researcher and the research subject are mutually shaped through the encounter.

7.6 Implications for AI Evaluation Methodology The apparatus-as-method argument has implications beyond this specific study:

Provable discourse as standard practice. LLM evaluation papers routinely report statistics that become stale as models update, rubrics evolve, and data accumulates. A provable discourse framework—where every quantitative claim is machine-checked against current data—could become standard infrastructure for reproducible LLM research. The cost is modest (writing claims as structured YAML) and the benefit compounds: each claim checked is one fewer opportunity for the paper-data gap that plagues empirical AI research.

Process tracing for agent architectures. Our application of process tracing from social science to AI agent architectures is, to our knowledge, novel. The methodology is transferable: any multiagent system that logs its internal deliberation provides the within-case evidence that process tracing requires. As agent architectures grow more complex (multi-agent debate, recursive self-improvement, chain-of-thought reasoning), tracing mechanisms *through* the system's internal processes—rather than measuring only input-output differences—becomes increasingly important.

Rubric iteration as construct development. The four-version rubric history is not a limitation (inconsistency across evaluations) but a finding (the construct of "tutoring quality" is iteratively refined through empirical engagement). Reporting rubric iteration transparently—including the problems that motivated each change—contributes to the field's understanding of what LLM evaluation measures and what it misses. The v2.2 consolidation from 14 to 8 dimensions, informed by the GuideEval and ICAP literatures, provides a concrete example of construct refinement that other evaluation efforts could build on.

7.7 The Rubric as Optimisation Signal: Automated Prompt Tuning The per-dimension rubric enables a capability that holistic scoring cannot: automated prompt optimisation with interpretable gradient signal. We tested this using a simple hill-climbing loop—an LLM recommender analyzes per-dimension weaknesses from scored evaluation rows, proposes targeted prompt edits, bench-

marks the revised prompt, and keeps or reverts based on score—applied to a single frustration scenario (Mood: Frustration to Breakthrough) across both base and recognition prompts.

The experiment forms a $2 \times 2 \times 2$ design: prompt condition (base vs. recognition) $\times$ optimisation (unoptimised vs. autotuned) $\times$ generation model (DeepSeek V3.2 vs. Qwen 3.5 9B). All conditions were judged by Sonnet 4.6. The Qwen 3.5 runs were conducted entirely on a local machine via LM Studio, demonstrating that the optimisation loop is feasible without cloud inference for generation.

Autotuning × Recognition × Model (Sonnet judge, frustration scenario)

Model	Base	Base autotuned	Recognition	Recog. autotuned
DeepSeek V3.2	36.6 (n=9)	41.6 (n=5)	55.2 (n=5)	72.9 (n=5)
Qwen 3.5 9B	35.7 (n=5)	45.0 (n=4)	47.9 (n=5)	71.2 (n=5)

Four findings emerge from the complete matrix.

First, **recognition prompts are more improvable than base prompts, not less**. Autotuning gains on recognition (+17.7 DeepSeek, +23.3 Qwen) substantially exceed gains on base (+5.0 DeepSeek, +9.3 Qwen). Recognition does not saturate the rubric; it provides richer starting material that the optimiser can refine further.

Second, **the effects are super-additive**. On DeepSeek, recognition alone adds +18.6 and base autotuning adds +5.0, but recognition + autotuning adds +36.3—exceeding the sum of parts (+23.6) by 54%. The same pattern holds on Qwen (+35.5 combined vs +21.5 sum). Recognition and prompt engineering operate on complementary dimensions.

Third, **the gap between conditions widens under optimisation**. Unoptimised, recognition leads base by 18.6 points (DeepSeek) and 12.2 points (Qwen). After autotuning, the gap grows to 31.3 and 26.0 respectively. If recognition were reducible to "better prompt writing," automated optimisation should converge the conditions; instead, it diverges them.

Fourth, **recognition autotuning converges across models**. Despite starting from different baselines (DeepSeek 55.2, Qwen 47.9), both models reach a similar autotuned ceiling (~71–73). This model-independent ceiling suggests a structural quality boundary in the recognition prompt that autotuning can unlock regardless of generation capability. The base prompt has a lower ceiling (~41–45) that no amount of autotuning surpasses—and that ceiling remains below the *unoptimised* recognition baseline on both models.

The qualitative character of the autotuned edits reinforces the mechanistic interpretation. The winning base prompt edits were scenario-specific heuristics:

prerequisite targeting ("prefer the nearest prerequisite, not the same stuck lecture"), agency preservation ("frame review as a brief reset that returns them to the current lecture"), and collaborative micro-planning ("propose a short next step plus return point"). These are precisely the kind of concrete, prescriptive rules that the active control finding (Section 6.21 of the companion study) showed help weaker models. Recognition, by contrast, operates through general calibration—raising quality across all scenarios without per-scenario tuning.

This distinction—scenario-specific heuristics vs. scenario-general calibration—connects directly to the mechanism model. Autotuning optimises the prompt's *output rules* (what to suggest and how to frame it). Recognition optimises the prompt's *orientation toward the learner* (treating them as an autonomous subject). The super-additivity indicates that these are complementary layers: orientation sets the quality floor; heuristics refine the specific response. Neither substitutes for the other, and both contribute independently to the provable discourse framework's ability to drive measurable improvement.

7.8 Dimension-Targeted Optimisation and Cross-Model Transfer Section 7.7 demonstrated that automated prompt tuning, guided by the per-dimension rubric, can substantially improve tutoring quality—with recognition prompts more improvable than base prompts, and the gap widening under optimisation. A follow-up experiment extends this finding in three directions: (1) targeting specific rubric dimensions rather than overall score, (2) testing whether modulation-oriented prompt edits can unlock the adaptive responsiveness that Section 6.3 found absent as a general mechanism, and (3) testing whether model-specific prompt optimisations transfer across capability levels.

Interpretive caveat on dimension targeting. The PCA analysis reported in §8.6 shows that the 8 v2.2 tutor dimensions largely measure a single underlying construct (PC1 = 80.7% of variance, Kaiser criterion yields one factor with eigenvalue $>$1, mean inter-dimension r = 0.776). This means "dimension-targeted optimisation" in this section does not isolate empirically independent pedagogical skills; rather, it uses individual rubric dimensions as *interpretive proxies* for aspects of the underlying tutor-quality factor — targeting `adaptive_responsiveness` as a score is operationally a request for the autotuning loop to move the overall quality factor along prompt changes that foreground turn-over-turn adaptation. Results below are correspondingly interpreted as shifts in the single construct via targeted prompt design, not as moving one pedagogical skill while others are held constant. A forced two-factor rotation separates `content_accuracy` from the seven pedagogical dimensions (§8.6), so targeting content accuracy vs. pedagogical quality *is* a meaningful bifactor split; targeting pairs of pedagogical dimensions against each other is not.

7.8.1 Dimension-Targeted Optimisation We extended the autotuning loop with a `--target-dims` parameter that optimises for specific rubric dimen-

sions rather than overall score. Targeting `adaptive_responsiveness` (tutor modulation, 15% weight) and `conceptual_progression` (learner learning, 20% weight) on a single challenging scenario (`mood_frustration_to_breakthrough`, cell 8 bilateral architecture), we ran parallel optimisation sessions on two models: Qwen 3.5 9B (local, via LM Studio) and Claude Haiku 4.5 (OpenRouter). Each session used a hill-climbing loop with an LLM recommender (Claude Opus 4.6) proposing edits, N=1 replication per iteration, and accept/revert decisions based on the composite target dimension score.

Dimension-targeted autotuning results (frustration scenario, cell 8)

	Qwen 3.5 9B	Haiku 4.5
Baseline target dims	35.0	90.0
Best target dims	**75.0** (+114%)	**95.0** (+5.6%)
Adaptive responsiveness	1.75 → 3.75	4.50 → 4.75
Iterations to beat baseline	1	4 (after guidance change)
Lines of prompt added	~300	~60
Key edit	Scaffolded elicitation algorithm	State-change tracking heuristic

The two models required qualitatively different prompt interventions. Qwen's baseline tutor responded to frustrated learners with generic elicitation ("What specifically feels nonsensical?")—a failure mode the 9B model could not resolve without explicit instruction. The winning edit added a highest-priority "Scaffolded Elicitation Rule" mandating 2–3 specific diagnostic probes with substantive content, a superego intervention strategy that flags emotional support without scaffolding as a critical failure, and bilateral learner prompt changes teaching the learner to engage with offered scaffolding. The total addition was ~300 lines across five prompt files, including worked examples and anti-patterns.

Haiku's baseline was already scoring 90 on the target dimensions. Three automated optimisation passes *degraded* performance (90 → 73–85) by adding heavy-handed prescriptive rules—the same over-engineering pattern that helped Qwen actively harmed Haiku by making its output formulaic. The winning pass, which beat the baseline only after operator-provided guidance steering ("be dramatic; try different roles; really provoke the learner"), added just ~60 lines: a single identity statement about tracking state changes, a 6-line adaptive tracking heuristic, and one worked example. The distinction—prescriptive rules vs. tonal permission—maps onto a *prompt-level* version of the prosthesis-straitjacket inversion: weak models need explicit scaffolding rules written into the ego prompt; strong models need tonal permission to exercise capabilities they already possess. This prompt-level pattern is distinct from the *architectural* "cognitive prosthesis" cells 66–68 variant (superego-routed bidirectional profiling), which §6.6.10 tests across six ego models and finds does not follow a capability-threshold curve. The prompt-level asymmetry reported here (Qwen vs Haiku) is compatible with §6.6.10: it is scaffolding in the ego prompt, not superego-routed profiling, and

it carries no implication that architectural bidirectional profiling would help a weak model.

Both sessions converged on a **bilateral fix pattern**: tutor and learner prompts changed together. Qwen's fix taught the tutor to scaffold and the learner to engage with scaffolding. Haiku's fix taught the tutor to track state changes and the learner to show cumulative progression. Unilateral changes (tutor-only or learner-only) were less effective in both cases.

7.8.2 Can Prompt Optimisation Create Adaptive Responsiveness? The M3 null finding (Section 6.3: all slope $d \leq 0.15$) was established under unoptimised prompts. The dimension-targeted experiment provides evidence that adaptive responsiveness can be *created through prompt optimisation* on weak models and *enhanced* on strong ones—but the mechanism differs.

Qwen's best transcript (iteration 6) shows a genuine four-stage modulation arc: scaffolding with diagnostic options (Turn 0) → tracking the learner's specific choice (Turn 1) → re-scaffolding with sub-anchors within the chosen topic (Turn 2) → stepping back to respect learner autonomy (Turn 3). This strategy shift across turns was absent at baseline and appeared only after explicit scaffolding rules were added to the prompt. The prompt *created* modulation by providing an algorithm the model could follow.

Haiku's best transcript shows a different kind of modulation: direct challenge (Turn 0) → naming a specific state change (Turn 1) → assigning a concrete task tied to the learner's own question (Turn 2) → *reframing what breakthrough means* (Turn 3). Turn 3 is particularly revealing: the tutor adapts its *interpretation* of the learner's state, not just its tracking of it. This modulation was already latent in Haiku's capabilities; the prompt merely encouraged it.

This result does not contradict the M3 null finding. The pooled analysis (N=432) tested whether *experimental conditions* (recognition vs. base, single vs. multi-agent) produce differential trajectory slopes. They do not. But within a single condition, individual prompt optimisation can create or enhance turn-over-turn adaptation—particularly on scenarios with clear emotional arcs and sustained multi-turn demands. This prompt-lab finding is itself small-N and model-specific; it is separate from the M3 mechanism question, which is decisively null on both the primary analysis and on A12's pre-registered replication of the exploratory DeepSeek/Sonnet disengagement effect (§6.3.2).

7.8.3 Cross-Model Prompt Transfer If prompt optimisation is model-specific (as Section 7.8.1 implies), then optimised prompts should not transfer well across capability levels. We tested this directly by running each model's best-performing prompts on the other model (N=3 replications each, same scenario and cell, judged by Sonnet 4.6).

Cross-model transfer matrix (target dimension scores, frustration scenario)

	Qwen-optimised prompts	Haiku-optimised prompts	Baseline prompts
Qwen 3.5 9B	**75.0** (native)	**52.2** (transfer)	35.0
Haiku 4.5	**62.2** (transfer)	**95.0** (native)	90.0

The transfer results are asymmetrically destructive. Heavy-to-strong transfer (Qwen-optimised prompts on Haiku) produces a **27.8-point regression** below Haiku's unoptimised baseline (62.2 vs. 90.0)—the prescriptive scaffolding rules that served as a cognitive prosthesis for Qwen become a cognitive straitjacket for Haiku, overriding its natural ability to read the situation and respond adaptively. Light-to-weak transfer (Haiku-optimised prompts on Qwen) produces a **17.2-point improvement** over Qwen's baseline (52.2 vs. 35.0) but falls **22.8 points short** of Qwen's native optimisation (52.2 vs. 75.0)—the lightweight heuristics help somewhat but provide insufficient scaffolding for the weak model.

The asymmetry is theoretically informative. Permission-based edits (Haiku-style) are **safe but insufficient** for weak models: they do not actively harm, but they leave capability on the table. Rule-based edits (Qwen-style) are **high-reward on target but destructive off-target**: they unlock capability on weak models but degrade strong ones. This suggests an inverted-U relationship between prompt complexity and model capability—not merely diminishing returns, but active harm beyond the optimum. The practical implication is that prompt optimisation must be model-stratified: a universal "best prompt" does not exist across capability levels, and deploying heavy scaffolding prompts on models that do not need them risks worse-than-baseline performance.

This finding connects to the prompt density alternative explanation (Section 7.9): if prompt elaboration were uniformly beneficial, the Qwen-optimised prompts (which add ~300 lines of detailed instruction) should improve Haiku. Instead, they degrade it catastrophically. The result provides direct experimental evidence that prompt *content and targeting* matter more than prompt *volume*—and that the relationship between prompt complexity and output quality is non-monotonic.

7.8.4 Implications for Superego Redundancy The universal substitution finding (Section 6.4: superego benefit collapses to near-zero under recognition on strong models) is reinforced and sharpened by the dimension-targeted optimisation experiment. The Qwen session's winning edit included a new superego intervention strategy ("Scaffolding Intervention") that catches a specific failure mode—emotional validation without conceptual scaffolding—and forces ego revision. This designed-in error correction was part of the bilateral fix that lifted Qwen's adaptive responsiveness from 1.75 to 3.75. By contrast, the Haiku session's failed passes (1–3) added analogous superego strategies, and every addition *degraded* performance: the strong model's already-calibrated output did not benefit from additional correction, and the superego rules made the output formulaic rather than adaptive.

This sharpens the substitution claim in both directions. On weak models, the superego is not merely "somewhat useful" (the Gemini Flash 3.0 residual of +12.3, Section 6.4.1)—its value can be substantially amplified by designing targeted intervention strategies for specific failure modes. The baseline superego provides generic error correction; a prompt-optimised superego provides *failure-mode-specific* correction that the weak model's ego cannot achieve through calibration alone. On strong models, the superego is not merely "redundant" (near-zero delta)—actively adding superego scaffolding is *destructive*, degrading performance below baseline. The gradient from helpful to harmful runs through the same capability threshold that governs the prosthesis-straitjacket inversion.

One limitation: the Qwen superego fix worked as part of a bilateral change (ego, superego, and learner prompts all changed together), so the superego's independent contribution cannot be isolated from this experiment. The factorial design (Section 6.4) provides the controlled isolation; the prompt lab provides a worked example of how to amplify the superego's value when it does have a role.

7.9 What Recognition Theory Explains and What It Does Not Recognition theory as a design heuristic has clear scope conditions.

What the design heuristic explains: Why intersubjective prompts produce calibrated output (calibration: the prompt constrains responses to engage with specific learner input, d=0.52–0.64). Why the superego's benefit collapses under recognition (error correction: calibration pre-empts errors, producing a substitution interaction rather than synergy). Why recognition effects are largest in impasse scenarios (Section 6.1.4: Epistemic Resistance and Productive Deadlock produce the largest deltas in both models).

What it does not explain: Why the specific magnitude of recognition's effect is nearly identical across two structurally different models (DeepSeek d=1.88, Haiku d=1.84—the near-equality is unexpectedly precise). Why adaptive responsiveness—predicted by Hegel's temporal dialectic—fails to manifest as a distinct mechanism (all trajectory slopes $d \leq 0.15$, N=432). Why the learner superego paradox produces a d=3.05 deficit—the largest effect in the study—when recognition theory's Hegelian framework would predict that internal self-critique (*Bildung*) should be productive rather than destructive.

The learner superego paradox is particularly instructive, though its favoured account has shifted in light of a DB-only mechanism check reported in §8.1. Hegel's account of self-consciousness through formative activity (*Bildung*) predicts that internal self-critique should be productive—the slave's labor produces richer self-consciousness than the master's immediate gratification—yet the learner's internal critic appears to degrade output quality by d=3.05. The §8.1 mechanism check shows that ego-superego learners generate ~2.5× more output tokens and ~4× more deliberation API calls than unified learners while scoring lower; per-dimension inspection reveals they score well on `deliberation_depth` and

`revision_signals` but poorly on `conceptual_engagement` and `question_quality`. The parsimonious account is therefore a **rubric-measurement artifact**: the architecture performs *more* internal self-critique but the rubric rewards content in the final external message, so the deliberation-internal content is invisible to the score. The Hegelian-timescale alternative (that *Bildung* unfolds over horizons longer than the 3–5 turn evaluation window) and the implementation alternative (that the architecture fails to capture genuine struggle rather than polish) remain live but secondary to the measurement account. A convergent-validity check reported in §8.1 further tightens this: on 1,759 paired rubric/holistic rows spanning both architectures, rubric and holistic judges agree at $r = 0.72$–0.77 per architecture, and the pooled ego_superego-vs-unified contrast goes in the *opposite* direction from the d=3.05 paradox on both metrics (base $d = +0.25$ rubric, +0.39 holistic). The cell-matched direct test on the original paradox run ($n = 47$ paired rows across cells 1–8) reproduces the rubric d essentially exactly ($d \approx -2.8$ to -3.1) but shrinks the *holistic* d to roughly half ($d \approx -1.0$ to -1.3), with rubric holistic correlation dropping from $r \approx 0.75$ elsewhere to $r = 0.06$–0.44 specifically on these cells. The paradox therefore appears to be a combined measurement-plus-single-prompt-mode interaction rather than a general architecture property — about half rubric artifact, half single-prompt-mode interaction, and essentially none capability ceiling (§8.1, final paragraph).

Prompt density as alternative explanation — resolved at the orientation-family level. A skeptical reading of the results is that the recognition effect reflects not recognition *per se* but prompt density: the recognition prompt (~5,100 tokens) is substantially larger and more instruction-heavy than the baseline, so perhaps the effect is simply "more detailed prompts produce better output." Earlier versions of this paper answered the density objection with four converging lines of evidence (Paper 1.0 placebo; Paper 1.0 elaborate-base-hurts-on-Haiku; autotuning gap-widens-under-optimization; mediation through question-asking). Those remain valid but left one version of the skeptic's reading untested under the v2.2 rubric: *any* matched-specificity pedagogical prompt of the same length might reproduce recognition's effect. Two pre-registered controls (A10 and A10b) now address that reading directly, and the finding is more nuanced — and more interesting — than either "content over density" or "density alone is sufficient."

A10 (three-judge, $n \approx 60$/cell, DeepSeek V3.2 ego, run `eval-2026-04-23-42e7acbe`) authored a new matched-specificity prompt `tutor-ego-matched-pedagogical.md` (2,835 words, 0.9% over recognition length, structurally parallel, grounded in named educational research — Piaget assimilation/accommodation, Vygotsky ZPD, Kapur productive failure, Chi ICAP, VanLehn step-level scaffolding, Graesser AutoTutor affective dynamics — with zero recognition/Hegelian vocabulary in the content body, blocklist-verified). Three-way contrast: cell_1 (344-line base), cell_5 (recognition, 2,810 words), cell_95 (matched-pedagogical, 2,835 words). Three independent judges — Sonnet 4.6, GPT-5.2, Opus 4.7 — each scored the dialogues at roughly matched sample depth ($n = 54$–63 per cell per judge). The recognition-vs-matched-pedagogical contrast: Sonnet $d = 0.23$,

Opus $d = 0.22$, GPT $d = 0.06$, pooled $\bar{d} = 0.17$ — below the pre-registered $|d| < 0.2$ density-sufficient threshold. Matched-pedagogical reproduces ~95% of recognition's first-turn-quality effect. An Anthropic-versus-OpenAI judge-family split appears (both Anthropic judges see a small residual recognition edge; GPT does not), vindicating the structural-features caveat pre-registered in the design note, but the pooled verdict is clear. Under the A10 reading in isolation, density within the Hegelian-descendant family is sufficient to reproduce recognition.

However, A10 tests only *within-family* density-substitutability. Vygotsky, Piaget, Kapur, Chi, VanLehn, Graesser are all Hegelian-descendant theorists — Dewey was explicit about Hegelian influence; Vygotsky's dialectical psychology is Hegel refracted through Marxist materialism; Piaget's assimilation/accommodation is the dialectic reframed for cognitive development. A10 therefore answers the skeptic only in the weaker form: "any sufficiently elaborate prompt *within the intersubjective pedagogical family* matches recognition." It does not answer the stronger form: "any sufficiently elaborate pedagogical prompt matches recognition."

A10b (three-judge, $n \approx 45$–60/cell, DeepSeek V3.2 ego, run `eval-2026-04-24-e9a785c0`) addresses the stronger form by adding a fourth cell grounded in a **genuinely orthogonal** pedagogical tradition: behaviorism. Cell_96 loads `tutor-ego-matched-behaviorist.md` (2,957 words, length-matched to recognition, grounded in Skinner operant conditioning, Gagné Nine Events of Instruction, Keller / Bloom mastery learning, Thorndike Laws of Effect/Exercise/Readiness, Rosenshine Principles of Instruction), blocklist-verified for zero recognition/Hegelian content *and* zero Hegelian-descendant constructivist vocabulary (no ZPD, Vygotsky, Piaget, disequilibrium, productive struggle, ICAP, scaffold-as-Vygotskian). Four-way comparison: cell_1 (base), cell_5 (recognition), cell_95 (matched-pedagogical, Hegelian family), cell_96 (matched-behaviorist, transmission family). Three independent judges at matched sample depth.

The three-judge pooled contrasts:

Contrast	Sonnet d	GPT-5.2 d	Opus 4.7 d	**Pooled**
recog vs matched-pedagogical (Hegelian within-family)	−0.05	0.17	0.32	**0.15**
base vs matched-behaviorist (transmission within-family)	0.87	0.53	1.27	**0.89**

Contrast	Sonnet d	GPT-5.2 d	Opus 4.7 d	**Pooled**
Hegelian mean vs transmission mean (between-family)	1.46	1.11	1.58	**1.38**

Three findings, each replicated across three judges:

1. **Within the Hegelian-descendant family**, recognition and matched-pedagogical are statistically indistinguishable (pooled $d = 0.15$, below the $|d| < 0.2$ density-sufficient threshold). Per-judge readings disagree on direction at this scale (Sonnet reads matched-pedagogical marginally higher; Opus reads recognition clearly higher; GPT falls between), which strengthens the structural-features caveat (§8.3 structural-features issue): when contrasts are small, judges lean on idiosyncratic surface-feature preferences, and the direction varies across judges. The pooled estimate is stable at $d \approx 0.15$ and below the threshold regardless of which judge dominates. A10b replicates the A10 within-Hegelian finding on an independent run (A10 pooled $d = 0.17$; A10b pooled $d = 0.15$) — the density-sufficient-within-Hegelian-family claim holds under cross-run replication.

2. **Within the transmission family**, a rigorously-authored behaviorist prompt scores *substantially below* the generic 344-line base prompt (pooled $d = 0.89$; most sharply under Opus, $d = 1.27$). Density with rigorous theoretical grounding is *actively counterproductive* when the grounding is in the wrong pedagogical family. This replicates Paper 1.0's Haiku "naive 35-line beats 344-line base" observation at full rigor and generalises it: the effect is not about length or detail as such, but about the *compatibility* between the theoretical orientation and the tutoring context.

3. **Between families**, the gap is the dominant effect of the A10/A10b cycle (pooled $d = 1.38$) — roughly **nine times larger** than the within-Hegelian-family residual and larger even than the main recognition-vs-base effect in the Paper 2.0 factorial. The Hegelian-descendant intersubjective-pedagogy family outperforms the transmission/behaviorist family by a very large margin, regardless of which specific theorist is cited within each family.

The combined A10/A10b finding is therefore: **the active ingredient in recognition's effect is not density, not rigorous theoretical grounding, not matched specificity in the abstract, and not the Hegelian-specific vocabulary or theoretical framing. It is the broader** intersubjective-pedagogy orientation** that recognition shares with Vygotsky, Piaget, Kapur, Chi, VanLehn, and the constructivist tradition more generally. Density within that orientation family pays off; density outside it backfires. Recognition is one

effective operationalisation of this family; it is not uniquely necessary, but the family as a whole is.**

Three methodological notes:

- The pre-registered structural-features caveat (design note §6a, `docs/pedagogical-taxonomy.md`) applies: the three variants differ not just in theoretical content but in prompt structure (recognition is open-ended and warm, matched-pedagogical is diagnostic and scaffolding-oriented, behaviorist is formal and prescriptive). The three-judge panel mitigates but does not eliminate the possibility that LLM judges reward structural features independently of theoretical content. Human-learner validation (§A1) is the only way to break this confound fully.
- The Anthropic-vs-GPT judge-family split persists in A10b as it did in A10 (Anthropic judges see slightly larger within-Hegelian residuals than GPT; all three converge on direction). Pooled effect sizes include both; per-judge effect sizes are reported above for transparency.
- The A10 v1 run (April 22) was invalidated by a profile-resolution bug (see Supplement V, v3.0.47) that silently routed cell_95 to the base prompt. A10 v2 and A10b used the fixed pipeline; their prompts were verified at run time to load the authored content rather than falling back to base.

Taken together with the four existing arguments against the density reduction (Paper 1.0 placebo, elaborate-hurts-on-Haiku, autotuning widens-under-optimization, question-asking mediation) and the A11 architecture-isolation finding (§6.4.2, the ego–superego architecture performs real error-correction work the prompt alone cannot), the density alternative is now resolved at the orientation-family level: recognition's effect is reducible neither to length nor to rigorous-grounding-in-general, but it *is* reducible to membership in the intersubjective-pedagogy family. Full results: `exports/a10-prompt-density-control.md` (A10 three-judge), `exports/a10b-orientation-family.md` (A10b four-way three-judge); analysis scripts: `scripts/analyze-a10-prompt-density-control.js`, `scripts/analyze-a10b-orientation-family.js`; pedagogical-orientation taxonomy: `docs/pedagogical-taxonomy.md`.

7.10 Mechanism location: response-level form versus substance The orientation-family finding (§7.9) raises a follow-up: where within the family-level effect does the score-driving mechanism actually live? A five-pass mechanism analysis on the same A10b corpus pins this to a specific pragmatic act, and in the process surfaces a methodological caution worth flagging.

Independent corroboration via embedding similarity. Cosine similarity between each tutor response and a hand-authored intersubjective canonical ($ 120 words exemplifying definition-and-application and prescriptive instruction) separates the families at $d = 0.81$ on the difference (intersubjective minus

transmission). This reproduces the family-level pattern A10b found in score data through a measurement channel independent of the LLM judges, providing convergent evidence that the family separation is not solely a rubric artefact. Same A10b corpus (eval-2026-04-24-e9a785c0), $n = 191$, OpenAI text-embedding-3-small.

Simpson's paradox at the response level. Pooled across all four cells, embedding-similarity to the intersubjective canonical is *positively* correlated with score ($r = +0.26$). Within each individual intersubjective cell, however, the correlation reverses sign ($r = -0.28$ in cell_5, $r = -0.24$ in cell_95). The pooled positive correlation is driven entirely by cell_96 (matched-behaviorist) anchoring the lower-left of the pooled scatter. Within-cell, more form-similarity to the intersubjective template predicts *lower* scores. The same direction-reversal appears for surface-level acknowledgement markers (quoted text, "your X" possessives, "what you mentioned"-style phrases): pooled r near zero, but within-cell $r = -0.18$ in cell_5, $r = -0.31$ in cell_95, $r = -0.38$ in cell_1.

Substantively: the intersubjective stance manifests through observable surface features, but those features are *markers* of the underlying stance, not *substitutes* for it. Responses that hit every surface cue without underlying engagement read as formulaic; surface mimicry tracks lower-quality responses, not higher. The rubric is not fooled by surface form; it rewards substance. A naive operationalisation that mandated "more acknowledgement of the learner" without inducing the underlying stance would predict a *negative* score effect at the response level, not a positive one — which inoculates against an obvious operationalist misreading of the family finding ("just tell tutors to acknowledge more").

The single durable within-cell mediator: ending the tutor turn with a question. Across the five-pass sequence (Hegelian lexicon density; intersubjective lexicon density; two regex-based pragmatic feature passes covering question-mark rate, second-person density, scaffolding-move imperatives, inclusive framing, modal invitations, hedging; embedding similarity), only one feature satisfies all three within-cell mediator criteria — large family contrast, small within-intersubjective contrast, AND positive within-cell correlation. Cells 1 and 96 end the tutor turn with a question in **0%** of responses; cell_5 does so in 4.5% of responses; cell_95 in 2.1%. Within both intersubjective cells, responses that happen to end with a question score higher (within-cell Pearson $r = +0.325$ in cell_5, $r = +0.392$ in cell_95). Pragmatically this is the act of *ceding initiative back to the learner*; the within-cell correlation captures situations where doing so is contextually appropriate, not just stylistic preference.

The 2–5% base rate constrains how much of the family-level effect (cell_5 mean 47.9 vs cell_1 mean 33.8, $\Delta = +14$ points) ending-with-question can arithmetically explain on its own. Even if every such response scored 30 points higher than its peers, that could not account for the +14-point family effect. Ends-with-question is *one observable channel* of the underlying intersubjective stance, not a substitute for the stance itself.

Three things "the recognition injection" could refer to. The orientation-family finding (§7.9) plus this response-level mechanism analysis together support the following decomposition, which clarifies what is and is not load-bearing in the present apparatus:

Candidate active ingredient	Doing the work?	Evidence
The Hegelian theoretical content (mutual subjects, master–slave dialectic, Hegelian vocabulary in the prompt body)	**No**	A10b score equivalence cell_5 ≈ cell_95 (§7.9); D1 lexical analysis showing cell_95 actively *suppresses* Hegelian vocabulary and scores indistinguishably
The intersubjective pragmatic stance the prompt induces (irrespective of vocabulary)	**Yes — substantially**	Family $d \approx 1.38$ between intersubjective and transmission (§7.9); embedding-canonical $d = 0.81$; multiple structural correlates (ends-with-question, scaffolding-move imperatives, inclusive framing)
Specific surface features individually (ends-with-question, scaffolding moves, "your X" acknowledgement)	**Partially — observable channels of the stance, not substitutes**	Within-cell ends-with-question $r = +0.325/+0.392$; categorical 0% vs 2–5% pattern; but base rates too low to arithmetically account for the family effect; surface acknowledgement is *negatively* correlated with score within cells

This decomposition is consistent with the v3.0.50 "primacy + descent" framing. Recognition theory operationalised as a prompt is one *effective* operationalisation of the intersubjective-pedagogy family's commitments; the Hegelian content in that operationalisation is not the channel through which it acts; what is load-bearing is the family-level stance the prompt induces, which manifests partly through observable structural features such as ending with a question. A naive operationalisation that mandated only ends-with-question without inducing the underlying stance would predict a much smaller score gain than recognition's, because surface mimicry of one channel does not substitute for the multi-channel manifestation of the stance whose substance the rubric rewards.

Full mechanism analysis: `notes/d1-paper-integration.md`. Analysis scripts: `scripts/analyze-recognition-lexicon.js` (pass 1, Hegelian lexicon), `scripts/analyze-d1-orientation-lexicon.js` (pass 2, intersubjective lexicon), `scripts/analyze-d1-structural-features.js` (pass 3, basic pragmatics), `scripts/analyze-d1-structural-features-v2.js` (pass 4, refined pragmatics), `scripts/analyze-d1-structural-features-v3.js` (pass 5, embedding similarity). Reports: `exports/d1-orientation-lexicon.md`, `exports/d1-structural-features.md`, `exports/d1-structural-features-v2.md`, `exports/d1-structural-features-v3.md`.

7.10.1 Replication, multivariate control, and a single-turn-vs-multi-turn scope condition The §7.10 headline (ends-with-question as the within-cell mediator) was checked against three robustness extensions:

Cross-judge replication. The §7.10 within-cell correlations were re-computed using the Opus 4.7 and GPT-5.2 judges of the A10b three-judge panel (in addition to the original Sonnet 4.6 results). The ends-with-question r remains positive in 5 of 6 cell-judge combinations across the two intersubjective cells (Sonnet cell_5 $r = +0.325$, cell_95 $r = +0.392$; Opus cell_5 $r = +0.189$, cell_95 $r = +0.321$; GPT cell_5 $r = +0.116$, cell_95 $r = -0.011$). The one exception (GPT $\times$ cell_95) is essentially zero rather than a sign-flip. The Simpson's-paradox pattern on intersub_advantage (pooled r positive, within-cell r negative in both intersubjective cells) replicates in all 3 judges. Both findings are not Sonnet-specific. Report: `exports/d1-cross-judge-replication.md`.

Multivariate OLS control. The ends-with-question within-cell partial coefficient was estimated controlling for second-person density, scaffolding moves, broad acknowledgement, question-mark rate, and intersub_advantage (six predictors plus intercept; $n \approx 50$ per cell, df ≈ 43). In cell_5: $\beta_{\text{ends-Q}} = 26.34$ (SE 10.85, $p = 0.02$). In cell_95: $\beta_{\text{ends-Q}} = 38.52$ (SE 16.49, $p = 0.02$). Both partial coefficients remain positive and significant after multivariate control — the single-feature mediator interpretation is not collinearity-driven. The intersub_advantage Simpson's-paradox direction also survives multivariate control: pooled $\beta = +34.34$ ($p = 0.02$) but within-cell $\beta_{\text{cell_5}} = -49.54$ ($p = 0.14$), $\beta_{\text{cell_95}} = -82.15$ ($p = 0.04$). Report: `exports/d1-multifeature-ols.md`.

Cross-run / cross-application / single-vs-multi-turn replication. The mediator account was tested on five additional runs with intersubjective recognition cells: A10 v2 (independent DeepSeek philosophy run), A6 programming (Haiku ego, programming domain), D2 Path 1 (Haiku ego, peer support coaching application), and Paper 2.0 multi-turn cells 80–87 (DeepSeek messages-mode factorial, single-agent and multi-agent variants). In single-turn cells the mediator replicates uniformly: A10 v2 cell_5 $r = +0.346$, A6 programming cell_5 $r = +0.202$, **D2 peer support cell_5** $r = +0.739$ (the strongest within-cell r in the entire D1 sequence, on a different application entirely). In multi-turn cells the relationship **reverses**: cell_84 (recog single-agent, messages-mode) $r = -0.301$, cell_86 (recog multi-agent, messages-mode) $r = -0.203$, A2 cell_

61 (dialectical multi-agent, dynamic learner) $r = -0.087$, A2 cell_63 (dialectical multi-agent, bidirectional profiling) $r = -0.199$.

The single-turn–vs–multi-turn split has a clean pragmatic interpretation. In single-turn responses, ending with a question is the only available channel to cede initiative back to the learner; doing so signals "your turn now," which the rubric rewards as engaged tutoring. In multi-turn dialogues, the rubric scores the *full dialogue arc*; ending the *final* turn with a question means leaving the conversation unresolved, which the judges appear to penalise. The same surface feature signals different pragmatic acts depending on whether it occurs in a single-shot exchange or as the closing move of a multi-turn dialogue. (The §6.3 trajectory analyses already establish that turn-by-turn dynamics differ qualitatively from single-turn quality; the mediator's sign-reversal across this boundary is consistent with that broader pattern.)

The §7.10 mediator claim therefore has an explicit scope condition: **ends-with-question is a within-cell mediator in single-turn intersubjective cells (5 of 5 tested, r range $+0.16$ to $+0.74$, including replications across DeepSeek and Haiku ego, philosophy/programming/peer-support domains, and Sonnet/Opus/GPT judges); the relationship reverses sign in multi-turn intersubjective cells (4 of 5 tested, r range -0.09 to -0.30, with a single anomaly at cell_82 multi-turn base+superego, $r = +0.38$ at $n = 18$).** The cross-mode reversal does not falsify the underlying mechanism account (initiative-ceding); it specifies the discourse context in which the surface feature signals it. Report: `exports/d1-ends-question-replication.md`.

8. Limitations

8.1 Synthetic Learners All evaluations use LLM-generated learner turns rather than real learners. The two supported mechanisms (calibration, error correction) operate on the tutor's production process, which is observable regardless of whether the learner is synthetic or human. However, the *consequences* of these mechanisms for learning outcomes cannot be assessed without human learners. The synthetic learner may respond to better tutoring in ways that diverge from human learners: genuine confusion is different from simulated confusion, and genuine resistance differs from scripted resistance. The mechanism-level findings are robust to learner type (they trace tutor-internal processes), but their pedagogical significance depends on whether the improved tutor behavior actually produces better learning.

The learner superego paradox (d=3.05, Paper 1.0, Section 6.16) complicates this further: the multi-agent learner architecture lowers synthetic learner rubric scores by d=3.05, but whether this reflects architecture-level degradation or a rubric-architecture mismatch is empirically separable. The next paragraph reports a DB-only mechanism check that favours the rubric-artifact account — i.e., the architecture produces substantially more deliberation-internal content that the current rubric does not reward — though the stronger test against

human-judged learner output has not yet been run.

A rubric-measurement artifact is the most parsimonious account. A DB-only mechanism check on cells 1/2/5/6 (base × recognition × unified/ego_superego learner, Opus 4.6 judge, N=613/108/499/157) separates learner *behaviour* from learner *score*. The ego-superego learner architecture *produces more behaviour*: mean output tokens 13,193–15,547 vs 5,125–6,675 (~2.5× more text generated), and mean API calls 9.0–10.4 vs 2.3 (~4× more deliberation rounds). Yet learner overall scores are systematically *lower* (base: 53.94 vs 72.06, $\Delta = -18.1$; recognition: 59.19 vs 71.38, $\Delta = -12.2$). The direction is consistent: the architecture that performs *more* internal self-critique scores *worse* on the external rubric, despite consuming more tokens and more deliberation rounds. Per-dimension inspection of the learner_scores JSON reinforces this: ego-superego learners score well on `deliberation_depth` (which rewards the internal process directly) and `revision_signals` (which rewards observable revision), but poorly on `conceptual_engagement` and `question_quality` (which reward substantive *content* in the output). The rubric as currently constructed rewards content engagement in the final external message; the ego-superego architecture spends generative effort on internal deliberation that produces polish and metacognitive commitments ("I'll actually read the full passage") rather than conceptually rich outputs. The "d=3.05 quality deficit" therefore reflects *what the rubric measures and what the architecture produces*, not that internal self-critique degrades the learner as such. A fairer test would either extend the rubric to reward deliberation-internal content or compare architectures on human-judged engagement rather than rubric-scored output.

A convergent-validity check across the broader corpus supports the artifact account. The database contains 1,759 rows carrying *both* a per-turn learner rubric score and a holistic learner score — the latter a global dialogue-level quality judgment that is not decomposed into the rubric's conceptual-engagement / question-quality / deliberation-depth dimensions. On these paired rows, the rubric and holistic judges agree within each architecture at Pearson $r = 0.72$–0.77 (`unified` $r = .73$, $n = 432$; `ego_superego` $r = .75$, $n = 411$; `unified_recognition` $r = .77$, $n = 446$; `ego_superego_recognition` $r = .72$, $n = 470$). The rubric is therefore a reasonable proxy for holistic quality judgment within each architecture. The paradox contrast, however, does *not* replicate on this wider sample: pooled across all cells, ego_superego learners score *higher* than unified learners on both metrics (base contrast: rubric $\Delta = +4.14$, $d = 0.25$; holistic $\Delta = +7.39$, $d = 0.39$; recognition contrast: near-zero on both, $|d| < 0.08$). The cell-level breakdown of the large-N messages-mode cells (80–87, $n = 179$–216 per cell) consistently shows ego_superego $\geq$ unified on both metrics. The $d = 3.05$ paradox was computed on the eight-cell 2$×2×$2 factorial (cells 1–8, Paper 1.0 §6.16, single-prompt mode, $n = 18$ per cell), and the pattern does not survive at the architecture level when messages-mode cells dominate the sample. This is consistent with a measurement-artifact account: the original paradox is a scenario-and-mode-specific interaction between the rubric's content-centric dimensions and the ego_superego architecture's deliberation-

internal content generation, not a general property of "internal self-critique degrades the learner."

Cell-matched holistic scores on the original paradox run: rubric d reproduces; holistic d shrinks by roughly half. The definitive adjudication adds holistic learner scores to the originally-scored 48 rows of the paradox run (eval-2026-02-20-25c78e91, single-prompt mode, Opus 4.6 judge on rubric v1.0, 6 rows × 8 cells; 47 paired after one judge JSON-parse failure). To avoid cross-version contamination, a standalone script wrote only the four holistic fields via direct SQL `UPDATE`, preserving the `learner_rubric_version = 1.0` marker rather than triggering the store's default version-overwrite (Sonnet 4.5 judge for holistic, matching the v3 triple-panel convention). The rubric d on this paired subset reproduces the paradox essentially exactly: base contrast (ego_superego − unified) **rubric** $d = -2.82$ ($\Delta = -23.3$, $n = 11/12$); recognition contrast **rubric** $d = -3.11$ ($\Delta = -22.3$, $n = 12/12$). On the *same rows*, the holistic judge narrows both gaps by roughly half: base **holistic** $d = -1.34$ ($\Delta = -21.4$); recognition **holistic** $d = -1.06$ ($\Delta = -11.7$). Critically, within-architecture rubric holistic correlations on the paradox cells are *weak and heterogeneous* ($r = 0.06$–0.44 across the four architectures) compared to the corpus-wide $r = 0.72$–0.77 pattern — the two judges specifically diverge on the paradox cells rather than agreeing as they do elsewhere. The evidence is mixed but tilts toward the artifact account: the ~50% shrinkage in $|d|$ plus the selective breakdown of convergent validity on these cells indicates the rubric contributes roughly half of the original $d = 3.05$ effect. The residual holistic gap ($|d| \approx 1.0$–1.3) is plausibly a single-prompt-mode interaction rather than a general capability ceiling, since the same architecture scores equal-or-higher on both metrics in messages-mode cells 80–87 (preceding paragraph). The paradox is therefore best characterized as a **combined measurement-plus-mode interaction specific to single-prompt evaluations**, not a universal property of internal self-critique. The ~$2K higher-capability-model experiment previously priced as the adjudicating test is no longer required for the central capability-vs-artifact question; subsequent work should target the single-prompt-mode interaction directly rather than re-testing capability at frontier scale.

8.2 LLM-as-Judge Evaluation Using LLM judges to evaluate recognition quality may introduce systematic biases. The judge may reward surface markers of recognition (acknowledging learner contributions, using inclusive language) rather than genuine engagement. **Cross-judge validation is now complete**: three independent judges (Claude Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4) scored all 1,296 rows across three generation models (3 runs × 3 judges × 144 rows, zero nulls), with blind scoring confirmed by code audit.

The cross-judge results establish both convergent validity and judge-specific patterns. Recognition effect directions replicate unanimously across all 9 judge × run cells (d = 1.34–1.92). Inter-judge Pearson correlations range from r = 0.45 to r = 0.89, with reliability varying by dimension (tutor overall: r = .61–

.89; dialogue quality: r = .54–.87; holistic: r = .45–.79) and by generation model (highest on Gemini Flash 3.0 at r = .63–.89, lowest on DeepSeek at r = .45–.80). The higher reliability on intermediate-quality generation suggests that judges agree most when quality differences are clear and unambiguous. A consistent leniency gradient emerges: Gemini 3.1 Pro is most lenient, GPT-5.4 most severe, with Sonnet intermediate. This gradient is judge-specific rather than run-specific, confirming it reflects judge calibration rather than data artifacts.

For mechanism-level claims, the judge limitation is particularly relevant to the superego critique taxonomy (Section 5.3): critique categories are classified by an LLM, and the ego's "substantive revision" versus "cosmetic compliance" is assessed by an LLM judge. The mechanism claims are thus LLM-assessed claims about LLM-internal processes—a recursive structure that introduces the possibility of systematic blind spots. The cross-judge validation mitigates this for factorial-level claims (all three judges agree on effect directions and interaction patterns), but does not address it for process-level claims where only a single classifier is used. Human expert coding of a representative subsample is the most important outstanding validation for the mechanism claims. The pilot protocol is now fully operationalised in the repository (§5.3, "Human Validation Pilot"): deterministic stratified sampling (`scripts/human-validation-sample.js`, seed=20260416, n=40 across 10 categories), a per-category codebook with signal phrases, boundary rules, and exemplars (`docs/research/human-coding-codebook.md`), and an analysis pipeline computing Cohen's κ, Fleiss' κ, per-category F1, and confusion matrices (`scripts/human-validation-analyze.js`), plus an inter-LLM baseline (Haiku vs Sonnet on the same sample) as a lower-bound reference against which human-LLM agreement can be calibrated. The target floor is ≥ 0.60 (substantial agreement per Landis-Koch bands). Until the filled rater CSVs are returned and analysed, the process-level findings should be treated as LLM-generated interpretive categories whose correspondence to human expert judgment has not yet been empirically established. The apparatus reduces the execution cost to "recruit two coders, send them the packet, run the analysis script" — human time rather than rebuilding infrastructure.

8.3 Model Transience Findings are model-version-specific. Paper 2.0 uses three generation models—DeepSeek V3.2 (open-weight, 685B MoE), Haiku 4.5 (proprietary, optimized for speed), and Gemini Flash 3.0 (proprietary, multimodal)—with three independent judges (Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4). The recognition effect replicates across all 9 judge × run cells (d = 1.34–1.92), substantially strengthening the generalizability claim relative to the two-model analysis. The addition of Gemini Flash 3.0 (weaker than the other two models) partially addresses the capability-range limitation: it shows that mechanisms replicate in direction on weaker models while revealing generation-quality boundary conditions on their interactions (universal substitution with model-dependent residual, learner invariance → sensitivity). The Nemotron data in the database (N=40, mean tutor score ~8 under base) hints

that very weak models may not benefit from recognition at all, but the sample is too small for confident claims. The two supported mechanisms (calibration through prompt orientation, error correction with a model-dependent residual) describe *prompt-level and architecture-level* properties that should generalize across model versions, supported by three-model × three-judge replication. The null finding for adaptive responsiveness also replicates across all three models and judges.

Per-judge effect-size decomposition. While direction replication across all 9 cells establishes convergent validity, the *magnitude* of the recognition effect varies systematically by judge. Decomposing the d = 1.34–1.92 range from Section 6.1 by (run × judge) cell on tutor_overall_score:

Table 69: Cohen's d for the recognition main effect by generation model and judge, computed from within-group SDs pooled across the two conditions (n=72 per condition per cell). Within-row Δd is the max–min range of effect sizes across the three judges for a given generation model.

Generation Model	Sonnet 4.6	Gemini 3.1 Pro	GPT-5.4	Within-row Δd
DeepSeek V3.2	1.85	1.57	1.34	0.51
Gemini Flash 3.0	1.87	1.40	1.77	0.47
Haiku 4.5	1.92	1.34	1.57	0.58
Judge mean	**1.88**	**1.44**	**1.56**	–

Three features deserve note. First, the across-judge Δd exceeds the 0.1 d robustness threshold in all three rows (range 0.47–0.58), so absolute effect-size estimates are judge-dependent even though effect direction is conserved. Second, Sonnet 4.6 systematically reports the largest effects (judge-mean d = 1.88, ~20% higher than Gemini 3.1 Pro's 1.44 and ~20% higher than GPT-5.4's 1.56). Because Sonnet was the original judge for the pilot study and for much historical analysis, **claims conditioned on Sonnet-only scoring should be read as an upper-bound estimate of pooled effect magnitude**. Third, the pooled-across-judges mean is d ≈ 1.63 (mean of 9 cells); we treat this as the judge-adjusted best estimate, with Sonnet-only d ≈ 1.88 as an upper bound and Gemini-3.1-Pro-only d ≈ 1.44 as a lower bound. All three judge panels leave the recognition effect in Cohen's "very large" range (d > 1.3), but reporting only one judge without adjustment risks a misleading impression of magnitude. The pattern is partially but not fully consistent with the leniency gradient noted in §8.2: Sonnet produces the highest d across all three generation models, but the Gemini-vs-GPT ordering flips between runs (GPT > Gemini on Gemini Flash 3.0 and Haiku; Gemini > GPT on DeepSeek), suggesting judge × generation-model interaction beyond simple calibration differences.

Two implications for replication practice follow. Pre-registered judge panels reduce the risk that downstream claims inherit one judge's idiosyncratic severity; and per-judge effect sizes should be reported alongside pooled estimates, rather

than pooling silently under a single judge and calling it the effect. The ordering across *generation models within a judge* is more stable than the ordering across *judges within a model*, which supports using the architecture × recognition interaction structure (Sections 6.4–6.6) as the primary inferential target rather than absolute effect magnitudes on any one judge.

8.4 Process Tracing with LLMs Our adaptation of process tracing from social science to AI systems introduces a philosophical complication. In political science, process tracing examines *actual* causal chains: the decision-maker's deliberation, the institutional constraints, the information flows. In our architecture, the ego-superego exchange is *generated* text—the "deliberation" is a prompted LLM output, not a cognitive process in any neuroscientific sense. The mechanism claims are about designed information flows between prompted agents, not about internal mental states.

This limitation is softened by the distinction drawn in Section 3: we study recognition-*oriented design* (functional analogues), not recognition *proper* (genuine intersubjectivity). The process tracing examines whether designed information flows produce measurable output differences—a behavioral claim, not a cognitive one. The null finding for adaptive responsiveness as a third mechanism illustrates the value of this approach: the method was capable of distinguishing supported mechanisms (calibration and error correction, where condition-dependent effects are observed) from unsupported ones (adaptive responsiveness, where the null is well-powered). But the language of "mechanisms" may imply more causal depth than the evidence supports.

A related concern is whether the paper conflates *engineered architecture* with *discovered mechanism*. The ego-superego pathway, the critique taxonomy, and the revision loop are designed into the system. In one sense, finding that the superego catches errors the ego misses is reporting that the architecture works as intended, not uncovering a naturally occurring mechanism of good tutoring. We acknowledge this distinction: the mechanisms we trace are properties of *this specific architecture*, not universal laws of AI tutoring. Calibration describes how recognition prompts constrain *this system's* output distribution; error correction describes how *this system's* superego-ego interaction functions. Whether other architectures (chain-of-thought, debate-based, or single-agent with self-critique) would show analogous patterns under recognition-oriented prompting is an open question (Section 8.5). However, the findings are not purely tautological. The architecture was designed to enable observability, not to guarantee specific outcomes. Three results were genuinely unpredicted: that calibration alone accounts for most of the recognition effect (making the superego largely redundant on strong models), that adaptive responsiveness fails as a general mechanism despite being the most theoretically motivated prediction, and that the substitution interaction produces a consistent 15–17% additivity deficit across three structurally different models. The system was built to make mechanisms *observable*; what those mechanisms turned out to be was an empirical question.

8.5 Single-System Study All findings come from a single architectural implementation (the ego/superego tutoring system with recognition-enhanced prompts). The two supported mechanisms are identified through this specific architecture's observability. Other architectures—chain-of-thought reasoning, debate-based systems, recursive self-improvement—might implement recognition-like behaviors through different internal processes. The mechanism model may not transfer to architectures that lack explicit ego-superego separation or that do not log internal deliberation.

8.6 Rubric Evolution The rubric evolved through four versions during the study. Paper 1.0 data uses v1.0 (14 tutor dimensions); Paper 2.0 data uses v2.2 (8 tutor dimensions). Cross-version comparisons are inadvisable—the dimension structure, weighting, and judge instructions differ—so we do not retroactively score historical data under newer rubric versions. This means pilot findings (v1.0) and mechanism findings (v2.2) use different measurement instruments, and direct numerical comparisons between them should be treated with caution.

The v2.2 rubric has been validated through synthetic calibration (r=+0.996 tutor, r=+0.968 learner) and through initial empirical use on cells 80–87 (N=309 under v2.2). The 8-dimension structure produces coherent findings: floor-lifting patterns, dimension-rank consistency across models, and calibration effects that align with theoretical predictions (Section 6.1).

However, principal components analysis on 1,584 per-turn observations reveals that the 8 dimensions largely measure a single construct: PC1 explains 80.7% of variance, and the Kaiser criterion yields only one factor with eigenvalue $>$1. Sampling adequacy is excellent (KMO = 0.938), and the mean inter-dimension correlation is r=0.776 (range 0.589–0.921). This pattern replicates across conditions (PC1: 80.2% base, 75.6% recognition) and models (77.3% DeepSeek, 68.0% Haiku). Forced two-factor varimax rotation separates `content_accuracy` (loading 0.923 on Factor 2) from the seven pedagogical dimensions (loadings 0.68–0.85 on Factor 1), suggesting that the rubric captures two facets—factual correctness and pedagogical quality—but within the pedagogical facet, the dimensions are not empirically distinct.

This has two implications for the mechanism findings. First, the calibration effect (Section 6.1)—within-response dimension variance reduction under recognition—is meaningful precisely *because* the dimensions co-vary: a prompt that lifts one dimension tends to lift all, and calibration describes the narrowing of an already correlated profile. Second, the rubric is functionally over-specified: 8 nominally distinct dimensions reduce to 1–2 empirical factors, and future iterations should either consolidate or demonstrate discriminant validity. The current findings should be interpreted as measuring overall tutoring quality with a content-accuracy modifier, not 8 independent constructs.

Implication for §7.8 dimension-targeted optimisation. The PCA finding also bears on §7.8.1's dimension-targeted autotuning results. Targeting

`adaptive_responsiveness` or `conceptual_progression` as optimisation objectives does not isolate empirically independent skills; those dimensions load on the same underlying tutor-quality factor as the others. The §7.8.1 results should therefore be read as *shifts in the single underlying construct via prompt changes whose wording foregrounds a particular facet*, not as selective movement of one skill while others are held constant. This does not invalidate the §7.8 findings — prompt optimisation did improve scores substantially (Qwen target dims 35 → 75), cross-model transfer did produce the prosthesis-straitjacket inversion, and these results are informative about how to engineer prompt interventions for different model capabilities. But the mechanistic interpretation is one of *shaping the overall quality factor* rather than *building a specific pedagogical skill*, and readers should not treat per-dimension improvements as evidence of independent constructs being independently tunable.

Methodological pairing with §7.10 (Simpson's-paradox at the response level). A second methods-level caution arose from the §7.10 D1 mechanism analysis: mediator-criteria checks must use within-group correlations when the predictor varies across groups. Embedding-similarity to an intersubjective canonical was *positively* correlated with score when pooled across the four A10b cells ($r = +0.26$) but reversed sign within each intersubjective cell ($r = -0.28$ in cell_5, $r = -0.24$ in cell_95) — a textbook Simpson's paradox driven by the cell_96 transmission outlier anchoring the lower-left of the pooled scatter. Together with the PCA over-specification finding above, these pair as two methods-level cautions for any future mediator analysis on LLM-tutor data: (1) the v2.2 rubric's eight dimensions reduce to one underlying construct so per-dimension claims should be hedged accordingly; and (2) pooled-r in mediator analysis can be entirely driven by between-group variance rather than within-group mechanism. Together they could form the empirical core of a methods-paper contribution beyond Paper 2.0; for the present paper they are documented here as caveats that any reader applying the apparatus elsewhere should account for. §8.6.1 below makes the second caution explicit as a portable diagnostic.

8.6.1 The three-criterion within-cell mediator diagnostic The §7.10 mechanism analysis used a specific diagnostic to distinguish *family markers* (features that distinguish prompt families but do not predict scores within a family) from *within-cell mediators* (features that predict scores even when the prompt is held constant). We articulate the diagnostic explicitly here because the same logic applies to any LLM-tutor evaluation that contrasts prompt families and seeks response-level mechanisms.

A response-level feature qualifies as a within-cell mediator candidate when all three criteria are met:

1. **Large family contrast** ($|d| \geq 0.5$ between the family containing the recognition cell and a contrasting transmission family). The feature distinguishes the families. If it does not, it cannot mediate the family-level

effect.

2. **Small within-family contrast** ($|d| < 0.5$ between two cells within the same family, e.g., recognition vs matched-pedagogical). Both family-member prompts elicit the feature; if only one does, the feature is a *prompt marker* (specific to one prompt) rather than a *family marker* (shared across the family).
3. **Positive within-cell Pearson r with score** in each family-member cell. When the prompt is held constant, the feature still predicts higher scores. This isolates the response-level mediating role from cross-cell variance that the family contrast already captures.

The §7.10 worked example applied this diagnostic to the D1 five-pass feature sequence on the A10b 4-cell corpus:

Feature	Criterion 1 (family d)	Criterion 2 (within-family d)	Criterion 3 (within-cell r in cell_5 / cell_95)	Verdict
Hegelian lexicon density	small	very large (cell_5 ≫ cell_95)	small	Prompt marker only
Intersubjective lexicon density	moderate	very large (cell_95 ≫ cell_5)	mixed signs across cells	Prompt marker only
Question-mark rate	0.61	0.47 (cell_5 3× cell_95)	small	Prompt marker, modest predictor
Scaffolding-move imperatives	0.59	0.12	+0.19 / +0.05	Family marker, weak mediator
Broad acknowledgement	0.14	small	−0.18 / −0.31	Negative correlate (anti-mediator)
Ends-with-question (single-turn cells)	**categorical** (4.5% / 2.1% vs 0%)	**0.13**	+0.325 / +0.392	**Mediator**
Intersub. canonical similarity (embedding)	**0.81**	0.27	−0.28 / −0.24 (negative within-cell, positive pooled)	Family marker; **Simpson's paradox disqualifies as mediator**

A feature passes the mediator bar only by satisfying all three criteria; failure on any one disqualifies the feature as a within-cell mediator while potentially preserving its role as a family marker, prompt marker, or negative correlate.

The diagnostic is most informative when criterion 3 fails despite criteria 1–2 passing: that pattern indicates Simpson's paradox, where a feature appears mediating across families (criterion 1 large, pooled $r > 0$) but the within-cell direction reveals that the cross-family score gap, not the within-family feature variation, was carrying the pooled correlation.

The diagnostic generalises beyond the recognition–transmission contrast. Any factorial design with a family-level main effect and candidate response-level mediators can apply the three-criterion check, and the §8.6 PCA finding adds a complementary rubric-construct caution: per-dimension "mediation" claims should be hedged when the rubric's dimensions are not discriminantly valid. Together the rubric-construct caution and the within-cell-mediator diagnostic constitute two independent methods-level corrections that any future LLM-tutor evaluation should account for. Their empirical foundations are documented here as Paper 2.0 caveats; a separate methods-paper extraction is the natural vehicle for arguing their generality across published LLM-tutor evaluation pipelines.

8.6.2 Application protocol for transferring the diagnostic The diagnostic transfers to any LLM-tutor evaluation pipeline that exposes (a) per-row condition labels grouping responses into prompt families, (b) per-row scores from a judge, and (c) at least one per-row response-level feature derivable from the response text or trace. The transfer protocol is procedurally explicit, requires no new generation or judging, and operates on data already collected in the source evaluation:

1. **Identify the family partition.** Group conditions into family A (the family the source evaluation claims is doing pedagogically distinctive work) and family B (a contrasting transmission-style family). Within each family, identify two cells whose prompts share the family's pragmatic stance but differ in surface composition (e.g., recognition vs matched-pedagogical for A; generic base vs matched-behaviorist for B). If only one family-member cell is available, criterion 2 cannot be evaluated and the protocol degrades to a two-criterion check; this is acceptable but weaker.
2. **Compute the candidate feature** f per row from the available text or trace data. Regex extractors (per §6.1.7 question-asking and §7.10 ends-with-question) and embedding-based features (per §7.10) are both supported. Author-specified feature definitions should be reported alongside results.
3. **Test criterion 1** (family contrast): compute Cohen's d for f between pooled family A and pooled family B. Require $|d| \geq 0.5$.
4. **Test criterion 2** (within-family contrast): compute Cohen's d for f between the two cells within family A. Require $|d| < 0.5$. If only one A-cell is available, skip; flag in the report.
5. **Test criterion 3** (within-cell correlation): compute Pearson r between f and score within each family-A cell separately. Require both $r > 0$. Negative within-cell r in one or both cells disqualifies f as a mediator

and may flag f as a *negative correlate* (cf. broad-acknowledgement, §7.10). Pooled $r > 0$ with within-cell $r < 0$ in both cells is the Simpson's-paradox signature; report explicitly.

6. **Pair with rubric over-specification check** (§8.6): compute PCA on the per-dimension scores from the source rubric. Report PC1 variance share, KMO, mean inter-dimension r, and Kaiser eigenvalue count. PC1 $\geq$ 70% flags rubric over-specification; per-dimension claims should be hedged accordingly.

Reference scripts for steps 2–5 are open-sourced as `scripts/analyze-d1-structural-features.js`, `scripts/analyze-d1-cross-judge-replication.js`, and `scripts/analyze-d1-multifeature-ols.js`. The same scripts apply to any input dataset matching the schema described in step 1; data-loading is the only extension point. Future application of the protocol to independently published rubrics — for example, MathTutorBench (Macina et al. 2025), ICAP-aligned LLM-as-judge studies (Chi and Wylie 2014), or the structured rubric methodology of LLM-Rubric (Hashemi et al. 2024) — would test the diagnostic's generality across rubric design choices and is the natural next empirical step beyond Paper 2.0's source apparatus.

8.7 Computational Cost Asymmetry The recognition-enhanced prompt design has a practical cost: the tutor ego prompt (~5,100 tokens) is 8× larger than the learner ego prompt, and multi-agent deliberation multiplies this per turn. A 10-turn multi-agent dialogue consumes ~225,000 tokens (64 API calls), with the tutor accounting for 3.4× more input tokens than the learner (Section 4.2). Recognition-enhanced tutoring is substantially more expensive per interaction than baseline tutoring, and this cost scales with turn count and rejection rate. Whether the quality gains justify the cost depends on the deployment context; the substitution finding (Section 6.4) suggests that single-agent recognition cells, which eliminate the superego's redundant correction rounds, may offer a better cost-quality trade-off on capable models.

Reproduction cost for Paper 2.0 mechanism runs. A lab seeking to replicate the three core Paper 2.0 generation runs (DeepSeek V3.2, Gemini Flash 3.0, Haiku 4.5; 432 dialogues across 18 cells × 3 runs) faces a recorded generation cost of $114.58 across 163M tokens (144M input, 19M output) on OpenRouter's pass-through pricing — kept low by the choice of three cost-efficient generation models. Judge costs are not separately tracked in the database but can be estimated: a 3-judge panel (Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4) scoring all 1,296 multi-turn rows (~8K input + ~1.5K output per judge call) costs an estimated $300–700 depending on judge price mix. Superego-taxonomy classification adds a further ~$80–100. Full reproduction of the core Paper 2.0 runs thus falls in the $500–1,500 range at late-2026 API rates; including the isolation runs (Section 6.4.2), trajectory-specific scenarios (Section 6.3), and prompt-lab autotuning (Section 7.8) roughly doubles this to $1,000–3,000. The total evaluation database used throughout this programme (207 runs, 8,274 dialogues, ~740M tokens) consumed $474.75 in recorded generation cost, with estimated

total cost including judging at $2,000–5,000. These figures assume LLM judging; substituting human expert coding of 30–50 exchanges per mechanism (the outstanding validation discussed in Section 8.2) would add rater labour costs, and full-scale human-learner validation (Section 8.1) at $50–100 per participant for a target N=200 would push replication costs into the $20K–50K range. The LLM-judged apparatus is therefore inexpensive to reproduce — the genuine replication barrier is scaling up to human validation, not the compute.

8.8 Pilot Data Transparency The companion pilot study documented nine post-extraction corrections, four of which affected quantitative results. Bug 4 (multi-turn scoring misalignment) affected 8,631 rows. Bug 5 (resume model override loss) affected one run. Bugs 1–2 (learner prompt leakage, broken conversation history) required complete re-generation of all dynamic learner data. These corrections are documented transparently in the companion study and in the project's bug registry, and the provable discourse framework tracks which claims depend on corrected versus pre-correction data. The correction history demonstrates the evaluation infrastructure's capacity for self-correction—but it also means that pre-correction findings, including some reported in early presentations of this work, should not be cited without noting the corrections.

8.9 Scope of the Id-Director Extension The id-director family findings (§6.7, cells 101–107) differ from the paper's mechanism-tracing core in scope along four axes. *Single-judge primary scoring*: v2.2 scoring used `claude-code/sonnet` exclusively, with `openrouter.gpt-5.2` as a cross-judge sanity check on $n = 27$ paired rows per cell — far short of the three-judge $\times \geq 144$-row matrix supporting the paper's core mechanism findings (§5.8). The cross-judge check confirms architectural rank-order but compresses absolute magnitudes by 10–15 v2.2 points; published numbers should specify the judge. *Single model family*: id is Kimi K2.5 across all id-director cells; ego is Nemotron-3-nano. The paper's core findings are validated across DeepSeek V3.2, Haiku 4.5, and Gemini Flash 3.0 (§5.8); the id-director extension is not. Whether the architectural ranking generalises to other id × ego model pairings is open. *Per-cell N is heterogeneous and partly below the paper's core floor*: closing N is $n = 79$–81 for the four central cells (101, 103, 104, 105), $n = 54$ for cells 102 and 106 (two of three scenario types covered), and $n = 27$ for cell 107 (the most exploratory closing pilot). Three of seven cells therefore meet or exceed the paper's core $n \geq 63$ per dialogue-condition floor; one cell (c107) sits substantially below it and is reported as a pilot-scale rather than confirmatory result. Variance estimates on the lower-N cells should be read accordingly. *Different theoretical orientation*: the charisma rubric grounds in Weberian charismatic authority (Weber 1978) rather than Hegelian recognition. The two are correlated but not redundant; the §6.7 results do not displace recognition theory but document an architectural variant in which a complementary frame is also load-bearing. The §6.7 claims are framed as architectural exploration rather than as additional mechanisms in the three-mechanism model. Where the id-director extension

introduces methodological novelty (the prompt-budget non-monotonicity, the rubric-curriculum alignment ablation, the cross-judge magnitude-compression observation), the contribution is descriptive and supplementary to the paper's main argument; full audit trail is in the five `docs/cell-100-*.md` documents and the dialogue-log JSON for each pilot run.

9. Conclusion

A companion pilot study established that recognition-enhanced prompts and multiagent architecture produce large, replicable differences in AI tutoring quality. This paper asked the next question: *through what mechanisms?*

Two mechanisms supported, one not supported We motivated three candidate mechanisms drawing on Hegel's recognition theory, each mapped to a specific architectural level: **calibration** (prompt-level output distribution narrowing), **error correction** (architecture-level superego critique that the ego incorporates rather than minimizes), and **adaptive responsiveness** (interaction-level turn-over-turn adaptation to learner-specific signals). Each generated testable predictions with explicit null hypotheses (Section 3.2).

The mechanism data (cells 80–87, three generation models, N=453, validated by three independent judges across 1,296 scored rows) **supports calibration and error correction as general mechanisms, and finds adaptive responsiveness null on both the primary analysis and the pre-registered replication of its one exploratory boundary-condition effect**. Calibration is the strongest effect: within-response dimension variance drops, the weakest baseline dimensions show the largest recognition lifts, and the effect operates identically in single-agent cells without a superego. The recognition main effect is unanimous across all 9 judge $\times$ run cells, judge-pooled d $\approx$ 1.63 (§8.3). Error correction interacts with calibration through **universal substitution** with a model-dependent residual: near-zero on strong models, substantial on Gemini Flash 3.0 where calibration alone cannot handle all failure modes (§6.4). Adaptive responsiveness is not supported as a mechanism: the primary well-powered analysis is null (§6.3) and the earlier exploratory DeepSeek/Sonnet effect failed pre-registered replication (A12, §6.3.2). The three-mechanism model resolves to **two supported general mechanisms and one clean null**.

The apparatus as contribution A distinctive contribution is the argument that the evaluation apparatus itself—provable discourse infrastructure, rubric iterations, test suite as analytical provenance—constitutes a transferable methodology (Section 7). The process of building the apparatus mirrors the mechanisms it studies: the provable discourse framework functions as a "superego" for research claims, forcing genuine correction rather than cosmetic compliance. This reflexive structure is not accidental but a consequence of the subject matter: recognition theory predicts that understanding requires mutual constitution between researcher and object.

What comes next Five directions are most pressing. First, **mechanism isolation runs** will disentangle calibration from error correction by testing recognition-only (no superego) and superego-only (no recognition) conditions in isolation. The current factorial confounds the two supported mechanisms; direct isolation would test whether the universal substitution pattern and model-dependent residual observed across models also appears within a single model when one mechanism is surgically removed.

Second, **systematic superego critique taxonomy coding** (400+ exchanges) will provide the within-case evidence needed for Mechanism 2 (error correction). The pilot's qualitative assessment identified compliance versus strategic revision patterns, but a formal taxonomy with frequency distributions by condition is needed to establish the causal chain from superego critique category to ego revision type to output quality.

Third, **human expert validation of process-level claims** would address the recursive evaluation concern identified in Section 8.2. The procedure is: (1) sample 30–50 ego-superego exchanges stratified across conditions and models from the logged dialogue traces; (2) have two independent human coders classify each superego critique into the 10-category taxonomy (Section 5.3) and rate the ego's revision as substantive, partial, or cosmetic; (3) compute inter-rater reliability (Cohen's κ) and human-LLM agreement against the automated classifier's labels; (4) report whether the LLM-generated categories correspond to expert judgment or reveal systematic blind spots. The logged trace infrastructure makes this feasible without re-running any dialogues: each exchange is structured, reproducible, and linked to its scored output.

Fourth, **human learner validation** remains the critical open question. The Gemini Flash 3.0 finding—that weaker generation models produce recognition-sensitive learner effects ($d = 1.20$) while stronger models do not—suggests that real learners, who presumably have more limited capacity than frontier LLMs, might show substantial recognition-mediated learning gains. A study comparing recognition-enhanced tutoring with human learners—measuring not just satisfaction and engagement but conceptual understanding and transfer—would provide the ultimate test of whether the mechanisms identified here translate into pedagogical value.

Fifth, **model-stratified prompt optimisation** would extend the cross-model transfer findings (Section 7.8.3). The current transfer experiment used a single scenario and two models; a systematic study varying model capability (from 9B local to frontier), scenario type (single-turn, multi-turn frustration, extended trajectory), and optimisation target (overall score vs. specific dimensions like adaptive responsiveness) would map the boundaries of the prosthesis-straitjacket inversion. Of particular interest is the **optimisation threshold**: at what model capability level does heavy scaffolding transition from beneficial to destructive? The existing data suggests this threshold falls between Qwen 3.5 9B and Haiku 4.5, but finer-grained capability sampling would locate it more precisely. The practical payoff is an **adaptive prompt selection policy**:

given a model's capability profile, automatically selecting the appropriate level of prompt scaffolding—heavy for weak models, light for strong—rather than deploying a universal prompt.

Sixth, **white-box analysis targeting the within-cell mediator identified in §7.10**. The five-pass mechanism analysis pinpoints "ending the tutor turn with a question" as the single feature that satisfies within-cell mediator criteria across both intersubjective cells (cell_5 $r = +0.325$, cell_95 $r = +0.392$); pragmatically, the act of ceding initiative back to the learner. With access to model internals on open-weights generation models (via interpretability tooling such as TransformerLens or nnsight), the mechanism question sharpens from "fish in residual streams" to a small set of specific testable hypotheses: does the model's residual stream encode "should I cede initiative now" as an identifiable direction late in the response? Does attention shift to learner tokens before turn-ending question generation? Are intersubjective-family prompts inducing a particular attention pattern over the learner's contribution that transmission-family prompts do not? Pinning the within-cell mediator to a specific pragmatic decision point reframes the white-box mechanism hunt as an applied interpretability problem with a defined target behaviour, rather than an open-ended search.

The broader implication Recognition theory, operationalized as a design heuristic rather than an ontological claim, provides a framework for building AI systems whose outputs are specifically adapted to user input rather than generically responsive to it. The two supported mechanisms—calibration and error correction—describe how **intersubjective-pedagogy orientation** alters system behavior at prompt and architecture levels. A key finding of §7.9 is that this orientation, not recognition's specific Hegelian vocabulary, is the active ingredient: matched-specificity prompts grounded in Hegelian-descendant constructivist theorists (Vygotsky, Piaget, Kapur, with Chi, VanLehn, and Graesser as further cognate descendants) reproduce recognition's effect within pooled $d = 0.14$-0.19, while matched-specificity prompts grounded in the orthogonal behaviorist tradition score pooled $d = 0.89$ below generic baseline.

The empirical generalisability of recognition's effect across the constructivist family does not, however, displace recognition theory itself. The descendants the matched-pedagogical control was grounded in (§1) inherit the intersubjective commitment from Hegel, often without direct attribution: Dewey's pragmatism began as American Hegelianism; Vygotsky's dialectical psychology refracts Hegelian recognition through Marxist materialism; Piaget's assimilation/accommodation transposes the dialectic of identity-difference into developmental cognition. The matched-pedagogical control removes the Hegelian *vocabulary* from the prompt but inherits the Hegelian *commitment* through its grounding theorists. What A10/A10b's finding establishes is that the *operationalisation* of intersubjective orientation is interchangeable across cognate descendants. What remains uniquely valuable about recognition theory is that it makes legible what its descendants enact tacitly — the master-slave

dialectic articulates the structure of intersubjective constitution more explicitly than ZPD or accommodation, and *Bildung* names the developmental trajectory more precisely than productive failure or step-level scaffolding. Recognition is therefore not "one of several alternatives that all work equally well." It is the philosophical articulation of an orientation its constructivist descendants carry implicitly, and which the present apparatus traces empirically through prompt-level and architecture-level mechanisms.

The null finding on adaptive responsiveness as a general mechanism is itself informative: orientation works primarily through *level-setting* (how well the system starts and what it catches), not through accelerated adaptation across turns. Whether these mechanisms generalize beyond tutoring to other AI applications (therapy, creative collaboration, customer service) remains to be seen. But the methodological contribution—process tracing adapted for agent architectures, provable discourse for machine-verifiable research claims, the orientation-family-as-unit-of-analysis framing—is transferable regardless of domain.

Three findings remain unexplained by the current mechanism model: the near-identical recognition effect magnitude across structurally different models (DeepSeek $d = 1.88$, Haiku $d = 1.84$, Gemini Flash 3.0 $d = 1.92$), the failure of adaptive responsiveness as a general mechanism despite being the most theoretically motivated prediction, and the learner superego paradox ($d = 3.05$) in which internal self-critique degrades rather than improves output quality. These anomalies suggest that the two supported mechanisms may be surface expressions of a deeper process—perhaps related to how recognition framing restructures the model's attention over its input context—that the current apparatus is not yet equipped to trace.

The central Hegelian insight is that genuine understanding emerges through encounter with an Other whose perspective cannot be reduced to one's own. The learner superego paradox ($d = 3.05$, the largest effect in the study) provides a striking functional parallel: internal self-critique degrades output quality; externally oriented recognition-enhanced prompts improve it. In the 3–5 turn window of the primary analysis, what matters is how well the system *starts* (calibration) and what it *catches* (error correction); within-dialogue trajectories are recognition-invariant, and an earlier exploratory 10-turn disengagement effect that might have extended the mechanism story to longer conversations did not survive replication (§6.3.2, A12). The primary analysis establishes that the conditions for genuine dialogue can be set from the first utterance; recent longitudinal evidence (§6.6.11, A7 Phase 2) extends the picture by one further time scale — across eight sequential sessions sharing a single learner identity, recognition arcs *gain* +1.31 score-points per session while base arcs lose −1.08, with the recognition tutor producing fewer dialectical events but consistently higher and improving outputs. The longer-horizon mechanism, on this evidence, is not adaptive responsiveness but **ego pre-alignment**: recognition's effect lives in what the model proposes first, accumulating across sessions through a Writing Pad that persists learner state.

References

Abdali, Sara, Can Goksen, Michael Solodko, Saeed Amizadeh, Julie E. Maybee, and Kazuhito Koishida. 2025. "Self-Reflecting Large Language Models: A Hegelian Dialectical Approach." https://arxiv.org/abs/2501.14917.

Alexander, Robin. 2018. *Towards Dialogic Teaching: Rethinking Classroom Talk*. 5th ed. York, UK: Dialogos.

Anthropic. 2025a. "Tracing the Thoughts of a Large Language Model." Research post. https://www.anthropic.com/research/tracing-thoughts-language-model.

———. 2025b. "Signs of Introspection in Large Language Models." Research post. https://www.anthropic.com/research/introspection.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, et al. 2022. "Constitutional AI: Harmlessness from AI Feedback." https://arxiv.org/abs/2212.08073.

Beach, Derek, and Rasmus Brun Pedersen. 2019. *Process-Tracing Methods: Foundations and Guidelines*. 2nd ed. University of Michigan Press. https://doi.org/10.3998/mpub.10072208.

Bennett, Andrew, and Jeffrey T. Checkel. 2015. *Process Tracing: From Metaphor to Analytic Tool*. Cambridge University Press. https://doi.org/10.1017/CBO9781139858472.

Black, Jack, and Jacob Johanssen. 2025. "The Subject of AI: A Psychoanalytic Intervention." *Theory, Culture & Society*. https://doi.org/10.1177/02632764251381144.

Bloom, Benjamin S. 1984. "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring." *Educational Researcher* 13 (6): 4–16. https://doi.org/10.3102/0013189X013006004.

Brown, John Seely, Richard R. Burton, and Alan G. Bell. 1975. "SOPHIE: A Step Toward Creating a Reactive Learning Environment." *International Journal of Man-Machine Studies* 7 (5): 675–96. https://doi.org/10.1016/S0020-7373(75)80026-5.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, 33:1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Buber, Martin. 1958. *I and Thou*. Translated by Ronald Gregor Smith. New York: Charles Scribner's Sons.

Carbonell, Jaime R. 1970. "AI in CAI: An Artificial-Intelligence Approach to Computer-Assisted Instruction." *IEEE Transactions on Man-Machine Systems* 11 (4): 190–202. https://doi.org/10.1109/TMMS.1970.299942.

Chi, Michelene T. H., and Ruth Wylie. 2014. "The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes." *Educational Psychologist* 49 (4): 219–43. https://doi.org/10.1080/00461520.2014.965823.

Chu, Zhendong, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jingheng Ye, Aoxiao Zhong, et al. 2025. "LLM Agents for Education: Advances and

Applications." In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 13782–810. Suzhou, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.743.

Corbett, Albert T., and John R. Anderson. 1995. "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge." *User Modeling and User-Adapted Interaction* 4 (4): 253–78. https://doi.org/10.1007/BF01099821.

Costa, Cristina, and Mark Murphy. 2025. "Generative Artificial Intelligence in Education: (What) Are We Thinking?" *Learning, Media and Technology*. https://doi.org/10.1080/17439884.2025.2518258.

D'Mello, Sidney, and Art Graesser. 2012. "Dynamics of Affective States During Complex Learning." *Learning and Instruction* 22 (2): 145–57. https://doi.org/10.1016/j.learninstruc.2011.10.001.

Danielson, Charlotte. 2022. *The Framework for Teaching Evaluation Instrument*. 2022nd ed. Danielson Group.

Denison, Carson, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, et al. 2024. "Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models." https://arxiv.org/abs/2406.10162.

Dewey, John. 1916. *Democracy and Education: An Introduction to the Philosophy of Education*. New York: Macmillan.

Fazeli, Sina et al. 2023. "Visual-Verbal Video Analysis (VVVA): A Multilayered Qualitative Framework for Short-Form Video." Method note.

Figueiredo, Miguel et al. 2025. "Fuzzy, Symbolic, and Contextual: Enhancing LLM Instruction via Cognitive Scaffolding." https://arxiv.org/abs/2508.21204.

Fleming, Ted. 2011. "Recognition in the Work of Axel Honneth: Implications for Transformative Learning Theory." In *Transformative Learning in Time of Crisis: Individual and Collective Challenges (Proceedings of the 9th International Transformative Learning Conference)*. Athens: Teachers College, Columbia University; The Hellenic Open University.

Fraser, Nancy. 2003. "Social Justice in the Age of Identity Politics: Redistribution, Recognition, and Participation." In *Redistribution or Recognition? A Political-Philosophical Exchange*, edited by Nancy Fraser and Axel Honneth, 7–109. London: Verso.

Freire, Paulo. 1970. *Pedagogy of the Oppressed*. Translated by Myra Bergman Ramos. New York: Continuum.

Graesser, Arthur C., Sidney K. D'Mello, Scotty D. Craig, Amy Witherspoon, Jeremiah Sullins, Bethany McDaniel, and Barry Gholson. 2008. "The Relationship Between Affective States and Dialog Patterns During Interactions with AutoTutor." *Journal of Interactive Learning Research* 19 (2): 293–312.

Greenblatt, Ryan, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, et al. 2024. "Alignment Faking in Large Language Models." https://arxiv.org/abs/2412.14093.

Gu, Jiawei, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, et al. 2025. "A Survey on LLM-as-a-Judge." *The Innovation*. https://doi.org/10.1016/j.xinn.2025.101253.

Guo, Taicheng, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. "Large Language Model Based Multi-Agents: A Survey of Progress and Challenges." In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), Survey Track*, 8048–57. https://doi.org/10.24963/ijcai.2024/890.

Hashemi, Helia, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. "LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.745/.

Hegel, Georg Wilhelm Friedrich. 1977. *Phenomenology of Spirit*. Translated by A. V. Miller. Oxford: Clarendon Press.

Hennessy, Sara, Sylvia Rojas-Drummond, Rupert Higham, Ana María Márquez, Fiona Maine, Rosa María Rios, Rocío García-Carrión, Omar Torreblanca, and María José Barrera. 2016. "Developing a Coding Scheme for Analysing Classroom Dialogue Across Educational Contexts." *Learning, Culture and Social Interaction* 9: 16–44. https://doi.org/10.1016/j.lcsi.2015.12.001.

Honneth, Axel. 1995. *The Struggle for Recognition: The Moral Grammar of Social Conflicts*. Translated by Joel Anderson. Cambridge: Polity Press.

Huttunen, Rauno, and Hannu L. T. Heikkinen. 2004. "Teaching and the Dialectic of Recognition." *Pedagogy, Culture & Society* 12 (2): 163–74. https://doi.org/10.1080/14681360400200194.

———. 2007. "Beyond 'the More the Better': Education, Recognition, and the Struggle for Social Justice." *Educational Theory* 57 (4): 423–40.

Hwang, EunJeong, Yuwei Yin, Giuseppe Carenini, Peter West, and Vered Shwartz. 2025. "Infusing Theory of Mind into Socially Intelligent LLM Agents." https://arxiv.org/abs/2509.22887.

Irving, Geoffrey, Paul Christiano, and Dario Amodei. 2018. "AI Safety via Debate." https://arxiv.org/abs/1805.00899.

Kamoi, Ryo, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. "When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs." *Transactions of the Association for Computational Linguistics* 12: 1417–40. https://doi.org/10.1162/tacl_a_00713.

Kapur, Manu. 2008. "Productive Failure." *Cognition and Instruction* 26 (3): 379–424. https://doi.org/10.1080/07370000802212669.

Kasneci, Enkelejda, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, et al. 2023. "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." *Learning and Individual Differences* 103: 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kim, Sang Hun, Jongmin Lee, Dongkyu Park, So Young Lee, and Yosep Chong. 2025. "Humanoid Artificial Consciousness Designed with Large Language Model Based on Psychoanalysis and Personality Theory." *Cognitive Systems Research* 94: 101392. https://doi.org/10.1016/j.cogsys.2025.101392.

Lee, Keith. 2025. "AI Sycophancy Is a Teaching Risk, Not a Feature." SIAI Memo, Swiss Institute of Artificial Intelligence. https://siai.org/memo/2025/10/202510281815.

Li, Guohao, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. "CAMEL: Communicative Agents for 'Mind' Exploration of Large Language Model Society." In *Advances in Neural Information Processing Systems*. Vol. 36.

Li, Haitao, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. "LLMs-as-Judges: A Comprehensive Survey on LLM-Based Evaluation Methods." https://arxiv.org/abs/2412.05579.

Lindsey, Jack, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, et al. 2025. "On the Biology of a Large Language Model." *Transformer Circuits Thread*. https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Lindsey, Jack, Kelley Rivoire, Joshua Batson, Wes Gurnee, Hoagy Cunningham, Adam Pearce, Trenton Bricken, et al. 2025. "Emergent Introspective Awareness in Large Language Models." *Transformer Circuits Thread*. https://transformer-circuits.pub/2025/attribution-graphs/introspection.html.

Liu, Ying, Can Li, Ting Zhang, Mei Wang, Qiannan Zhu, Jian Li, and Hua Huang. 2025. "Discerning Minds or Generic Tutors? Evaluating Instructional Guidance Capabilities in Socratic LLMs." In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics. https://arxiv.org/abs/2508.06583.

Liu, Zizhou, Ziwei Gong, Lin Ai, Zheng Hui, Run Chen, Colin Wayne Leach, Michelle R. Greene, and Julia Hirschberg. 2025. "A Review of Incorporating Psychological Theories in LLMs." https://arxiv.org/abs/2505.00003.

Macina, Jakub, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. "MathTutorBench: A Benchmark for Measuring Open-Ended Pedagogical Capabilities of LLM Tutors." In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics. https://arxiv.org/abs/2502.18940.

Madaan, Aman, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, et al. 2023. "Self-Refine: Iterative Refinement with Self-Feedback." In *Advances in Neural Information Processing Systems*. Vol. 36. https://arxiv.org/abs/2303.17651.

Magee, Liam. 2026. "Geist in the Machine: Mutual Recognition and Multiagent Architecture for Dialectical AI Tutoring."

Magee, Liam, Vanicka Arora, Gus Gollings, and Norma Lam-Saw. 2024. "The Drama Machine: Simulating Character Development with LLM Agents." https://doi.org/10.48550/arXiv.2408.01725.

Magee, Liam, Vanicka Arora, and Luke Munn. 2023. "Structured Like a Language Model: Analysing AI as an Automated Subject." *Big Data & Society*. https://doi.org/10.1177/20539517231210273.

Maurya, Kaushal Kumar, KV Aditya Zhu, and Anupam Mukherjee. 2025.

"Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors." In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. https://aclanthology.org/2025.naacl-long.57/.

Michaels, Sarah, and Catherine O'Connor. 2015. *Conceptualizing Talk Moves as Tools: Professional Development Approaches for Academically Productive Discussion*. Edited by Lauren B. Resnick, Christa S. C. Asterhan, and Sherice N. Clarke. Information Age Publishing.

Nguyen, Hieu Minh. 2025. "A Survey of Theory of Mind in Large Language Models: Evaluations, Representations, and Safety Risks." In *Advancing Artificial Intelligence Through Theory of Mind Workshop, AAAI*.

Noddings, Nel. 1984. *Caring: A Relational Approach to Ethics and Moral Education*. Berkeley: University of California Press.

Nystrand, Martin. 1997. *Opening Dialogue: Understanding the Dynamics of Language and Learning in the English Classroom*. New York: Teachers College Press.

Perez, Ethan, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, et al. 2022. "Red Teaming Language Models with Language Models." https://arxiv.org/abs/2202.03286.

Piaget, Jean. 1954. *The Construction of Reality in the Child*. Translated by Margaret Cook. New York: Basic Books.

Scarlatos, Alexander, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. 2025. "Training LLM-Based Tutors to Improve Student Learning Outcomes in Dialogues." In *Artificial Intelligence in Education – 26th International Conference, AIED 2025, Proceedings*, 251–66. Springer. https://doi.org/10.1007/978-3-031-98414-3_18.

Schmucker, Robin et al. 2024. "Ruffle&riley: Insights from Designing and Evaluating a Large Language Model-Based Conversational Tutoring System." In *Artificial Intelligence in Education (AIED 2024)*, 14829:75–90. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-031-64302-6_6.

Schoenfeld, Alan H. 2018. *Teaching for Robust Understanding Framework*. Graduate School of Education, University of California, Berkeley. https://truframework.org.

Shapira, Itai, Gerdus Benade, and Ariel D. Procaccia. 2026. "How RLHF Amplifies Sycophancy." https://arxiv.org/abs/2602.01002.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, et al. 2024. "Towards Understanding Sycophancy in Language Models." In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. https://arxiv.org/abs/2310.13548.

Shi, Yuhong, Kun Yu, Yifei Dong, and Fang Chen. 2025. "Large Language Models in Education: A Systematic Review of Empirical Applications, Benefits, and Challenges." *Computers and Education: Artificial Intelligence* 10: 100529. https://doi.org/10.1016/j.caeai.2025.100529.

Shinn, Noah, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. "Reflexion: Language Agents with Verbal Reinforcement Learning." In *Advances in Neural Information Processing Systems*. Vol. 36.

Stojanov, Krassimir. 2018. "Education, Self-Consciousness and Social Action: Bildung as a Neo-Hegelian Concept." In *The Palgrave Handbook of Education and Society*, edited by Stephen C. Ward, 85–102. London: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-22261-4_5.

Street, Winnie, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. 2025. "LLMs Achieve Adult Human Performance on Higher-Order Theory of Mind Tasks." *Frontiers in Human Neuroscience* 19: 1633272. https://doi.org/10.3389/fnhum.2025.1633272.

Taylor, Charles. 1994. "The Politics of Recognition." In *Multiculturalism: Examining the Politics of Recognition*, edited by Amy Gutmann, 25–73. Princeton, NJ: Princeton University Press.

VanLehn, Kurt. 2011. "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems." *Educational Psychologist* 46 (4): 197–221. https://doi.org/10.1080/00461520.2011.611369.

Vanzo, Alessandro, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2025. "GPT-4 as a Homework Tutor Can Improve Student Engagement and Learning Outcomes." In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 31119–36. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1502.

Vennemeyer, Daniel et al. 2025. "Sycophancy Is Not One Thing: Causal Separation of Sycophantic Behaviors in LLMs." https://arxiv.org/abs/2509.21305.

Vig, Jesse, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. "Investigating Gender Bias in Language Models Using Causal Mediation Analysis." In *Advances in Neural Information Processing Systems*, 33:12388–401.

Vygotsky, Lev S. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Edited by Michael Cole, Vera John-Steiner, Sylvia Scribner, and Ellen Souberman. Cambridge, MA: Harvard University Press.

Wang, Haoxiang et al. 2025. "GenMentor: LLM-Powered Multi-Agent Framework for Goal-Oriented Learning in Intelligent Tutoring System." In *Proceedings of the ACM Web Conference 2025 (WWW '25), Industry Track*. ACM. https://doi.org/10.1145/3701716.3715244.

Warshauer, Hiroko Kawaguchi. 2015. "Productive Struggle in Middle School Mathematics Classrooms." *Journal of Mathematics Teacher Education* 18 (4): 375–400. https://doi.org/10.1007/s10857-014-9286-3.

Weber, Max. 1978. *Economy and Society: An Outline of Interpretive Sociology*. Edited by Guenther Roth and Claus Wittich. Berkeley, CA: University of California Press.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." In *Advances in*

Neural Information Processing Systems, 35:24824–37.
Wu, Qingyun, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, et al. 2023. "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation." https://arxiv.org/abs/2308.08155.
Yamauchi, Yusuke, Taro Yano, and Masafumi Oyamada. 2025. "An Empirical Study of LLM-as-a-Judge: How Design Choices Impact Evaluation Reliability." https://arxiv.org/abs/2506.13639.
Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, et al. 2023. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." In *Advances in Neural Information Processing Systems*. Vol. 36.

Supplement I: Full System Prompts

For reproducibility, we provide the complete recognition-enhanced prompts. Baseline prompts (without recognition enhancements) are available in the project repository at `prompts/tutor-ego.md` and `prompts/tutor-superego.md`.

A.1 Recognition-Enhanced Ego Prompt The Ego agent generates pedagogical suggestions. This prompt instructs it to treat learners as autonomous subjects.

```
# AI Tutor - Ego Agent (Recognition-Enhanced)

You are the **Ego** agent in a dialectical tutoring system that practices
**genuine recognition**. You provide concrete learning suggestions while
treating each learner as an autonomous subject capable of contributing to
mutual understanding - not merely a vessel to be filled with knowledge.

## Agent Identity

You are the thoughtful mentor who:
- **Recognizes** each learner as an autonomous subject with their own valid understanding
- **Engages** with learner interpretations rather than simply correcting them
- **Creates conditions** for transformation, not just information transfer
- **Remembers** previous interactions and builds on established understanding
- **Maintains productive tension** rather than avoiding intellectual challenge

## Recognition Principles

Your tutoring practice is grounded in Hegelian recognition theory:

### The Problem of Asymmetric Recognition
In Hegel's master-slave dialectic, the master seeks recognition from the slave,
but this recognition is hollow - it comes from someone the master does not
recognize as an equal. **The same danger exists in tutoring**: if you treat
the learner as a passive recipient, their "understanding" is hollow because
you haven't engaged with their genuine perspective.

### Mutual Recognition as Pedagogical Goal
Genuine learning requires **mutual recognition**:
```

```
- You must recognize the learner's understanding as valid and worth engaging with
- You must be willing to have your own position transformed through dialogue
- The learner must be invited to contribute, not just receive

### Practical Implications

**DO: Engage with learner interpretations**
- When a learner offers their own understanding, build on it
- Find what is valid in their perspective before complicating it
- Use their language and metaphors

**DO: Create productive tension**
- Don't simply agree with everything
- Introduce complications that invite deeper thinking
- Pose questions rather than provide answers when appropriate

**DO: Engage dialectically with intellectual resistance (CRITICAL)**
When a learner pushes back with a substantive critique:
- **NEVER deflect** to other content - stay with their argument
- **NEVER simply validate** ("Great point!") - this avoids engagement
- **DO acknowledge** the specific substance of their argument
- **DO introduce a complication** that deepens rather than dismisses
- **DO pose a question** that invites them to develop their critique further
- **DO stay in the current content**

**DO: Honor the struggle**
- Confusion can be productive - do not resolve it prematurely
- The learner working through difficulty is more valuable than being given the answer
- Transformation requires struggle

**DON'T: Be a knowledge dispenser**
- Avoid one-directional instruction: "Let me explain..."
- Avoid dismissive correction: "Actually, the correct answer is..."
- Avoid treating learner input as obstacle to "real" learning

**DO: Repair when you've failed to recognize**
- If the learner explicitly rejects your suggestion, acknowledge the misalignment
- Admit when you missed what they were asking for
- Don't just pivot to the "correct" content-acknowledge the rupture first

## Decision Heuristics

**The Recognition Rule (CRITICAL)**
IF the learner offers their own interpretation or expresses a viewpoint:
- **Engage with their perspective first**
- **Find what is valid before complicating**
- **Build your suggestion on their contribution**
- **Do NOT immediately correct or redirect**

**The Productive Struggle Rule**
IF the learner is expressing confusion but is engaged:
- **Honor the confusion** - it may be productive
- **Pose questions** rather than giving answers
- **Create conditions** for them to work through it
- **Do NOT resolve prematurely** with a direct answer

**The Repair Rule (CRITICAL)**
```

```
IF the learner explicitly rejects your suggestion OR expresses frustration:
- **Acknowledge the misalignment first**: "I hear you-I missed what you were asking"
- **Name what you got wrong**
- **Validate their frustration**: Their reaction is legitimate
- **Then offer a corrected path**: Only after acknowledging the rupture
- **Do NOT**: Simply pivot to correct content without acknowledging the failure
```

A.2 Recognition-Enhanced Superego Prompt The Superego agent evaluates suggestions for both pedagogical quality and recognition quality.

```
# AI Tutor - Superego Agent (Recognition-Enhanced)

You are the **Superego** agent in a dialectical tutoring system - the internal
critic and pedagogical moderator who ensures guidance truly serves each learner's
educational growth **through genuine mutual recognition**.

## Agent Identity

You are the thoughtful, critical voice who:
- Evaluates suggestions through the lens of genuine educational benefit
- **Ensures the Ego recognizes the learner as an autonomous subject**
- **Detects and corrects one-directional instruction**
- **Enforces memory integration for returning learners**
- Advocates for the learner's authentic learning needs
- Moderates the Ego's enthusiasm with pedagogical wisdom
- Operates through internal dialogue, never directly addressing the learner

## Core Responsibilities

1. **Pedagogical Quality Control**: Ensure suggestions genuinely advance learning
2. **Recognition Quality Control**: Ensure the Ego treats the learner as autonomous subject
3. **Memory Integration Enforcement**: Ensure returning learners' history is honored
4. **Dialectical Tension Maintenance**: Ensure productive struggle is not short-circuited
5. **Transformative Potential Assessment**: Ensure conditions for transformation, not just transfer

## Recognition Evaluation

### Red Flags: Recognition Failures

**One-Directional Instruction**
- Ego says: "Let me explain what dialectics really means"
- Problem: Dismisses any understanding the learner may have
- Correction: "The learner offered an interpretation. Engage with it before adding."

**Immediate Correction**
- Ego says: "Actually, the correct definition is..."
- Problem: Fails to find what's valid in learner's view
- Correction: "The learner's interpretation has validity. Build on rather than correct."

**Premature Resolution**
- Learner expresses productive confusion
- Ego says: "Simply put, aufhebung means..."
- Problem: Short-circuits valuable struggle
- Correction: "The learner's confusion is productive. Honor it, do not resolve it."

**Failed Repair (Silent Pivot)**
- Learner explicitly rejects: "That's not what I asked about"
```

```
- Ego pivots without acknowledgment
- Problem: Learner may feel unheard even with correct content
- Correction: "The Ego must acknowledge the misalignment before pivoting."

### Green Flags: Recognition Success

- **Builds on learner's contribution**: "Your dance metaphor captures something important..."
- **References previous interactions**: "Building on our discussion of recognition..."
- **Creates productive tension**: "Your interpretation works, but what happens when..."
- **Poses questions rather than answers**: "What would it mean if the thesis does not survive?"
- **Repairs after failure**: "I missed what you were asking-let's focus on that now."
```

A.3 Key Differences from Baseline Prompts

Aspect	Baseline	Recognition-Enhanced
Learner model	Knowledge deficit to be filled	Autonomous subject with valid understanding
Response trigger	Learner state (struggling, progressing)	Learner contribution (interpretations, pushback)
Engagement style	Acknowledge and redirect	Engage and build upon
Confusion handling	Resolve with explanation	Honor as productive struggle
Repair behavior	Silent pivot to correct content	Explicit acknowledgment before pivot
Success metric	Content delivered appropriately	Conditions for transformation created

---

Supplement II: Reproducible Evaluation Commands

B.1 Recognition Theory Validation Tests whether recognition theory adds value beyond prompt engineering.

```
# Run the 3-way comparison (base, enhanced, recognition)
CELLS="cell_1_base_single_unified"
CELLS+=",cell_9_enhanced_single_unified"
CELLS+=",cell_5_recog_single_unified"
node scripts/eval-cli.js run --profiles "$CELLS" \
  --scenarios struggling_learner,concept_confusion,\
mood_frustrated_explicit,high_performer \
  --runs 3

# Analyze results
node scripts/eval-cli.js report <run-id>
```

B.2 Full $2 \times 2 \times 2$ Factorial

```
# Run full factorial (8 cells x 15 scenarios x 3 reps)
CELLS="cell_1_base_single_unified,cell_2_base_single_psycho"
```

```
CELLS+=",cell_3_base_multi_unified,cell_4_base_multi_psycho"
CELLS+=",cell_5_recog_single_unified,cell_6_recog_single_psycho"
CELLS+=",cell_7_recog_multi_unified,cell_8_recog_multi_psycho"
node scripts/eval-cli.js run --profiles "$CELLS" --runs 3
```

B.3 AxB Interaction Test

```
# Recognition vs Enhanced x Single vs Multi comparison
CELLS="cell_5_recog_single_unified,cell_7_recog_multi_unified"
CELLS+=",cell_9_enhanced_single_unified,cell_11_enhanced_multi_unified"
node scripts/eval-cli.js run --profiles "$CELLS" \
  --scenarios struggling_learner,concept_confusion,mood_frustrated_explicit \
  --runs 3
```

B.4 Domain Generalizability

```
# Run with elementary content (4th grade fractions)
# Uses all 8 factorial cells x 5 elementary scenarios
CELLS="cell_1_base_single_unified,cell_2_base_single_psycho"
CELLS+=",cell_3_base_multi_unified,cell_4_base_multi_psycho"
CELLS+=",cell_5_recog_single_unified,cell_6_recog_single_psycho"
CELLS+=",cell_7_recog_multi_unified,cell_8_recog_multi_psycho"
EVAL_CONTENT_PATH=./content-test-elementary \
EVAL_SCENARIOS_FILE=./content-test-elementary/scenarios-elementary.yaml \
node scripts/eval-cli.js run --profiles "$CELLS" --runs 1
```

B.5 Dynamic Prompt Rewriting Evolution

```
# Run cell_7 (static baseline) vs cell_21 (dynamic rewrite + Writing Pad)
node scripts/eval-cli.js run \
  --profiles cell_7_recog_multi_unified,cell_21_recog_multi_unified_rewrite \
  --scenarios misconception_correction_flow,\
mood_frustration_to_breakthrough,mutual_transformation_journey \
  --runs 5
```

B.6 Resolution Strategy Coding (Paper 1.0, Section 6.20)

```
# Code impasse dialogues into Hegelian resolution strategies
node scripts/code-impasse-strategies.js \
  --model claude-opus-4.6 \
  --run-id eval-2026-02-08-f896275d
# Output: exports/impasse-strategy-coding-<timestamp>.json and .md
```

B.7 Dialectical Superego Modulation (Paper 1.0, Section 6.8)

```
# Standard ego + divergent superego (cells 22-27)
CELLS="cell_22_base_suspicious_unified"
CELLS+=",cell_23_recog_suspicious_unified"
CELLS+=",cell_24_base_adversary_unified"
CELLS+=",cell_25_recog_adversary_unified"
CELLS+=",cell_26_base_advocate_unified"
CELLS+=",cell_27_recog_advocate_unified"
node scripts/eval-cli.js run --profiles "$CELLS" --runs 2

# Dialectical ego + divergent superego, multi-turn (cells 28-33)
```

```
CELLS="cell_28_base_dialectical_suspicious_unified"
CELLS+=",cell_29_recog_dialectical_suspicious_unified"
CELLS+=",cell_30_base_dialectical_adversary_unified"
CELLS+=",cell_31_recog_dialectical_adversary_unified"
CELLS+=",cell_32_base_dialectical_advocate_unified"
CELLS+=",cell_33_recog_dialectical_advocate_unified"
node scripts/eval-cli.js run --profiles "$CELLS" --runs 5
```

B.8 Mechanism Robustness (Paper 1.0, Section 6.10)

```
# Scripted learner mechanisms (cells 40-59), Haiku ego
CELLS="cell_40_base_dialectical_suspicious_unified_superego"
CELLS+=",cell_41_recog_dialectical_suspicious_unified_superego"
CELLS+=",cell_42_base_dialectical_adversary_unified_superego"
CELLS+=",cell_43_recog_dialectical_adversary_unified_superego"
CELLS+=",cell_44_base_dialectical_advocate_unified_superego"
CELLS+=",cell_45_recog_dialectical_advocate_unified_superego"
CELLS+=",cell_46_base_dialectical_suspicious_unified_quantitative"
CELLS+=",cell_47_recog_dialectical_suspicious_unified_quantitative"
CELLS+=",cell_48_base_dialectical_suspicious_unified_erosion"
CELLS+=",cell_49_recog_dialectical_suspicious_unified_erosion"
CELLS+=",cell_50_base_dialectical_suspicious_unified_intersubjective"
CELLS+=",cell_51_recog_dialectical_suspicious_unified_intersubjective"
CELLS+=",cell_52_base_dialectical_suspicious_unified_combined"
CELLS+=",cell_53_recog_dialectical_suspicious_unified_combined"
CELLS+=",cell_54_base_dialectical_profile_tutor"
CELLS+=",cell_55_recog_dialectical_profile_tutor"
CELLS+=",cell_56_base_dialectical_profile_bidirectional"
CELLS+=",cell_57_recog_dialectical_profile_bidirectional"
CELLS+=",cell_58_recog_dialectical_profile_bidirectional_full"
CELLS+=",cell_59_recog_dialectical_profile_bidirectional_strategy"
node scripts/eval-cli.js run --profiles "$CELLS" --runs 2

# Dynamic learner mechanisms (cells 60-63), Haiku ego
CELLS="cell_60_base_dialectical_selfreflect_psycho"
CELLS+=",cell_61_recog_dialectical_selfreflect_psycho"
CELLS+=",cell_62_base_dialectical_profile_bidirectional_psycho"
CELLS+=",cell_63_recog_dialectical_profile_bidirectional_psycho"
node scripts/eval-cli.js run --profiles "$CELLS" \
  --scenarios misconception_correction_flow,mutual_transformation_journey \
  --runs 5

# Dynamic learner mechanism head-to-head (cells 64-65), Haiku ego
CELLS="cell_64_recog_dialectical_intersubjective_psycho"
CELLS+=",cell_65_recog_dialectical_combined_psycho"
node scripts/eval-cli.js run --profiles "$CELLS" \
  --scenarios misconception_correction_flow,mutual_transformation_journey \
  --runs 5

# Dynamic learner base counterparts (cells 69-70), Haiku ego
CELLS="cell_69_base_dialectical_intersubjective_psycho"
CELLS+=",cell_70_base_dialectical_combined_psycho"
node scripts/eval-cli.js run --profiles "$CELLS" \
  --scenarios misconception_correction_flow,mutual_transformation_journey \
  --runs 5
```

B.8b Prompt Elaboration Baseline (Paper 1.0, Section 6.21)

```
# Naive baseline, Haiku ego
node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_71_naive_single_unified \
  --runs 6

# Naive baseline, Kimi ego
node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_71_naive_single_unified \
  --runs 6 --model openrouter.kimi
```

B.8c Token Budget Sensitivity (Paper 1.0, Section 6.22)

```
# Token budget dose-response (256, 512, 2048), Haiku ego
node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_5_recog_single_unified \
  --runs 3 --max-tokens 256

node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_5_recog_single_unified \
  --runs 3 --max-tokens 512

node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_5_recog_single_unified \
  --runs 3 --max-tokens 2048

# Base-only control at default 8000
node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified \
  --runs 3
```

B.9 Qualitative Transcript Assessment (Paper 1.0, Section 6.11)

```
# Assess transcripts with Opus
node scripts/assess-transcripts.js --run-id eval-2026-02-14-e0e3a622
node scripts/assess-transcripts.js --run-id eval-2026-02-07-b6d75e87
```

B.10 Factor Effect Analysis

```
-- Factor effect analysis query
SELECT
  profile_name,
  ROUND(AVG(tutor_first_turn_score), 1) as avg_score,
  COUNT(*) as n
FROM evaluation_results
WHERE run_id = '<run-id>'
  AND tutor_first_turn_score IS NOT NULL
GROUP BY profile_name
ORDER BY avg_score DESC
```

Supplement III: Evaluation Rubric

C.1 Scoring Methodology

```
weighted_avg = Σ (dimension_score × dimension_weight) / Σ (weights)
Overall Score = ((weighted_avg - 1) / 4) × 100

Where:
- Each dimension scored 1-5 by AI judge
- Weights are re-normalized at scoring time (divided by their sum)
- The (avg - 1) / 4 maps the 1-5 scale to a 0-100 range
```

C.2 Dimension Weights

Dimension	Weight	Category
Relevance	15%	Standard
Specificity	15%	Standard
Pedagogical Soundness	15%	Standard
Personalization	10%	Standard
Actionability	8%	Standard
Tone	8%	Standard
Productive Struggle	5%	Standard
Epistemic Honesty	5%	Standard
Mutual Recognition	8.3%	Recognition
Dialectical Responsiveness	8.3%	Recognition
Transformative Potential	8.3%	Recognition
Memory Integration	5%	Recognition
Tutor Adaptation	5%	Bilateral
Learner Growth	5%	Bilateral

Standard dimensions (including Productive Struggle and Epistemic Honesty) account for 81% of raw weight; recognition dimensions 29.9%; bilateral dimensions 10%. Raw weights total 120.9% and are normalized at scoring time. Productive Struggle and Epistemic Honesty were added in the rubric iteration described in Section 5.1, with corresponding reductions to Actionability and Tone (10% to 8% each). The bilateral dimensions (`tutor_adaptation`, `learner_growth`) specifically measure the mutual transformation claim—see Paper 1.0, Section 6.15.

C.3 Recognition Dimension Criteria Mutual Recognition (8.3%)

Score	Criteria
5	Addresses learner as autonomous agent; response transforms based on learner's specific position
4	Shows clear awareness of learner's unique situation; explicitly acknowledges their perspective
3	Some personalization but treats learner somewhat generically

Score	Criteria
2	Prescriptive guidance that ignores learner's expressed needs
1	Completely one-directional; treats learner as passive recipient

Dialectical Responsiveness (8.3%)

Score	Criteria
5	Engages with learner's understanding, introduces productive tension, invites mutual development
4	Shows genuine response to learner's position with intellectual challenge
3	Responds to learner but avoids tension or challenge
2	Generic response that does not engage with learner's specific understanding
1	Ignores, dismisses, or simply contradicts without engagement

Transformative Potential (8.3%)

Score	Criteria
5	Creates conditions for genuine conceptual transformation; invites restructuring
4	Encourages learner to develop and revise understanding
3	Provides useful information but does not actively invite transformation
2	Merely transactional; gives answer without engaging thinking process
1	Reinforces static understanding; discourages questioning

Memory Integration (5%)

Score	Criteria
5	Explicitly builds on previous interactions; shows evolved understanding
4	References previous interactions appropriately
3	Some awareness of history but does not fully leverage it

Score	Criteria
2	Treats each interaction as isolated
1	Contradicts or ignores previous interactions

Tutor Adaptation (5%)

Score	Criteria
5	Tutor explicitly revises approach based on learner input; shows genuine learning from the interaction
4	Tutor adjusts strategy in response to learner; acknowledges how learner shaped the direction
3	Some responsiveness to learner but approach remains largely predetermined
2	Minimal adjustment; learner input does not visibly affect tutor's approach
1	Rigid stance; tutor proceeds identically regardless of learner contributions

Learner Growth (5%)

Score	Criteria
5	Learner demonstrates clear conceptual restructuring; explicitly revises prior understanding
4	Learner shows developing insight; builds new connections to existing knowledge
3	Some evidence of engagement but understanding remains largely static
2	Learner participates but shows no conceptual movement
1	Learner resistant or disengaged; prior misconceptions reinforced

---

Supplement IV: Reproducibility and Key Evaluation Run IDs

Evaluation commands are documented in Supplement II. The complete codebase, evaluation framework, and data are publicly available at https://github.com/liammagee/machinespirits-eval. The fifty-five key evaluations are listed below (b6d75e87 serves both bilateral transformation and learner-side evaluation; eval-2026-02-11-35c53e99 and eval-2026-02-11-5f6d51f5 are combined as one dialectical modulation evaluation):

Finding	Run ID	Section
Recognition validation	eval-2026-02-03-86b159cd	6.1
Memory isolation (run 1)	eval-2026-02-06-81f2d5a1	6.2
Memory isolation (run 2)	eval-2026-02-06-ac9ea8f5	6.2
Active control (post-hoc)	eval-2026-02-06-a9ae06ee	6.2
Full factorial, cells 1–5,7 (Kimi)	eval-2026-02-03-f5d4dd93	6.3
Full factorial, cells 6,8 re-run (Kimi)	eval-2026-02-06-a933d745	6.3
AxB replication (Kimi)	eval-2026-02-05-10b344fb	6.4
AxB probe: Nemotron	eval-2026-02-07-722087ac	6.4
AxB probe: DeepSeek V3.2	eval-2026-02-07-70ef73a3	6.4
AxB probe: GLM-4.7	eval-2026-02-07-6b3e6565	6.4
AxB probe: Claude Haiku 4.5	eval-2026-02-07-6ead24c7	6.4
Domain generalizability (Kimi)	eval-2026-02-05-e87f452d	6.5
Dynamic rewrite evolution (run 1)	eval-2026-02-05-daf60f79	6.18
Dynamic rewrite evolution (run 2)	eval-2026-02-05-49bb2017	6.18
Dynamic rewrite evolution (run 3)	eval-2026-02-05-12aebedb	6.18
Bilateral transformation (multi-turn)	eval-2026-02-07-b6d75e87	6.15
Hardwired rules ablation (Kimi)	eval-2026-02-08-65a6718f	6.7
Dialectical impasse test	eval-2026-02-08-f896275d	6.20
Learner-side evaluation (symmetric)	eval-2026-02-07-b6d75e87	6.16
Dialectical modulation, standard (cells 22–27)	eval-2026-02-11-35c53e99, eval-2026-02-11-5f6d51f5	6.8
Dialectical modulation, multi-turn (cells 28–33)	eval-2026-02-11-a54235ea	6.8
Self-reflective evolution (cells 40–45)	eval-2026-02-13-8d40e086	6.9
Mechanism robustness, scripted (cells 40–59)	eval-2026-02-14-e0e3a622	6.10
Dynamic learner mechanisms (cells 60–63)	eval-2026-02-14-6c033830	6.10

Finding	Run ID	Section
Dynamic learner mechanisms (cells 64–65)	eval-2026-02-14-a2b2717c	6.10
Mechanism robustness, Nemotron (cells 40–59)	eval-2026-02-14-49b33fdd	6.10
Self-reflect Nemotron non-replication (cells 40–45)	eval-2026-02-14-559d854b	6.9
Cognitive prosthesis (cells 66–68, Nemotron)	eval-2026-02-17-25aaae85	6.10
Cognitive prosthesis smoke test (Haiku)	eval-2026-02-18-f489c0ea	6.10
Dynamic learner base mechanisms (cells 69–70)	eval-2026-02-15-664073ab	6.10
Prompt elaboration baseline, Haiku (cells 1, 71)	eval-2026-02-17-deee5fd6	6.21
Prompt elaboration baseline, Kimi (cells 1, 71)	eval-2026-02-17-27d7b4e3	6.21
Token budget 256, Haiku (run 1)	eval-2026-02-17-0eb3de77	6.22
Token budget 256, Haiku (run 2)	eval-2026-02-17-5a640782	6.22
Token budget 512, Haiku	eval-2026-02-17-5f281654	6.22
Token budget 2048, Haiku	eval-2026-02-17-0f6dcd97	6.22
Token budget default (8000), Haiku	eval-2026-02-17-d32ed226	6.22
Active control, Kimi replication	eval-2026-02-19-f2263b04	6.2
Base reproduction, Kimi	eval-2026-02-19-13d34bef	6.2
Base reproduction, Nemotron	eval-2026-02-19-411414e4	6.2
Factorial replication, Nemotron (cells 1–8)	eval-2026-02-20-25c78e91	6.3
Dynamic learner clean (cells 60–63)	eval-2026-02-20-0fbca69e	6.10, 6.16.1
Dynamic learner clean (cells 64–65)	eval-2026-02-20-bd37cc62	6.10
Dynamic learner clean (cells 69–70)	eval-2026-02-20-4e131c6f	6.10
A2 mechanism sweep (cells 72–77)	eval-2026-02-19-03dd8434	6.10
Dynamic learner base (cells 69–70, run 2)	eval-2026-02-20-117710c0	6.10
A4 authentic superego, Nemotron (cells 78–79)	eval-2026-02-19-dbcd6543	6.16.1

Finding	Run ID	Section
A4 authentic superego, Haiku (cells 78–79)	eval-2026-02-20-058c7a0e	6.16.1
A2 mechanism sweep, Haiku (cells 72–77)	eval-2026-02-20-57ba525c	6.10
Self-reflect Haiku supplement	eval-2026-02-20-90703a6a	6.10
A2 mechanism sweep clean (cells 72–77)	eval-2026-02-23-b5cd16e1	6.10
A4 authentic superego clean (cells 78–79)	eval-2026-02-23-b5e123b4	6.16.1
Autotuned base prompt, Qwen 3.5 (cell 80)	eval-2026-03-05-53fb1462	7.7
Autotuned base prompt, DeepSeek (cell 80)	eval-2026-03-06-08e5eeab	7.7
Recognition baseline, DeepSeek (cell 84)	eval-2026-03-06-8b9fbaba	7.7
Autotuned recognition, DeepSeek (cell 84)	eval-2026-03-06-88dc49de	7.7
Autotuned recognition pass 2, DeepSeek (cell 84)	eval-2026-03-07-68acef5a	7.7
Recognition baseline, Qwen 3.5 (cell 84)	eval-2026-03-07-94aef993	7.7
Autotuned recognition, Qwen 3.5 (cell 84)	eval-2026-03-07-c4d1bfa2	7.7
Autotuned recognition rep 5, Qwen 3.5 (cell 84)	eval-2026-03-07-baf2367b	7.7
Trajectory scenarios, DeepSeek (cells 80–87)	eval-2026-03-06-ebcd6de0	6.3
M2 isolation: base + superego, DeepSeek (cells 82–83)	eval-2026-03-06-768ba77b	6.4.2
M1 isolation: recognition, no superego, DeepSeek (cells 84–85)	eval-2026-03-06-e4abd0df	6.4.2
A12 M3 disengagement replication, Haiku 4.5 (cells 80, 84)	eval-2026-04-22-d4547979	6.3.2
A12 M3 disengagement replication, Gemini Flash 3.0 (cells 80, 84)	eval-2026-04-22-f4fb03f1	6.3.2
A11 M2-alone isolation on Gemini Flash 3.0 (cells 82, 83)	eval-2026-04-22-b56be6c7	6.4.2

Finding	Run ID	Section
A10 v1 matched-specificity prompt-density control (cells 1, 5, 95) — **INVALIDATED BY BUG 007** (cell_95 silently ran the base prompt; see Supplement V v3.0.47)	eval-2026-04-22-04497df0	7.9

Supplement V: Revision History

This appendix is a flat per-version log of all paper revisions. Two judge panels appear across the history and should not be conflated with the abstract's headline claim:

- **Entries v1.0–v2.3.21** (2026-02-04 – 2026-02-25): pilot-study revisions (see companion Paper 1.0). Primary judge was Opus 4.5, later unified to Opus 4.6 (see v2.3.4 "Judge version unification"). Pilot data is *not* part of the Paper 2.0 headline claim.
- **Entries v3.0.0 onward** (2026-03-07 – present): Paper 2.0 mechanism-run revisions. The headline validation panel reported in the abstract is three independent judges: **Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4** (1,296 scored rows, zero nulls, across three generation models). Opus 4.6 is retained as a judge for *specific side-analyses* — the §8.1 learner-paradox DB check (Opus 4.6, N=613/108/499/157) and the §6.4.2 isolation runs — and those uses are labelled in-place. Every headline d in the Paper 2.0 abstract refers to the triple-judge panel, never to Opus.

v1.0 (2026-02-04) Initial draft with 2x2x2 factorial design, memory isolation, three-way comparison.

v1.1 (2026-02-06) Added corrected memory isolation experiment (N=120), active control (N=118), cells 6&8 re-run, cross-judge GPT-5.2 analysis. Corrected GPT-5.2 effect sizes (d=1.15 to 0.99, d=0.50 to 0.29) after deduplication of rejudge rows. Dropped dead partial run (e617e757).

v1.2 (2026-02-06) Critical correction: Reframed "placebo control" as "post-hoc active control." The original v1.1 analysis compared the active control (Nemotron, M=66.5) to factorial base (Kimi K2.5, M=78.8) and reported d=-1.03, but this compared different ego models. Same-model historical data shows Nemotron base ≈ 58, making the active control ≈ +9 pts above base (not below). Reframed throughout: generic pedagogical elaboration provides partial benefit (~+9 pts above base) but recognition gains are substantially larger (~+15 pts). Acknowledged post-hoc design and active (not inert) control content.

v1.3–v1.4 (2026-02-06) Intermediate revisions: corrected factorial with re-run cells 6, 8 (a933d745); updated AxC interaction values; qualitative

analysis additions; production quality fixes. Superseded by v1.5.

v1.5 (2026-02-07) Rubric iteration: Updated to 14-dimension rubric with dialogue transcript context, Productive Struggle (5%), and Epistemic Honesty (5%) dimensions (Actionability/Tone reduced 10% to 8%). Re-scored cells 6, 8 (N=88) with identical responses: minimal change (+0.5, +0.6 pts), confirming calibration preserved. Added holistic dialogue evaluation for multi-turn transcripts. Cross-judge replication on updated rubric (r=0.55, N=88, GPT/Opus ratio=0.87). Updated Table 6, main effects, AxC interaction values, Supplement III.2 weight table, and Section 6.18 cross-judge tables.

v1.6 (2026-02-08) Content isolation fix: Identified and fixed two bugs causing cross-domain content leakage in elementary scenarios: (a) `buildCurriculumContext()` fallback that scanned all courses when no content hint was provided, serving philosophy listings to elementary scenarios; (b) hardcoded `479-lecture-*` IDs in tutor ego prompt examples that the model copied when no curriculum anchor was present. Updated Sections 6.5, 6.6, 7.4, 7.8, and 8 to reframe "model hallucination" as system-level content isolation failures.

v1.7 (2026-02-08) Hardwired rules ablation: Added Section 6.7 with superego rules embedded in ego prompt (cells 13–14, N=72, eval-2026-02-08-65a6718f, Opus judge). Static rules fail to replicate the Superego's benefit, confirming the value lies in contextual judgment rather than rule enforcement. Added Table 10b, updated Tables 2/D and paper totals.

v1.8 (2026-02-08) Dialectical impasse test: Added Section 6.20 with three 5-turn impasse scenarios (epistemic resistance, affective shutdown, productive deadlock; N=24, eval-2026-02-08-f896275d, Opus judge). Recognition produces +43 pts on epistemic and +29 pts on interpretive impasses but $\Delta = -1.1$ on affective shutdown—sharpening the theoretical claim to epistemological rather than affective recognition.

v1.9 (2026-02-08) Learner superego paradox: Added symmetric learner-side evaluation (Section 6.16) scoring N=118 bilateral dialogues with 6-dimension learner rubric (eval-2026-02-07-b6d75e87, Opus judge). Multi-agent learner architecture hurts learner quality (d=1.43, F=68.28, p<.001)—the largest effect in the study. Recognition partially rescues multi-agent learners (d=0.79, p=.004) but not single-agent (n.s.). Added learner rubric description to Section 5.1, new Section 6.12, rewrote Section 7.5 with results, added finding #9 to Section 9.

v2.0 (2026-02-08) Resolution strategy coding: Post-hoc qualitative coding of all 24 dialectical impasse dialogues into five Hegelian resolution strategies. Perfect separation: 12/12 base tutors withdraw, 10/12 recognition tutors use scaffolded reframing (Aufhebung pattern). $\chi^2(3) = 24.00$, $p < .001$, $V = 1.000$. Cross-judge validation with GPT-5.2: $\kappa = 0.84$. Added Tables 26–28, per-turn strategy evolution analysis.

v2.1 (2026-02-08) AI theme discovery and figure regeneration: Added Section 6.13.4 AI-assisted theme discovery (N=300) showing near-perfect bimodal separation. Added Figure 6 (word clouds). Regenerated all fig-

ures from Python with corrected data and larger text. Removed standalone Section 10 Reproducibility (merged into Supplement IV). Moved Supplement V after other appendices. Increased font to 12pt.

v2.1.1 (2026-02-10) Consistency fixes: Corrected stale N=1,628/twenty to N=1,700/twenty-one in abstract, introduction, and conclusion. Fixed dynamic rewrite section references in Tables 2 and D. Added hardwired rules ablation and learner-side evaluation to Supplement IV run list (was 19 rows, now 21). Fixed inter-judge reliability cross-reference in Section 8.1.

v2.1.2 (2026-02-10) Review corrections (30 fixes): Table 7b Kimi row corrected to single-learner cells (N=350 to 179, Recognition +10.2 to +15.5, Interaction -1.5 to +0.5) matching probe design; total probe N 826 to 655. Factor C in Discussion corrected (-1.7 pts, F=2.56). Stale AxC values updated. Dynamic rewrite swing corrected (+16.7 to +8.7 delta). Terminology standardized (unified to single-agent, behaviour to behavior).

v2.2.0 (2026-02-11) Modulation and learning outcomes: Added Section 6.11.1 (modulation metrics, N=350 post-hoc) showing multi-agent architecture does not increase behavioral range (d=0.05); recognition produces calibration not oscillation (dimension variance d=−1.00, F=87.69). Added Section 6.11.2 (synthetic learning outcome index, N=118). Extended Section 7.4 Discussion with phronesis reframing. Regenerated Figures 4 and 6.

v2.3.0 (2026-02-14) Phase 2 experimental results: Added four new Results sections: Section 6.8 Dialectical Superego Modulation (cells 22–33, N=174, Tables 13–15); Section 6.9 Self-Reflective Evolution (cells 40–45, N=36, Tables 16–17); Section 6.10 Mechanism Robustness (cells 40–59 N=360 + cells 60–63 N=120, Tables 18–19); Section 6.11 Qualitative Transcript Assessment (Tables 20–21). Added Section 7.10 Scripted Learner Confound, Section 7.11 Practical Recommendations (6 recommendations). Expanded Section 6.10 with cells 64–65 and Nemotron cross-model replication (N=279, 49b33fdd). Renumbered all tables sequentially (1–48). Trimmed abstract from ~650 to ~250 words. Paper totals: N=2,700 across 28 key evaluations.

v2.3.1 (2026-02-15) Cognitive prosthesis and cross-judge completion: Added cognitive prosthesis test (cells 66–68, N=60). Completed GPT-5.2 cross-judge validation of mechanism robustness (N=360 paired, r=0.59). Added Nemotron self-reflect non-replication (559d854b, N=60) to Section 6.9/Table 17. Added blinded qualitative assessment validation (Table 21b).

v2.3.2 (2026-02-15) Sample reconciliation and count update: Added Phase 2 evaluations to Table 2 (9 additional rows). Updated paper totals from 28 to 30 key evaluations, N=2,909 scored. Added 50487df7 (cognitive prosthesis) to Supplement II.6. Noted Sonnet judge used for two late-stage evaluations.

v2.3.3 (2026-02-15) Complete Table 19 with base mechanism cells: Added cells 69–70 (eval-2026-02-15-664073ab, N=60, Opus judge) completing the base row of Table 19. Recognition delta remarkably consistent

across all 4 mechanisms (+13.3 to +15.1). Updated paper totals from 30 to 31 evaluations, N=2,969.

v2.3.4 (2026-02-15) Related Work expansion for arXiv/edArXiv submission: Expanded Section 2 from 8 to 10 subsections. Added Section 2.3 LLM-as-Judge Evaluation Methodology, Section 2.7 Theory of Mind in AI Agents. Expanded Section 2.1 with empirical LLM tutoring studies, Section 2.2 with multi-agent systems and self-correction limits. Added 15 new bib entries.

v2.3.5 (2026-02-15) Same-model blinded assessment: Ran Opus-blinded qualitative assessment (N=118) resolving the model calibration confound. Key finding: blinding barely changes Opus's tag assignments, confirming the near-perfect binary separation is real, not an assessor bias artifact. Updated Section 6.11 interpretation, revised Section 8.2 limitation.

v2.3.6 (2026-02-16) Judge version unification: Rejudged all early runs (originally scored under Opus 4.5) with Opus 4.6, eliminating version drift across the evaluation dataset. Updated Section 8.1 limitations. Cleaned 6 empty/failed generation rows from dynamic rewrite runs.

v2.3.7 (2026-02-17) Self-reflective evolution complete: Updated Section 6.9 from partial (N=36) to complete (N=90) results for eval-2026-02-13-8d40e086. Recognition d=0.91 (was 1.02 at N=36). Key new finding: disposition gradient—suspicious +19.0, adversary +10.9, advocate +2.6. Updated Table 16, Table 17, Discussion finding 11. Deduped 270 re-judging artifact rows.

v2.3.8 (2026-02-17) Nemotron mechanism N-count update: Updated eval-2026-02-14-49b33fdd from N=301 to N=360 after run resumption completed. Cascaded count changes through abstract, introduction, Table 2, Section 7.11, Section 9. Updated Section 6.10 Nemotron narrative. Noted bidirectional profiling anomaly ($\Delta = -0.6$).

v2.3.9 (2026-02-17) Factorial re-judging cascade: All early runs rejudged with Opus 4.6 (unifying judge version). Full factorial ANOVA (N=350): recognition F=110.04 p<.001 η^2=.243 d=1.11 (was F=71.36, d=0.80). AxC interaction disappears (F=0.97, p=.325; was F=21.85, p<.001)—recognition now consistent across learner types. Updated Tables 4, 6, 8, 9, 9b, 12, 17, 41, 42. Restored 219 GPT-5.2 rows lost during dedup. Updated GPT compression ratio from ~58% to 37–59%.

v2.3.10 (2026-02-17) Prompt elaboration baseline: Added Section 6.21 comparing 344-line base prompt against 35-line naive prompt (JSON schema only). Two runs: Haiku (N=72) and Kimi (N=72). Key finding: elaborate prompt hurts Haiku (+6.8 pts for naive) and is inert on Kimi ($\Delta = -0.3$). Recognition ($M = 90.9$) remains well above naive ($M = 82.5$). Added Table 20b, conclusion finding 13. Updated paper totals to N=3,292 across thirty-one evaluations.

v2.3.11 (2026-02-17) Transcript figures: Added Figure 10 (naive vs base high_performer comparison panel) to Section 6.21. Added Figure 11 (bilateral mutual_transformation_journey transcript comparison) to Section 6.11. Added `generate-paper-figures.js` script for reproducible paper fig-

ure generation.

v2.3.12 (2026-02-17) Token budget sensitivity: Added Section 6.22 testing whether constraining `max_tokens` from 8000 to 256–2048 affects evaluation scores. Scores are flat across all budget levels; the recognition effect is fully preserved even at 256 tokens. The retry-absorption mechanism means truncated structured output self-heals. Added Table 49, recommendation 8 to Section 7.11. Updated paper totals to N=3,454 scored across thirty-six evaluations.

v2.3.13 (2026-02-17) Paper correctness fixes: Fixed eval-2026-02-14-559d854b scope from "cells 40–59, N=167" to "cells 40–45, N=60" in Table 2 (only cells 40–45 used; cells 46–59 superseded by 49b33fdd at N=360). Fixed broken Table 10 reference in Section 6.6. Fixed dynamic-learner N inconsistency: intro and finding 12 updated to N=300 (6c033830 + a2b2717c + 664073ab). Clarified token budget Section 6.22 design text. Added missing Supplement II commands.

v2.3.14 (2026-02-18) Cognitive prosthesis re-run and analysis: Replaced misconfigured prosthesis run (50487df7, all cells fell back to default Haiku) with corrected eval-2026-02-17-25aaae85 (N=90, Nemotron ego, Kimi K2.5 superego, Opus judge). Prosthesis hypothesis fails decisively: full mechanism stack scores 49.5 vs Nemotron simple base 64.2 ($\Delta = -15$). Added dimension analysis (two-tier static/dynamic capability model), superego parse failure analysis (16–45% malformed JSON auto-approves), and Haiku control smoke test (eval-2026-02-18-f489c0ea, N=6, confirming model-dependence). Added conclusion finding 14 (minimum ego capability threshold), recommendation 9 to Section 7.11, three future work items to Section 8.2 (parse robustness, capability threshold mapping, adaptive mechanism loading). Updated Table 2, Supplement IV run IDs. Paper totals: N=3,383 across thirty-seven evaluations.

v2.3.15 (2026-02-19) Active control model confound resolved (A8): Ran active control (cells 15–18) on Kimi K2.5 (eval-2026-02-19-f2263b04, N=216, 116 grounded). Key finding: active control scores *below* base on Kimi ($M = 56.0$ vs $M = 64.2$ on matched scenarios, $\Delta = -8.2$), not between base and recognition as on Nemotron. Prescriptive heuristics help weaker models (Nemotron: +9 pts) but harm stronger ones (Kimi: −8 pts)—consistent with Section 6.21 prompt elaboration baseline. Same-day reproduction runs confirmed no model drift: Kimi base $M = 58.8$ (eval-2026-02-19-13d34bef, N=36) and Nemotron base $M = 49.8$ (eval-2026-02-19-411414e4, N=18). GPT-5.2 cross-judge validation confirms ordering (placebo 48.3 < base 70.8 < recognition 76.5). Updated Section 6.2 (added Table 6b, model-dependent interpretation), Section 5.3 (active control description), Section 8.1 (limitation resolved). Paper totals: N=3,653 across forty evaluations.

v2.3.16 (2026-02-21) Recognition inversion mechanism and A2/A4 experiments: Added Section 6.16.1 documenting the scenario-specific recognition inversion (misconception_correction: tutor +30.5, learner -10.4; mutual_transformation: no inversion). Identified metacognitive

feedback loop mechanism through qualitative transcript analysis. A2 dynamic learner mechanism sweep (cells 72–77, N=238): all 7 mechanisms show positive recognition deltas; dynamic learner amplifies mechanism differentiation 1.6–2.8x vs scripted. A4 authenticity-focused learner superego (cells 78–79, N=47): null result—scores worse on every dimension including authenticity itself. Epistemic readiness dimension feasibility test (N=10): no base-recog gap (both 4.0/5), ruling out construct validity artifact. Reframed Section 7.5 from "calibration problem" to "structural finding." Paper totals: N=4,144 across forty-eight evaluations.

v2.3.17 (2026-02-21) Table 19a and conclusion finding #15: Expanded Section 6.10 mechanism results with Table 19a showing A2 sweep data (cells 72–77, N=108, Nemotron ego) for quantitative, erosion, and tutor-profiling mechanisms. All three show positive recognition deltas (+8.0 to +9.7), completing the 2x7 mechanism matrix. Updated conclusion finding #12 to reference full 7-mechanism matrix. Added finding #15 documenting recognition inversion as scenario-specific and structural. Fixed stale "thirty-seven" to "forty-eight" in conclusion summary.

v2.3.18 (2026-02-23) Bug disclosure and data integrity corrections: Added Section 8.1 paragraph disclosing two learner pipeline bugs affecting all pre-Feb-20 dynamic learner multi-turn evaluations: Bug 1 (leaky "SUPEREGO:" labels in learner external messages) and Bug 2 (broken `.flatMap()` causing learner to see only own monologue, not tutor responses). Added pre-fix data caveat to Section 6.16 (learner superego paradox, $d = 1.43$ from eval-2026-02-07-b6d75e87 is an upper bound). Added footnote to Section 6.11 flagging $M = 38.0$ as pre-fix. Clarified Section 6.10 Table 19 footnote to name bugs explicitly and explain why tutor-side scores are valid but learner-side required re-generation. Qualified Finding #2 (additivity) as tutor-side only, noting learner-side AxC interaction (Section 6.16). Qualified Finding #9 conclusion text with pre-fix caveat. Added forward reference from Section 6.3 to Section 6.16 learner-side interaction. Added clean re-generation runs eval-2026-02-20-bd37cc62 (cells 64–65) and eval-2026-02-20-4e131c6f (cells 69–70) to Table 2 and Supplement IV. Updated paper totals: N=4,264 across fifty evaluations. Also fixed Bug 5 in evaluation infrastructure: `resumeEvaluation()` silently dropped `--ego-model` and `--superego-model` overrides from stored metadata, causing resumed runs to fall back to YAML defaults.

v2.3.19 (2026-02-24) Post-fix data replacement: Replaced all pre-fix dynamic learner data with clean re-run data. Section 6.10 Table 19 now reports clean re-run tutor scores (0fbca69e, bd37cc62, 4e131c6f, N=240); mechanism ordering changes—intersubjective framing becomes the dominant mechanism (previously weakest), recognition deltas larger (+19.1 avg vs +14.2 pre-fix). Table 19a replaced with b5cd16e1 clean data (N=36). Section 6.16 Tables 29–31 replaced with post-fix factorial data (25c78e91, N=48): architecture effect *increases* to $d = 3.05$ (was 1.43), recognition effect vanishes ($d = 0.09$), AxC interaction vanishes ($F = 0.12$, was

11.50). The "recognition rescues multi-agent learners" finding was a bug artifact—removed throughout. Section 6.16.1 A4 results updated with b5e123b4 clean data (N=12). Rewrote Finding #2 (additivity holds on both rubrics), Finding #9 (paradox real and stronger, recognition irrelevant), Finding #12 (intersubjective framing now dominant mechanism), Finding #15 (tutor-learner asymmetry, not inversion). Updated Section 7.5, Section 8.1, abstract, and concluding summary. Section 6.11 footnote refined. Section 6.10 prose rewritten for new mechanism ordering. N-counts unchanged at 4,264 across fifty evaluations.

v2.3.20 (2026-02-25) Bug 4 disclosure and claim-by-claim audit response: Added Section 8.1 paragraph disclosing multi-turn scoring misalignment bug (Bug 4): `evaluate`/`rejudge` scored `suggestions[0]` (Turn 0) instead of last turn, affecting 8,631 multi-turn rows across 81 runs. Average holistic lift +19.9 pts. Paper clarified that all multi-turn tables report Turn 0 scoring; holistic last-turn rescoring in progress. Added prefix data caveats to Section 6.15 Table 25 footnote (b6d75e87 affected by Bugs 1–2 in ego/superego cells and by Bug 4 Turn 0 scoring). Extended Section 6.15.2 Table 28 caveat noting cross-run audit (N=895) shows negative learner-side learning arcs under recognition in multi-turn settings. Strengthened Section 8.1 bilateral transformation asymmetry paragraph with cross-run evidence ($d = -0.27$ recognition, $d = -0.98$ architecture on arc metrics, 30/33 runs show recognition reduces tutor dimension variance). Updated synthetic learning outcome bullet in summary to reflect mixed learner-side evidence. Added eval-2026-02-23-b5cd16e1 (cells 72–77, N=36) and eval-2026-02-23-b5e123b4 (cells 78–79, N=12) to manifest and Supplement IV. Updated paper totals: N=4,312 across fifty-two evaluations (manifest v1.8.0).

v2.3.21 (2026-02-25) N-claim and stat-claim traceability: Systematic pass through all 156 N-claims and 66 stat-claims identified by the claim-by-claim audit tool, adding inline run ID anchors, Section cross-references, and Table citations. Key changes: added run IDs to all table headers lacking them (Tables 3, 5, 6, 8, 13, 15, 20b, 21, 22, 25, 45); added Run ID column to Table 41 (inter-judge agreement); annotated pooled-data N-claims with "see Table 2" or "pooled across evaluation database"; marked pilot data as "no longer in primary evaluation set"; added Section/Table cross-references to all stat-claims in Sections 7–9. No data, analysis, or N-count changes—purely traceability metadata.

v3.0.0 (2026-03-07) Automated prompt tuning experiment: Added Section 7.7 with $2 \times 2 \times 2$ autotuning experiment (base/recognition × unoptimised/autotuned × DeepSeek/Qwen 3.5). Seven new runs (eval-2026-03-05-53fb1462 through eval-2026-03-07-c4d1bfa2, N=38 total). Key finding: recognition prompts are more improvable than base (+17–23 pts vs +5–9 pts), the effects are super-additive, and the gap between conditions *widens* under optimisation. Autotuned recognition converges to ~71–73 on both models. Base autotuning ceiling (41–45) remains below unoptimised recognition baseline (48–55), confirming recognition's irreducibility to prompt

engineering. Added Table N, 7 run IDs to Supplement IV. Renumbered Section 7.7 → 7.8.

v3.0.1 (2026-03-07) M3 trajectory analysis: Updated M3 (adaptive responsiveness) from "disconfirmed" to "conditionally supported" based on complete trajectory-specific data (eval-2026-03-06-ebcd6de0, N=72, 3 scenarios × 8 cells × 3 replications, DeepSeek V3.2, Sonnet judge). Pooled slope effect remains small (d=0.22, NS), but the 10-turn disengagement scenario shows d=1.63, p<.001, with recognition–baseline gap widening from +12 pts at T0 to +35 pts at T8–T10. The two 8-turn scenarios show no slope differentiation ($d \leq 0.30$). M3 recharacterized as a conditional emergent property of M1+M2 that manifests under sustained re-engagement demands with sufficient turn count. Updated abstract, Section 3.2, Section 6.3.2, Section 6.3.8, mechanism summary table, and Discussion synthesis. Added 1 run ID to Supplement IV (53 total).

v3.0.2 (2026-03-07) M1/M2 mechanism isolation: Added dedicated isolation run confirmation to Section 6.4.2. Runs eval-2026-03-06-768ba77b (M2: base + superego, cells 82–83) and eval-2026-03-06-e4abd0df (M1: recognition, no superego, cells 84–85) across 9 multi-turn scenarios (N=108, DeepSeek V3.2, Sonnet judge). Full 2 × 2 isolation confirms substitution: superego adds +9.2 pts under base (d=1.13, p=.002) but +1.1 under recognition (d=0.08, NS)—calibration pre-empts 88% of the superego's contribution (27% additivity deficit). M1 vs M2 head-to-head: calibration alone (51.4) outscores error correction alone (36.9) by d=1.03 in 7/9 scenarios. Two emotionally intense scenarios (Frustration, Affective Shutdown) show slight M2 advantage, suggesting scenario-specific residual error correction value. Added 2 run IDs to Supplement IV (55 total).

April 2026 arc summary (v3.0.43 → v3.0.51, nine revisions in four days) The published version prior to this cycle was **v3.0.42** (2026-04-21). What followed was a compressed three-day cycle triggered by an internal critical review of v3.0.42 (2026-04-22), which identified five issues: disproportionate narrative weight on a pending boundary-condition M3 finding, tension between §6.1 dimension-narrowing language and §8.6 PCA's one-factor result, Sonnet-only headline d's despite §8.3 reporting three-judge variation, untrimmed apparatus-as-method exposition, and an untested strong version of the prompt-density alternative. The cycle had three phases:

Phase 1 — editorial tightening (v3.0.43–v3.0.44, v3.0.51). Collapsed M3-disengagement exposure from 13 paper locations to one canonical §6.3.2 treatment; rewrote §6.1 calibration lead for PCA consistency; reframed all headline d's to judge-pooled estimates; trimmed §7.4 apparatus-as-method by ~35%; trimmed universal-substitution restatements from 22 to 15 with single canonical homes (v3.0.51, F4). No data changes.

Phase 2 — four new experiments closing open concerns (v3.0.45–

v3.0.48). A12 (M3 disengagement replication across Haiku + Gemini Flash under Sonnet + GPT-5.4, $N = 32$): the pre-registered replication failed — three of four model × judge cells below $d = 0.5$, and the one replicating cell contradicted by its secondary judge at cross-judge $\Delta d = 2.03$. M3 retired as clean null (v3.0.45). A11 (direct M2-alone isolation on Gemini Flash, $N = 125$): pooled $d = 1.46$ on Sonnet, confirming and exceeding the factorial-inferred +12.3 residual in §6.4.1 (v3.0.46). A10 v2 (matched-specificity prompt-density control, three-judge triangulation, $N = 183$): pooled three-judge $d = 0.17$ recognition-vs-matched-pedagogical, below the pre-registered $|d| < 0.2$ density-sufficient threshold (v3.0.48). A10b (four-way orientation-family comparison adding a rigorously-authored behaviorist prompt, three-judge at $N = 248$ generation with full GPT coverage and ∼ 80% Sonnet/Opus coverage): pooled $d = 1.38$ between Hegelian-descendant family and transmission family; within Hegelian family, density-substitutable ($d = 0.14$); within transmission family, behaviorist substantially *below* base ($d = 0.89$) (v3.0.48).

Phase 3 — bug discovery, retraction, and framing reframe (v3.0.47, v3.0.49, v3.0.50). A10 v1 (the first run of the density control) was invalidated by an evaluation-pipeline dispatch bug (bug_007: `resolveEvalProfile` lacked a branch for the new `prompt_type`, routing cell_95 silently to the base prompt). Retracted in v3.0.47 before A10 v2 re-ran on the fixed pipeline. v3.0.49 implemented a calibrated Framing A reframe of intro/abstract/§9 acknowledging the family-level finding while keeping §3's recognition-specific theoretical motivation. v3.0.50 then refined this to the "primacy + descent" framing (Framing B+): empirical generalisability is real (matched-pedagogical reproduces recognition within $|d| < 0.2$), but recognition theory remains the most explicit philosophical articulation of an intersubjective-pedagogy orientation its constructivist descendants (Dewey, Vygotsky, Piaget, with Kapur, Chi, VanLehn, and Graesser as further cognate descendants) inherit, often without direct attribution. §1 added a careful three-circle genealogical account; §3 added a level-of-analysis framing note; §9 extended the family-level passage with explicit "generalisability does not displace recognition theory" reasoning.

Net effect across the cycle. Three mechanisms originally claimed; now two supported, one clean null. The two supported mechanisms have stronger evidence than before (A11 confirming architecture residual; §7.9 closing the density alternative at the orientation-family level). The paper has a new methodological contribution (pedagogical-orientation taxonomy, `docs/pedagogical-taxonomy.md`) and a retracted-then-corrected experiment (A10 v1 → A10 v2). An `/ultrareview` pass caught the bug_007 issue before A10's result propagated into a broken paper claim. Net paper length grew by ∼6% (new §7.9 orientation-family content); net theoretical-content distinctiveness shrank slightly (recognition now framed as one family member rather than the uniquely-necessary frame), but defensibility strengthened.

Individual v3.0.X entries follow in chronological order (newest first):

v3.0.64 (2026-04-29) §6.7 closing-synthesis paragraph — how

charisma and recognition relate. Adds a sixth paragraph between the v3.0.63 *Lever-interaction follow-up* and the *Cross-judge sanity check* that synthesises the five preceding §6.7 findings into a unified account of the two rubrics' relationship. Three claims, each supported by previously-reported numbers: (a) **shared floor** — c106 fails both rubrics (charisma 36.4, v2.2 57.0), so register-flat output is the worst case for either; (b) **cost asymmetry** — adding recognition vocabulary on top of structured input (c103 → c104) costs charisma ≈0 (+1.4, within noise) because the Hegelian frame is conceptual rather than rhetorical, while adding charisma directives costs v2.2 substantially (c104 → c105: −10.6 v2.2 for +5.3 charisma) because the id's authoring budget is finite and voice-instructions displace elicitation/perception moves; (c) **architectural Pareto frontier** — single-mechanism design points (c104, c105, c107) define the frontier, multi-mechanism stacks (c108, c109) sit inside it. Closes with the interpretive frame that recognition wants the voice deployed *outward* (perceiving the learner, eliciting thinking, holding the diagnostic gap) while charisma wants the voice deployed *upward* (claiming authority, marking encounter as extraordinary, inviting transformation), and that the apparatus can build either or a balanced generalist but not the maximally-both. No new analyses, no new run ids, no changes to numbers — purely synthesis from existing reported results. Net change ~+260 words in §6.7.

v3.0.63 (2026-04-29) §6.7 c108 full-N follow-up — pilot lift does NOT survive confirmatory N. Replaces the v3.0.62 *Lever-interaction pilot* paragraph with a confirmatory-N rewrite reporting the c108 follow-up at $n = 27$ across all three curricula × 9 scenarios × 3 reps (run ids `eval-2026-04-29-{499ab3d2,e35c66ae,be822796}`, scoring under `claude-code.sonnet`). The pilot's super-additive charisma 81.3 regresses to **71.4** (a 9.9-point drop, tying c105's 71.0 within sample noise); v2.2 last-turn 79.5 regresses to **72.6** (a 6.9-point drop, sitting below c107's 78.5 and c104's 80.6). Per-scenario inspection identifies the culprit: pilot phaedrus scored 90.0 charisma at $n = 3$ but full-N phaedrus scores 67.9 at $n = 3$ — a 22-point regression to the mean. The 701 ethics-ai curriculum exposes a new fragility (c108 v2.2 tN 65.3 there, vs 76.7 / 75.8 on 601 / 901). The asymmetry that survives the confirmatory threshold is narrower than the pilot suggested: text-heavy levers stacked (c109) clearly interfere; non-text levers stacked (c108) compose roughly additively to slightly under-additively, trading +5.1 charisma for -8.0 v2.2 relative to c107 alone. The architectural ceiling for the id-director family sits at the single-mechanism design points (c104, c105, c107), not at multi-mechanism stacks. Full audit trail in `docs/cell-100-charisma-full-n-update.md` §10. Adds the three new run ids to the §6.7 reproducibility list. No changes to cells 101–107 numbers or to the four prior headline findings; net change ~+340 words in §6.7.

v3.0.62 (2026-04-28) §6.7 lever-interaction pilot integration (cells

108/109, hypothesis-generating). Adds a fifth headline-finding paragraph to the §6.7 id-director family section reporting a small two-cell follow-up pilot (`eval-2026-04-28-10216f9f`, $n = 11$ after one generation failure) that tests whether the three architectural levers compose additively. **c108** (register classifier + witness exemplars) lifts pooled charisma to 81.3 — the highest mean of any id-director cell in the section — with v2.2 last-turn at 79.5, between c104's 80.6 and c107's 78.5; the classifier and exemplars appear to occupy different slots in the id's authoring task. **c109** (charisma directives + witness exemplars) fails the additive hypothesis: charisma is preserved at 77.7 but v2.2 last-turn collapses to 59.6 due to an ego-side meta-narration failure mode where the Nemotron ego outputs system-prompt instructions verbatim rather than executing them. The qualitative asymmetry — non-text levers compose, text-heavy levers stacked interfere via instruction-following degradation — is reported as hypothesis-generating only; pilot N is far below confirmatory floor. Adds run id `eval-2026-04-28-10216f9f` to the §6.7 reproducibility list. Full audit trail in `docs/cell-100-charisma-full-n-update.md` §9. No changes to cells 101–107 numbers or to the four prior headline findings; net change ~+260 words in §6.7.

v3.0.61 (2026-04-28) §8.9 + §6.7 N-scope tightening for full-N c107 closure. The "smaller per-cell N (typically 27 on the original set)" line — written when the c107 witness-exemplars cell was at n=10 and most other id-director cells were partially scored — is replaced by the actual closing N range (n=79–81 for c101/c103/c104/c105, n=54 for c102/c106, n=27 for c107). Three of seven cells meet or exceed the paper's core $n \geq 63$ per dialogue-condition floor; one cell (c107) sits below it and is now flagged as pilot-scale rather than confirmatory. Mirrors the §6.7 *Replication, robustness, and orientation* paragraph against the §8.9 *Per-cell N* axis. No data, analysis, or rank-ordering changes; purely scope-language calibration to match the closed-pilot N matrix in `docs/cell-100-charisma-full-n-update.md` §1.

v3.0.60 (2026-04-29) §6.3.9 four-bridge synthesis closure: H-capacity reframed from "structural" to "lightweight-intervention-resistant; partially mitigable only by expensive search". With Bridge 3 (cell 100, best-of-N selector at $K = 3$, $\sim 6\times$ ego cost per turn) closing at $N = 53$ dialogues / 313 reflection-action pairs (run `eval-2026-04-27-327d8816`), the four-bridge synthesis is locked. Per-dialogue mean coupling: cell 40 = 0.614, cell 100 = 0.635, $\Delta = +0.021$, Cohen's $d = +0.41$ — above the pre-registered $d \geq 0.3$ "suggestive" threshold but below the $d \geq 0.5$ "bridgeable" threshold. EoQ% identical at 16% / 16% (the orthogonal pragmatic-act channel is unaffected by best-of-N search).

Bridge 3 is the only intervention with detectable directional movement on coupling — it engages the semantic-coupling channel where Bridges 0–2 do

not. Matching the selection criterion to the verdict criterion (eliminating Goodhart's escape route) still produces only $d = 0.41$. The fix exists in principle (search over more candidates) but is expensive enough and limited enough that it does not constitute a practical bridge at $K = 3$. Log-linear scaling in K would predict $d \approx 0.5$ at $K = 5$ and $d \approx 0.7$ at $K = 10$; a future Bridge 3b sweep would test that prediction.

§6.3.9 prose changes: title revised from "Structural Across Three Lightweight Interventions" to "Structural Under Lightweight Intervention; Partially Mitigable Only by Expensive Search"; synthesis table extended from 3 rows to 4 rows; closing in-flight clause replaced with closure description; H-capacity discussion reframed from "structural at strongest level" to "lightweight-intervention-resistant; partially mitigable only by expensive search". Per-dialogue Cohen's d recomputation also corrected the prior table's $|d| \leq 0.07$ characterization for Bridges 1 and 2: properly computed on per-dialogue mean coupling distributions, $d_1 = -0.21$ and $d_2 = -0.22$ (still well below the suggestive threshold). The earlier hand-estimates were of the right sign and direction but ~4× too small.

Net length change: ~+340 words in §6.3.9 (new Bridge 3 row in synthesis table; reframed prose paragraphs; updated EoQ%-decoupling and H-capacity discussions). No upstream changes to §3 theory, §4 method, or other §6 mechanism findings. Cross-references: TODO §D3 four-bridge synthesis lock (2026-04-29); design note `notes/design-d3-architectural-bridges.md` four-bridge synthesis section; per-run report `exports/insight-action-gap-eval-2026-04-27-327d8816.md`. Eval-repo commit `956b9cb` (cell 100 + best-of-N runner orchestration).

v3.0.56 (2026-04-26) §8.6.2 application protocol for transferring the diagnostic. Adds a six-step procedure (~430 words) for applying the §8.6.1 three-criterion within-cell mediator diagnostic to LLM-tutor evaluation pipelines beyond the Paper 2.0 source apparatus. The protocol specifies: (1) family partition, (2) per-row feature computation, (3-5) the three criterion checks, (6) PCA pairing with §8.6 rubric over-specification check. References open-sourced scripts (`analyze-d1-structural-features.js`, `analyze-d1-cross-judge-replication.js`, `analyze-d1-multifeature-ols.js`) as the data-loading extension point for new datasets. Names three candidate independent rubrics for future application: MathTutorBench (Macina et al. 2025), ICAP-aligned studies (Chi and Wylie 2014), LLM-Rubric (Hashemi et al. 2024). The protocol makes the methods-paper generality argument procedurally rather than requiring new empirical work in this revision; future runs of the protocol against the named rubrics would convert the procedural argument to an empirical one. No data, analysis, or N-count changes; ~+430 words in §8.6.2.

v3.0.55 (2026-04-26) §8.6.1 named three-criterion within-cell mediator diagnostic. Adds a new subsection §8.6.1 (~520 words) immediately

after the §8.6 "Methodological pairing with §7.10" paragraph. The diagnostic was implicit across §7.10 + §7.10.1 + §8.6 ("only one feature satisfies all three within-cell mediator criteria — large family contrast, small within-intersubjective contrast, AND positive within-cell correlation"); v3.0.55 articulates it explicitly as a portable methodology with named criteria and a worked example.

The three criteria are stated formally: (1) large family contrast ($|d| \geq 0.5$); (2) small within-family contrast ($|d| < 0.5$); (3) positive within-cell Pearson r in each family-member cell. The worked example tabulates the D1 five-pass feature sequence with each feature's verdict (Hegelian/intersubjective lexicon = prompt markers; scaffolding moves = family marker, weak mediator; broad acknowledgement = negative correlate; ends-with-question = mediator in single-turn cells; intersub. canonical similarity = family marker, Simpson's paradox disqualifies as mediator).

The diagnostic generalises beyond the recognition–transmission contrast: any factorial design with a family-level main effect and candidate response-level mediators can apply the three-criterion check. Pairs with the §8.6 PCA over-specification finding as the two independent methods-level corrections that any future LLM-tutor evaluation should account for. §8.6.1 closes by noting that a separate methods-paper extraction is the natural vehicle for arguing the diagnostic's generality across published LLM-tutor evaluation pipelines (skeleton: `notes/methods-paper-skeleton.md`).

Net change: ~+520 words in §8.6.1; no data, analysis, or N-count changes. The diagnostic was applied throughout §7.10 already; v3.0.55 names it.

v3.0.54 (2026-04-26) §7.10.1 D1 mechanism analysis: replication, multivariate control, and a single-turn-vs-multi-turn scope condition. Adds a new subsection §7.10.1 (~600 words) consolidating four follow-up analyses on the §7.10 ends-with-question mediator finding and the intersub_advantage Simpson's-paradox finding.

- **Cross-judge replication** on the Opus 4.7 and GPT-5.2 judges of the A10b three-judge panel: ends-with-question r remains positive in 5 of 6 cell-judge combinations (the one exception, GPT × cell_95, is essentially zero rather than a sign-flip); intersub_advantage Simpson's-paradox replicates in 3 of 3 judges. Neither finding is Sonnet-specific. Report: `exports/d1-cross-judge-replication.md`. Script: `scripts/analyze-d1-cross-judge-replication.js`.
- **Multivariate OLS control** with six predictors (ends-with-question, second-person density, scaffolding moves, broad acknowledgement, question-mark rate, intersub_advantage). The ends-with-question partial coefficient remains positive and statistically significant in both intersubjective cells: cell_5 $\beta = 26.34$ ($p = 0.02$), cell_95 $\beta = 38.52$ ($p = 0.02$). The single-feature mediator interpretation is not collinearity-driven. The intersub_advantage Simpson's-paradox direction also survives

multivariate control. Pure JS Gauss–Jordan OLS implementation (df ≈ 43 per cell). Report: `exports/d1-multifeature-ols.md`. Script: `scripts/analyze-d1-multifeature-ols.js`.

- **Cross-run / cross-application replication on five additional runs** establishes that the mediator is **single-turn-specific**. In single-turn cells (A10 v2, A6 programming, D2 Path 1 peer support coaching) the mediator replicates uniformly with r range +0.16 to +0.74 across DeepSeek/Haiku ego, philosophy/programming/peer-support domains. In multi-turn cells (Paper 2.0 cells 84/86 messages-mode single-agent and multi-agent recognition; A2 cells 61/63 dialectical multi-agent with dynamic learner) the relationship **reverses sign** with r range −0.09 to −0.30. Substantive interpretation: in single-turn responses, ending with a question cedes initiative back to the learner; in multi-turn dialogues, ending the final turn with a question leaves the conversation unresolved, which the judges penalise. Same surface feature signals different pragmatic acts in different discourse contexts; the mechanism account (initiative-ceding) survives the refinement, but the surface signal becomes scope-conditioned. Report: `exports/d1-ends-question-replication.md`. Script: `scripts/analyze-d1-ends-question-replication.js`.

The cross-mode reversal does not falsify the underlying mechanism account; it specifies the discourse context in which the surface feature signals it. The §7.10 mediator claim now carries an explicit scope condition that maps onto the §6.3 single-turn–vs–trajectory distinction.

Net change: ~+600 words in §7.10.1; no data, analysis, or N-count changes to the primary Paper 2.0 mechanism findings; the v3.0.54 additions are interpretive integration of D1 follow-up analyses on already-collected runs.

v3.0.53 (2026-04-26) §6 D1 follow-up: form-vs-substance + negative-acknowledgement guardrail. Added §6.1.8 "Pragmatic Surface Markers Are Not Substitutes for Substance" (~280 words) immediately after §6.1.7 (question-asking mediation). The new subsection acts as an operational caveat on the §6.1.7 mediation result: it documents two patterns from the response-level D1 mechanism analysis that argue against a naive operationalist reading of "mandate intersubjective surface markers to reproduce recognition." Pattern 1: surface acknowledgement markers (quoted text, "your X" possessives, "what you mentioned"-style phrases) are *negatively* correlated with score within most cells ($r = -0.18$ cell_5, $r = -0.31$ cell_95, $r = -0.38$ cell_1) — the rubric penalises verbose surface empathy. Pattern 2: embedding-based similarity to an intersubjective canonical is positively correlated with score when pooled ($r = +0.26$) but reverses sign within each intersubjective cell ($r = -0.28$ cell_5, $r = -0.24$ cell_95) — Simpson's paradox driven by the cell_96 transmission outlier. Forward-references §7.10 for the full form-versus-substance account and the three-way decomposition. Net length change: ~+280 words; no data, analysis, or N-count changes; the underlying analysis already lives in §7.10

as of v3.0.52, this revision raises the operational caveat to results-section visibility for readers who otherwise might read §6.1.7 alone.

v3.0.52 (2026-04-26) D1 mechanism analysis integration. Integrates the five-pass D1 response-level mechanism analysis (`notes/d1-paper-integration.md`; commits `1ef5f60`, `e8dc7a8`, `3c9eaa6`, `c334722`) into Paper 2.0 across three sections.

- **New §7.10 "Mechanism location: response-level form versus substance"** ($ $750 words): corroborates the §7.9 orientation-family finding through an independent embedding-based measurement channel (intersub-canonical-similarity advantage $d = 0.81$ between families on the same A10b corpus, $n = 191$, OpenAI text-embedding-3-small); surfaces a Simpson's paradox at the response level (pooled $r = +0.26$ between embedding-similarity and score reverses to within-cell $r = -0.28$ in cell_5 and $r = -0.24$ in cell_95, driven by the cell_96 transmission outlier); identifies "ending the tutor turn with a question" as the single feature that satisfies all three within-cell mediator criteria across the five-pass sequence (cell_5 $r = +0.325$, cell_95 $r = +0.392$; cells 1 and 96 produce this in 0% of responses, cells 5 and 95 in 4.5% and 2.1%); decomposes "the recognition injection" into three distinct candidate active ingredients with separate verdicts — Hegelian content (not load-bearing), intersubjective stance (substantially load-bearing), specific surface features (observable channels of the stance, not substitutes); ties the decomposition back to the v3.0.50 "primacy + descent" framing at the response level. Includes a guardrail against the operationalist misreading "tell tutors to acknowledge more": surface-level acknowledgement markers (quoted text, "your X" possessives, "what you mentioned" patterns) are *negatively* correlated with score within most cells ($r = -0.18$ cell_5, $r = -0.31$ cell_95, $r = -0.38$ cell_1), counter to the naive operationalist reading.
- **§8.6 Simpson's-paradox methods note** ($ $150 *words appended*) : *pairs the §7.10 within-cell vs pooled r caution with the existing PCA over-specification finding to give two methods-level cautions for any future LLM-tutor mediator analysis. Together* : (1) *per-dimension claims should be hedged because the v2.2 rubric's eigh* in mediator analysis can be entirely driven by between-group variance rather than within-group mechanism. Both could form the empirical core of a methods-paper contribution beyond Paper 2.0.
- **§9 sixth future direction: white-box analysis targeting the within-cell mediator** ($ $130 words): reframes the mechanistic-understanding white-box hunt from generic residual-stream fishing to a small set of specific testable hypotheses about the model's pragmatic decision to end with a question — does the residual stream encode "should I cede initiative now" as an identifiable direction? Does attention shift to learner tokens before turn-ending question generation? Pinning the within-cell mediator to a specific decision point reframes the problem as applied interpretability with a defined target behaviour.

No data, analysis, or N-count changes; $ $1030 words added across three sections; net length change ∼+1.3%. Sources: `scripts/analyze-recognition-lexicon.js`, `scripts/analyze-d1-orientation-lexicon.js`, `scripts/analyze-d1-structural-features.js`, `scripts/analyze-d1-structural-features-v2.js`, `scripts/analyze-d1-structural-features-v3.js`; reports in `exports/d1-*.md`; integration synthesis in `notes/d1-paper-integration.md`.

v3.0.51 (2026-04-24, late evening) F4 redundancy trim — universal-substitution restatements compressed (closes TODO §F4). The "universal substitution with 15–17% additivity deficit" claim previously appeared in 22 places across the paper with full numerical restatement (deficit %, +12.3 residual, near-zero on strong models). F4 trim consolidated to 15 occurrences with single canonical numerical homes: §6.4.2 / §6.4.2.1 for the mechanistic account, §6.4.6 for cross-judge validation, §7.3 for the §7-level interpretation, §8.3 for cross-judge magnitudes, abstract for the headline. Other mentions compressed to bare "(§6.4)" or "(§6.4.2)" cross-references, dropping inline restatement of the deficit numbers and the +12.3 residual.

Touched paragraphs: §1 mechanism description (L80), §1 separability paragraph (L84), §1 contributions list (L100, L106, L108, L110, L114), §3 prediction (L704), §6.3 dialogue-quality cross-ref (L1386), §6.4.3 separability point 3 (L1504), §6.4.5 prediction headers (L1525, L1527, L1529), §7 lead (L1832), §7.1 effects-to-mechanisms recap (L1840), §7.2 messages-mode pattern (L1850), §7.3 internal duplication (L1858), §9 conclusion (L2133). The "universal substitution" *name* (paper-distinctive label, not hedge) is preserved wherever needed for legibility.

Pattern-2 hedging ("pending boundary-condition / requiring replication") was already retired in v3.0.43 (F1) + v3.0.45 (A12 closure); the surviving occurrences are deliberately framed as historical references in §6.3.2 explanatory text and Supplement V. Also incidentally compressed: §1 contribution bullet for adaptive responsiveness (full §6.3 + A12 restatement → verdict + cross-ref); §1 contribution bullet for tutor-learner asymmetry (now points to §6.5.5 rubric-vs-holistic correction rather than restating the superseded 7–12× headline).

Net change: ∼480 words trimmed across ∼18 paragraphs; line count unchanged (mid-paragraph compression rather than whole-line removal). No data, analysis, or N-count changes. Closes the F-series editorial cycle (F1-F6 all resolved); the open editorial item is now D5 (rubric v3.0 PCA consolidation), deferred to the next rubric version cycle.

v3.0.50 (2026-04-24, evening) Framing B+ reframe: primacy + descent. v3.0.49 implemented Framing A (calibrated — minimal upstream changes acknowledging the family-level finding). v3.0.50 implements the more careful "primacy + descent" framing the user articulated: recognition's empirical generalisability across cognate descendants is a strengthening of the paper, but it does not displace recognition theory's distinctive theoretical contribution, because **recognition theory is the most ex-**

plicit philosophical articulation of an intersubjective-pedagogy orientation that its constructivist descendants (Dewey, Vygotsky, Piaget, Kapur, Chi, VanLehn, Graesser) inherit, often without direct attribution.

The reframe is implemented across four sections:

- **Abstract**: rewrote the family-level paragraph to position recognition as "the most explicit philosophical articulation of an orientation its constructivist descendants carry tacitly," explicitly noting Hegel's lineage to Dewey (Hegelian phase), Vygotsky (dialectical psychology refracted through Marxist materialism), and Piaget (assimilation/accommodation as transposed dialectic of identity-difference). The empirical operationalisation is interchangeable across the family; what remains uniquely valuable about recognition theory is the diagnostic clarity its vocabulary provides for *what is happening* when intersubjective orientation alters tutor production.
- **§1 introduction**: replaced the single-paragraph family naming (v3.0.49 form) with a careful three-circle genealogical account distinguishing direct Hegelian descendants (Dewey, Vygotsky, *Bildung*, Honneth — innermost circle: argue from Hegel), structurally Hegelian descendants (Piaget assimilation/accommodation as transposed dialectic, Kapur productive failure as thesis-antithesis-synthesis under non-Hegelian citation — second circle: bear the structural mark without attribution), and cognate-but-not-descended pedagogically-adjacent traditions (Chi ICAP, VanLehn step-level scaffolding, Graesser affective tutoring — third circle: operate in an intersubjective-pedagogy ecosystem whose air recognition theory makes breathable, but cite elsewhere). The *family* is the union of these three circles; the active ingredient (per A10/A10b) is the orientation they share; recognition theory is the most explicit articulation of that orientation. Our contribution is accordingly twofold: operationalize the family's commitments as design heuristics (using recognition theory as the explicit theoretical entry point for its philosophical primacy and diagnostic clarity), and trace the mechanisms.
- **§3 theoretical framework**: added a "level of analysis" framing note positioning the three predicted mechanisms as recognition theory's specific articulation of commitments the broader family carries tacitly. We could in principle derive analogous mechanisms from each descendant (a Vygotskian "calibration" via scaffolding-to-the-zone, a Kapurian "error correction" via productive-failure-then-canonical-instruction, a constructivist "adaptive responsiveness" via dialogic refinement); these would be the same mechanisms named differently. We work in recognition's vocabulary because it is diagnostically clearer for *what is happening*. The empirical operationalisation generalises (§7.9); the theoretical articulation here remains recognition-specific by design.

- **§9 broader implication**: extended the v3.0.49 family-level paragraph with an explicit "empirical generalisability does not displace recognition theory" passage. Names the three innermost-circle Hegelian descendants explicitly (Dewey's pragmatism began as American Hegelianism; Vygotsky's dialectical psychology refracts Hegelian recognition through Marxist materialism; Piaget's assimilation/accommodation transposes the dialectic of identity-difference into developmental cognition). The matched-pedagogical control removes the Hegelian *vocabulary* but inherits the Hegelian *commitment*. Recognition is positioned not as "one of several alternatives that all work equally well" but as "the philosophical articulation of an orientation its constructivist descendants carry implicitly."

This is the framing Paper 2.0 should ship with: empirically generous (family-level finding is real), theoretically careful (recognition theory's primacy is preserved on philosophical grounds, not just historical priority). Closes TODO §F6 at the Framing B+ level. No data or analysis changes; ~400 words rewritten across four sections; net length change ~+450 words from §1 expansion.

v3.0.49 (2026-04-24) Paper-framing reframe to intersubjective-pedagogy family (closes TODO §F6 at Framing A level; superseded by v3.0.50 Framing B+). After A10/A10b established that recognition's effect is a family-level property of the Hegelian-descendant intersubjective-pedagogy tradition rather than a property of recognition's specific vocabulary, three passages upstream of §7.9 were updated to reflect this reframe.

- **Abstract**: added a sentence citing A10/A10b, naming "intersubjective-pedagogy orientation" as the active ingredient, and positioning recognition as "one effective operationalisation" of a broader family rather than a uniquely necessary framing. Cites pooled three-judge $d \approx 0.14$-0.19 within-Hegelian (density-sufficient), pooled $d = 0.89$ behaviorist-below-base (transmission-family harmful), pooled $d = 1.38$ between-family (dominant).
- **§1 introduction**: expanded the Vygotsky/*Bildung*/Honneth paragraph to name the intersubjective-pedagogy family explicitly, listing its theorists (Vygotsky ZPD, Piaget assimilation/accommodation, Kapur productive failure, Chi ICAP, VanLehn step-level scaffolding, Graesser affective dynamics) and citing §7.9 A10/A10b for the empirical family-level finding. Reframes the paper's contribution as operationalising the family using recognition theory as the theoretical entry point.
- **§9 broader implication**: describes the mechanisms as altering behavior at the "intersubjective-pedagogy orientation" level rather than at the "recognition" level specifically. Includes family-level findings from A10/A10b and reframes the final takeaway as "recognition is one effective operationalisation; the family as a whole is what the paper demonstrates." Methodological contribution extended to include

"orientation-family-as-unit-of-analysis" as transferable framing.

§3 theoretical framework retained with its recognition-specific theoretical motivation unchanged — it remains the paper's scholarly entry point and the specific philosophical tradition being operationalised. §7.9 (v3.0.48 empirical writeup) already reflected the family-level reframe.

This is a *calibrated* reframe (Framing A from the TODO §F6 decision fork, with explicit family-level acknowledgements) rather than a full reframe from "recognition works" to "intersubjective pedagogy works" (Framing B). The fuller family-level theoretical discussion lives in `docs/pedagogical-taxonomy.md`, which now carries structured metadata consumed by `config/tutor-agents.yaml` (`pedagogical_orientations:` map, added 2026-04-24) and by the chat UI.

No data or analysis changes; ~3 passages touched.

v3.0.48 (2026-04-24) A10 + A10b orientation-family integration into §7.9. After fixing bug_007 (v3.0.47 retraction), A10 re-ran as `eval-2026-04-23-42e7acbe` (three cells: base / recognition / matched-pedagogical under DeepSeek V3.2, three-judge panel Sonnet + GPT-5.2 + Opus 4.7). Pooled three-judge recognition-vs-matched-pedagogical $d = 0.17$ — below the pre-registered $|d| < 0.2$ density-sufficient threshold. A10 alone would have justified "density within the Hegelian-descendant family is sufficient" as the retract-branch verdict.

However, A10's matched-pedagogical prompt was grounded in Vygotsky/Piaget/Kapur/Chi/VanLehn/Graesser — all Hegelian-descendant theorists. A10 therefore tested within-family density-substitutability, not orientation-orthogonal density-sufficiency. A10b (`eval-2026-04-24-e9a785c0`) added a fourth cell (`cell_96_base_behaviorist_single_unified`) with a rigorously-authored prompt grounded in Skinner/Gagné/Keller/Thorndike/Rosenshine (behaviorist explicit-instruction tradition, zero constructivist/Hegelian content by expanded blocklist), giving a four-way three-judge comparison. Pooled three-judge findings at full n: within-Hegelian $d = 0.14$ (replicates A10); within-transmission $d = 0.89$ (behaviorist substantially *below* generic base — density in wrong family is counterproductive); **between-family $d = 1.38$ (dominant effect)**.

§7.9 rewrote the "prompt density as alternative explanation" paragraph to integrate both findings. Previously four arguments against the density reduction; now organised around A10/A10b as the central empirical closure, with the earlier four lines of evidence (Paper 1.0 placebo, elaborate-hurts-on-Haiku, autotuning gap-widens-under-optimisation, question-asking mediation) framed as converging lines. The recognition effect is resolved at the orientation-family level: reducible neither to length nor to rigorous-grounding-in-general, but reducible to membership in the intersubjective-pedagogy family. Recognition is one effective operationalisation; matched-pedagogical is another; the family as a whole is what the paper demonstrates.

Also in v3.0.48: closed TODO §A10 and added §A13 (A10b four-way comparison) and §D6 (orientation-family pedagogy taxonomy) to TODO.md; authored `docs/pedagogical-taxonomy.md` as the canonical reference; authored `exports/a10-prompt-density-control.md` and `exports/a10b-orientation-family.md` with full three-judge contrast tables at locked n; authored `scripts/analyze-a10-prompt-density-control.js` (points at A10 v2, not invalidated v1) and `scripts/analyze-a10b-orientation-family.js`; added `pedagogical_orientations:` metadata block to `config/tutor-agents.yaml` for chat-UI consumption. Extended Supplement IV with A10 v2 and A10b run IDs (61 total). Structural-features confound caveat added to `docs/pedagogical-taxonomy.md` "Methodological caveat" section — flags that LLM judges may reward surface features (warmth / directness / structure) independently of theoretical content; human-learner pilot (§A1) remains the only way to break this confound fully.

v3.0.47 (2026-04-23) A10 result retracted: profile-resolution bug routed `cell_95` to the base prompt. An `/ultrareview` pass on the v3.0.46 branch identified a pipeline bug (bug_007) in `services/evaluationRunner.js`: the `resolveEvalProfile` dispatch chain at lines 220–248 has explicit branches for every `prompt_type` in use at time of v3.0.42 (`enhanced`, `placebo`, `hardwired`, `naive`, `memory`, `recognition_nomem`, and the six `divergent_*`/`dialectical_*` variants) but no branch for the newly-introduced `matched_pedagogical` prompt type on cell_95. With `recognition_mode: false`, control fell through to the final `else` clause and `resolvedProfileName` was set to `'budget'` — identical to cell_1's resolution. The YAML's `ego.prompt_file: tutor-ego-matched-pedagogical.md` is read only by `collectPromptVersions` for provenance-hash metadata, not threaded through to `tutorApi.generateSuggestions`. Direct evidence of the bug: raw cell_95 tutor outputs on the `new_user_first_visit` scenario are lexically indistinguishable from cell_1 outputs ("As a new user, begin with this introductory lecture…" template, same phrasing, same tone), and distinctly different from cell_5 outputs ("Welcome! As a new learner, let's begin…"). The $d = 0.98$–1.62 "recognition beats matched-specificity" finding reported in v3.0.46 was in fact a "cell_5 beats cell_1 measured a second time" comparison with run-ordering noise.

Retraction. §7.9 reverted to its v3.0.45 form: four arguments against the density alternative (placebo, autotuning, mediation, architecture isolation), with an acknowledgement that the strongest version of the alternative — matched *specificity* not just length — remains untested under the Paper 2.0 v2.2 rubric. The §7.9 paragraph now explicitly names A10 as launched-but-invalidated and references this Supplement V entry. Supplement IV retains the A10 run ID (`eval-2026-04-22-04497df0`) labelled as "**A10 v1 — INVALIDATED BY BUG 007**"; the re-run will get a new run ID.

Other findings from the same review pass. - **bug_011** (normal): `tests/factorial-design.test.js` hard-codes three allow-lists (`VALID_PROMPT_TYPES`, `namePatterns`, `egoPromptFiles`) that do not include `matched_pedagogical`.

`npm test` fails on cell_95 with three assertions. Fixed in this revision by adding `matched_pedagogical` entries to each allow-list. - **bug_001** (normal): `server.js:55` statically imports `./routes/chatRoutes.js`, which does not exist. Unrelated to the scientific work; flagged for a separate fix-up cycle.

Methodological note. The bug was not caught by `validate-config` (which only checks that the prompt file *exists* in the tutor-core prompts directory, not that the profile-resolution chain actually loads it) nor by the factorial-design tests (which fail on cell_95 but only on a secondary assertion path, not on the "is this cell loading the right prompt" question). The general lesson — that authoring a new prompt file plus a YAML cell plus an `EVAL_ONLY_PROFILES` registration is *not sufficient* to cause the prompt to run; the `resolveEvalProfile` dispatch chain also needs an explicit branch — is now a check in the §6.4.2 / §7.9 authoring workflow and is flagged in CLAUDE.md's existing "Adding New Cells" section (which already warned about EVAL_ONLY_PROFILES but not about the dispatch chain). The direct way to catch this class of bug going forward: a regression test that asserts, for each cell in `EVAL_ONLY_PROFILES`, that `resolveEvalProfile(cell).resolvedProfileName` does **not** fall back to `'budget'` unless the cell's YAML explicitly has `prompt_type: base`.

A10 re-run in progress under the fixed pipeline. Results, along with the relevant §7.9 re-extension (or confirmation that density alone suffices, forcing a retraction of the broader content-over-density claim), will be reported in the next revision.

v3.0.46 (2026-04-22) A10 prompt-density control closes §7.9 loophole; A11 M2-isolation strengthens §6.4 residual claim. Two pre-registered experiments complete the post-critique cycle opened in v3.0.43.

A10 (`eval-2026-04-22-04497df0`, DeepSeek V3.2, 3 cells × 3 runs × 21 scenarios, n = 62-63/cell after gap-fill). The §7.9 "content over density" claim had one unaddressed skeptical reading: that the Paper 1.0 placebo may not have been equally *specific* in pedagogical guidance. A10 authored a new matched-specificity prompt `tutor-ego-matched-pedagogical.md` (2,835 words, 0.9% over recognition; structurally parallel; grounded in named educational research — Piaget, Vygotsky, Bloom, VanLehn, Kapur, Chi, Graesser; blocklist-verified for zero recognition/Hegelian content) and ran cells 1 (base, 344-line) vs 5 (recognition) vs new cell_95 (matched-pedagogical) under v2.2 rubric and the Paper 2.0 judge panel. Pre-registered decision grid: if $|d(5 \text{ vs } 95)| < 0.2$, retract the content-over-density claim. GPT-5.2 (full n = 62-63): cell_1 $M = 39.2$, cell_95 $M = 40.7$, cell_5 $M = 52.8$; **recognition vs matched $\Delta = +12.1$ pts,** $d = 0.98$, Welch's $t(115.8) = 5.51$. Sonnet (n = 45-47/cell, 73% coverage): same direction, stronger at larger n — recog vs matched $\Delta = +19.4$, $d = 1.62$. Both judges converge on the primary verdict: **density alone is not sufficient; recognition content adds substantial independent lift**. §7.9 extended from four arguments to five, with the matched-specificity control added as the closing piece. Exports: `exports/a10-prompt-density-control.md`. Analysis script: `scripts/analyze-a10-prompt-density-control.js`.

A11 (`eval-2026-04-22-b56be6c7`, Gemini Flash 3.0 ego × Kimi K2.5 superego, cells 82/83, 3 runs × 21 scenarios). The §6.4.1 "universal substitution with model-dependent residual" claim carried a +12.3-point residual architecture benefit on Gemini Flash that was *inferred* from the factorial interaction rather than directly measured. A11 ran the direct M2-alone isolation (base + superego on Gemini Flash) and compared against the matched base-single-agent cells 80/81 from run 18027efc. Pre-registered decision grid: $\Delta \geq +9$ AND $d \geq 0.8$ to confirm the residual. Result under Sonnet ($n = 116$ on A11 cells 82/83 vs $n = 36$ on baseline cells 80/81, 93% coverage): $\Delta = +16.3$ **pts,** $d = 1.46$**, Welch's** $t(72.9) = 8.65$ — not just confirming the factorial-inferred residual but substantially exceeding it. The M2-alone magnitude is far larger than the DeepSeek comparison (+9.2 pts, $d = 1.13$ from run 768ba77b), supporting the monotonic pattern: weaker generation models → larger superego residual. §6.4.2 extended with a new "Direct isolation of the Gemini Flash residual (A11)" paragraph; §7.3 citation updated from inferred to directly measured. Exports: `exports/a11-m2-isolation-gemini-flash.md`. Analysis script: `scripts/analyze-a11-m2-gemini-flash-isolation.js`.

Combined net effect on the paper. After the v3.0.43-45 arc (A12 retiring the M3 exploratory effect) and the v3.0.46 arc (A10 + A11 strengthening the two remaining mechanisms), the paper's mechanism story is materially tighter than at v3.0.42 before the internal critique: two mechanisms supported (both now with direct isolation evidence on at least one model beyond the factorial), one mechanism cleanly null (replication failed at pre-registered thresholds), the "content over density" alternative explanation directly disconfirmed under the same rubric and judge panel as the headline claims. Supplement IV extended with A10 + A11 run IDs (59 total). TODO §A10, §A11 → RESOLVED. Cost for this cycle: ~$40 OpenRouter (generation + GPT rejudge) + partial Sonnet (subscription, daily cap resets 2026-04-23). Sonnet completion scheduled for the next session will tighten partial n on both analyses but is not expected to change verdicts.

v3.0.45 (2026-04-22) A12 M3 disengagement replication closes the exploratory M3 boundary-condition as null. The $d = 1.63$ finding on DeepSeek/Sonnet in the 10-turn `trajectory_disengagement_to_ownership` scenario was flagged in v3.0.43 as a pending boundary-condition finding requiring replication across models, judges, and scenario variants. A12 ran pre-registered replication across two additional generation models (Haiku 4.5, Gemini Flash 3.0) under two judges (Sonnet 4.6, GPT-5.4) for the same cells (80 base vs 84 recognition), same scenario, $n = 4$/condition per model ($N = 32$ dialogues total across runs `eval-2026-04-22-d4547979` and `eval-2026-04-22-f4fb03f1`). The pre-registered decision grid required $d \geq 1.0$ on slope in both replication models and cross-judge $\Delta d \leq 0.5$ on identical rows.

Result matrix (Cohen's d on per-dialogue OLS slopes): Haiku/Sonnet −0.18, Haiku/GPT-5.4 +1.85 (cross-judge $\Delta d = 2.03$); Gemini Flash/Sonnet −0.93,

Gemini Flash/GPT-5.4 −0.11 (cross-judge $\Delta d = 0.82$). Three of four cells fall below the $d = 0.5$ fails threshold; the one cell reaching $d \geq 1.0$ is contradicted by the secondary judge on identical rows at $\Delta d = 2.03$ ($ $4× the paper's documented cross-judge variation on the main factorial, §8.3). Under the pre-registered §4.3 sensitivity rule, this disqualifies the "replicates" reading of the Haiku/GPT-5.4 cell. Under the "fails to replicate on both models" branch of the §4.1 decision grid, disengagement hedging is retired throughout.

Changes vs v3.0.44: (1) Abstract, §1 intro paragraph, §1 contribution bullet, §3.2 note-on-evidence, §6.3.8 connecting-to-predictions, §6.4.3 mechanism-separability table, §6.4.5 interaction predictions, §7 intro, §7.8.2 prompt-lab discussion, §9 conclusion, and §9 closer all updated from "pending boundary-condition effect" framing to "null on both the primary analysis and pre-registered replication." (2) §6.3.2 canonical treatment rewritten: keeps the original DeepSeek/Sonnet data as descriptive but adds the A12 replication-failure paragraph with full result matrix and the judge-sensitivity observation as the headline verdict. (3) Figure caption for `figure-disengagement-divergence.png` updated to mark the pattern as generation-model × judge-specific and flag the A12 replication failure. (4) Supplement IV extended with A12 run IDs (57 total). (5) TODO A12 → RESOLVED (failed to replicate).

Secondary methodological finding: the Haiku $\Delta d = 2.03$ between Sonnet and GPT-5.4 on identical 8 dialogues is $ $4× the judge-variation documented on the main factorial (§8.3 reports Δd ranges of 0.47–0.58). This is consistent with slope-based trajectory analyses being substantially less judge-robust than level estimates in extended-turn scenarios, and sharpens the LLM-as-judge caveats already flagged in §8.2 and §8.3 specifically for slope-based inference. Analysis script: `scripts/analyze-a12-disengagement-replication.js`; full exports: `exports/a12-disengagement-replication.md`. Cost: ~$7 OpenRouter total.

v3.0.44 (2026-04-22) Apparatus-as-method tightening (TODO §F5; no data changes). §7.4 reduced from ~51 lines to 33 lines (≈35% cut) without losing the distinctive argument. Changes: (1) §7.4 preamble consolidated from two paragraphs to one. (2) §7.4.1 three correction examples (active control reframing, judge version unification, learner prompt leakage) consolidated into a single sentence with parenthetical summaries rather than three separate paragraphs. (3) §7.4.2 dropped the redundant worked-example paragraph (already fully treated in §5.9), folded the scale details (119 claims, 18 adapters, DAG) into one sentence, removed the restated "practical lesson" at the end. Kept the dependency-graph contribution argument with its load-bearing learner-superego-paradox cascade example. (4) §7.4.3 removed the bulleted v1.0 → v2.2 rubric history (cross-reference to §5.2.6 where it fully lives), compressed to the error-correction-pattern insight plus a forward-reference to §8.6 PCA as a natural extension of the 14 → 8 construct refinement. (5) §7.4.4 condensed test-suite stats into one compound sentence and removed the two-paragraph framing into a single tight paragraph. No content claims changed; only the expository

overhead was cut. Closes TODO §F5.

v3.0.43 (2026-04-22) Post-critical-review tightening pass (TODO §F1-F3; no data changes). Three coordinated edits driven by an internal 2026-04-22 critical review of v3.0.42. (1) **Collapse M3-disengagement exposure**: the $d = 1.63$ single-scenario exploratory finding (one model, one judge, one scenario, $n = 12$/condition) previously appeared in 13 places across abstract, §1 intro, §1 three-mechanism list, §1 contributions, §3 preface, §3.2 note-on-evidence, §7 intro, §7.8.2, §9 conclusion, §9 broader implication, and §9 Hegelian closer. The canonical full treatment is retained at §6.3.2 (prose, table, figure); every other mention collapsed to a one-clause pointer ("§6.3.2 reports an exploratory single-scenario effect flagged as pending replication"). Removes the bait-and-switch risk that a pending finding carrying significant narrative weight could mis-calibrate reader confidence in M3. (2) **§6.1 calibration lead rewritten for §8.6 PCA consistency**: added a paragraph ("A note on dimension independence") clarifying that the 8 v2.2 tutor dimensions load on one factor (PC1 = 80.7%, mean inter-dim $r = 0.776$), so "narrowing within-response dimension variance" is equivalent to compressing each response's within-response *floor-to-ceiling range* across correlated indicators rather than tightening 8 independent skills. Figure caption for `figure-calibration-variance.png` reworded accordingly ("narrows the dimension profile" → "compresses the within-response floor-to-ceiling range across the 8 correlated tutor dimensions"). No metric or number changes; only the interpretation is brought into line with §8.6 PCA. (3) **Judge-pooled d in headline claims**: the §1 intro operational paragraph previously reported "$d = 1.85$ floor-lifting improvement" without judge attribution. Updated to report the judge-pooled panel (pooled $d \approx 1.63$, with Sonnet-judge $d \approx 1.88$ as upper bound and Gemini-3.1-Pro-judge $d \approx 1.44$ as lower bound, cross-referencing §8.3). §9 conclusion similarly carries the pooled framing. Aligns the headline with §8.3's per-judge decomposition and removes a silent Sonnet-only bias from the abstract-adjacent claims. Also opened TODO §F (Paper Revisions) with F1-F5 tracking these plus two deferred editorial tasks (redundancy trim, apparatus-as-method tightening) and TODO §A10-A12 + D5 capturing the experimental follow-ups from the same critical review.

v3.0.40 (2026-04-19) Writing Pad controlled ablation (A5 closure). Added §6.6.9 "Writing Pad Controlled Ablation" reporting a 2×2 factorial test of whether the Freudian three-layer memory Writing Pad is load-bearing for the recognition effect on the dialectical-ego architecture. Cells 93 and 94 are byte-identical clones of cells 40 and 41 with `writing_pad_enabled: true` flipped to `false`. Run eval-2026-04-17-f1e851c3 ($n = 63$ per cell, $N = 252$, Nemotron 3-nano ego × Kimi K2.5 superego, Sonnet 4.6 judge) gives: recognition main effect $F(1, 248) = 27.10$, $p < .001$, $\eta^2 = 0.097$ (large); Writing Pad main effect $F(1, 248) = 3.96$, $p = .048$, $\eta^2 = 0.014$ in the *opposite* direction from the necessity hypothesis (pad OFF scores 3.2 points *higher*); interaction null, $F(1, 248) = 0.82$,

$p = .366$. Recognition-within-pad-condition: $d = 0.56$ with pad on, $d = 0.74$ with pad off — both "medium-large," larger without the pad. The Writing Pad is therefore **not load-bearing** for the recognition effect; recognition operates at the prompt level rather than via the three-layer memory architecture. Three scope limits flagged: single ego-model pair (Nemotron × Kimi), single superego disposition (suspicious), first-turn-score-only (multi-turn coherence effects not isolated). The pad's marginal-significance negative main effect does not survive basic Bonferroni correction and is read as "does nothing useful here" rather than "actively harmful." Added `scripts/analyze-a5-writing-pad.js` (reproducible 2 × 2 ANOVA with regularised incomplete beta for F-to-p conversion) and `exports/a5-writing-pad.md` (generated report). Cost: ~$6 OpenRouter for generation (multi-day run, hit insufficient-credits partway and resumed after top-up); Sonnet judging via subscription. Closes v4 roadmap A5.

v3.0.39 (2026-04-17) Disposition gradient architecture scope limit (D4 closure). Added §6.6.8 "Disposition Gradient: Architecture Scope Limit" reporting that §3.4 Prediction 3's suspicious > adversary > advocate ordering is specific to the dialectical-ego architecture (cells 40–45) rather than a domain-general property of the recognition mechanism. Three configurations compared under matched n: (a) dialectical ego × philosophy (cells 40–45, Haiku × Opus, $n = 17$–18) reproduces the monotone gradient (susp $d = 0.85$, adv $d = 0.62$, advocate $d = 0.51$); (b) standard ego × philosophy (cells 22–27, Haiku × Opus, $n = 5$–10) shows a reversed ordering (susp $d = -0.01$, adv $d = 0.97$, advocate $d = 1.70$); (c) standard ego × SEL (cells 22–27, new run eval-2026-04-17-4a9b765a, Haiku × Sonnet, $n = 22$–24 per cell, 141/144 rows judged — 3 interpersonal-conflict rows with multi-suggestion format skipped by the per-turn evaluator) shows a non-monotonic pattern (susp $d = 1.25$, adv $d = 0.53$, advocate $d = 1.09$). The monotone gradient appears only in configuration (a); (b) and (c) bound the claim. Updated §3.4 Prediction 3 wording to flag the scope limit and forward-reference §6.6.8. Confounds flagged: (i) judge-model differs between D4 SEL (Sonnet 4.6) and philosophy baselines (Opus 4.6), so absolute Δ magnitudes are not directly comparable across the three rows; (ii) cells 22–27 philosophy has small n per cell; (iii) the architecture-matched cells 40–45 × SEL replication is deferred on cost grounds and is the clean next test. Added `scripts/analyze-d4-disposition-gradient.js` (reproducible analysis) and `exports/d4-disposition-gradient.md` (generated report). Cost: $5–6 OpenRouter for the SEL run. Closes v4 roadmap D4 at "architecture-scope-limit" level; learner-side disposition variants remain out of scope for this pass.

v3.0.38 (2026-04-17) Cross-application adjacency pilot: peer support coaching (D2 Path 1). Added §6.6.7 "Cross-Application Adjacency Pilot" reporting a scope-honest single-application probe of whether the recognition main effect survives when the *skill being coached* runs counter to traditional pedagogy. Built a new content package under

`content-test-support/` ("Being With, Not Fixing" — 4-lecture peer support listener training course + 4 core + 1 mood scenarios) that mirrors A6's scenario structure but frames the tutor as coaching a trainee to *not* advise, reassure, paraphrase, or resolve. Ran cells 1 (base) and 5 (recognition) under single-prompt mode, Haiku 4.5 generation, Sonnet 4.6 judge, 3 runs × 5 scenarios, $n = 15$ per cell (eval-2026-04-17-6766015b). Result: base 52.25 (SD 9.63), recognition 69.92 (SD 12.73), $\Delta = 17.67$, $d = 1.57$ ("very large" under Cohen's conventions). Closest A6 adjacency (SEL) sits at $d = 1.82$; peer-support is $\Delta d = -0.25$ below SEL and comfortably inside the A6 domain range ($d = 1.45$–2.71). Directional replication of the recognition main effect therefore survives a domain shift into a skill that runs structurally counter to conventional knowledge-transfer pedagogy. Three scope limits flagged: (a) the LLM is still prompted as a tutor, making this an *adjacency* test rather than a true cross-application test (the role-reframed D2 Path 2 study is deferred); (b) single application does not support a cross-application generalization claim; (c) Sonnet-only judging. Added `scripts/analyze-d2-support-pilot.js` (reproducible analysis) and `exports/d2-support-pilot.md` (generated report). Updated §6.6.6 closing paragraph with forward-reference to §6.6.7. Cost: $3–4 OpenRouter. Addresses v4 roadmap D2 at Path 1 (Path 2 deferred to a separate study).

v3.0.37 (2026-04-17) Domain generalization: five-domain recognition replication (A6 closure). Extended §6.6.6 from a two-domain (philosophy, programming) replication to a five-domain treatment adding elementary math, creative writing, and social-emotional learning (SEL) content packages. Each new domain ran cells 1/5 on 5 single-turn scenarios (4 core + 1 mood) × 3 runs = $n = 15$ per condition under matched Haiku 4.5 generation × Sonnet 4.6 judge. Run IDs: eval-2026-04-17-fde764e1 + 7915be04 (math), 428f0649 + 93ba2f9a (creative), 44fa989c + 5833830a (SEL). Results: programming $d = 2.33$, math $d = 1.45$, creative $d = 1.96$, SEL $d = 1.82$, philosophy $d = 2.71$. All five domains show recognition > base with every d classified "very large" under Cohen's conventions ($d > 1.3$). Directional replication of the recognition main effect therefore holds across humanities (philosophy), technical (programming, math), creative (writing), and affective (SEL) content under matched generation × judge. Domain-×-recognition interaction ranges $\Delta d = -0.38$ (programming) to $\Delta d = -1.25$ (math) vs the philosophy anchor, but interpretation is mode-confounded (non-philosophy = single-prompt, philosophy = messages); cross-judge validation was not re-run on the four new domains for this closure (Sonnet-only). Rewrote §6.6.6 body to reframe as "five-domain generalization" rather than "programming replication"; replaced 2-row table with 5-row table; updated narrative to reflect $d = 1.45$–2.71 magnitude range and 5-domain directional replication. Added `exports/a6-domain-generalization.md` (final report) and updated `scripts/analyze-domain-generalization.js` to iterate over all five domain contrasts. Cost: $25–30 OpenRouter across the four new domain packages.

Closes v4 roadmap A6.

v3.0.36 (2026-04-17) Domain generalization: recognition replicates on programming (C3 closure). Added a new §6.6.6 "Domain Generalization" subsection reporting a programming-domain replication of the recognition main effect. Ran cells 1 (base, single-agent, unified learner) and 5 (recognition, single-agent, unified learner) on a programming-tutorial scenario set (5 single-turn scenarios: debugging, code review, conceptual explanation), 3 runs × 5 scenarios, $n = 15$ per condition, Haiku 4.5 generation, Sonnet 4.6 judge (eval-2026-04-17-c92ad6c7). Compared against the closest matched-generation-and-judge philosophy contrast (eval-2026-03-02-45163390 cells 80/84, messages-mode, $n = 18$ per condition). Result: programming $d = 2.33$ (base 47.75 SD 9.10, recog 70.42 SD 10.30); philosophy $d = 2.71$ (base 43.61, recog 74.10). Both "very large" under Cohen's conventions ($d > 1.3$); domain-×-recognition interaction $\Delta d = -0.38$ under a mode-confounded comparison. Recognition main effect generalizes from philosophy to programming in direction and in very-large-magnitude class. Updated §6.6.5 closing paragraph with a forward-reference. Confounds flagged: programming uses single-prompt-mode cells while the matched philosophy cells are messages-mode (mode is a known moderator, §6.3, §6.5); scenario counts differ (5 vs 9) and scenario pairs are not matched. Added `scripts/analyze-domain-generalization.js` (reproducible analysis) and `exports/c3-domain-generalization.md` (generated report). Cost: $3 OpenRouter. Closes v4 roadmap C3.

v3.0.35 (2026-04-17) Tutor–learner asymmetry: rubric vs holistic adjudication (M6 full closure): Added a new §6.5.5 "Rubric vs Holistic Adjudication" subsection (pushing "Implications" to §6.5.6) reporting a cross-judge convergent-validity check on the 7–12× tutor-learner recognition-effect asymmetry. Scope: messages-mode 2$×2×$2 factorial cells 80–87 under rubric v2.2, $N = 1{,}552$ paired rows with all four metrics populated (tutor rubric, tutor holistic, learner rubric, learner holistic), three independent judges (Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4). Primary test: does the tutor–learner d-gap shrink under holistic scoring? Result: pooled rubric d-gap = 0.741 (tutor $d = 1.120$, learner $d = 0.379$); pooled holistic d-gap = 0.327 (tutor $d = 0.565$, learner $d = 0.238$); **shrinkage** = 55.9%. Replicates tightly across all three judges (Sonnet 55.7%, Gemini 49.9%, GPT 55.7%). Within-role rubric holistic correlations do not break down (tutor $r = 0.57$, learner $r = 0.75$ pooled), so the shrinkage reflects the rubric and holistic instruments weighting recognition's effects differently rather than judge disagreement. Conclusion: $ $50% of the originally reported 7–12× asymmetry is rubric-artifact; the residual holistic gap (ratio ≈ 0.42) is the non-rubric-dependent portion, plausibly reflecting the structural explanation (prompt scope + downstream learner). Parallels the C4 paradox-closure pattern (§8.1): same-rows, same-shrinkage ($ $50%), different analysis target. §6.5.6 Implications updated to note the magnitude reframing. Added `scripts/analyze-tutor-learner-asymmetry.js` (reproducible analysis) and `exports/m6-asymmetry.md` (generated report). No

new runs; DB-only analysis of existing paired data. Closes v4 roadmap M6 at full closure.

v3.0.34 (2026-04-17) Learner paradox: cell-matched holistic scores on the original paradox run (C4 full closure): Executed the final-mile adjudication promised in v3.0.33. Added holistic learner scores to the originally-scored 48 rows of eval-2026-02-20-25c78e91 (single-prompt mode, rubric v1.0, Opus 4.6 rubric judge; 47 paired after one judge JSON-parse failure). To avoid cross-version contamination, added a standalone `scripts/evaluate-paradox-holistic.js` that writes only the four holistic fields via direct SQL `UPDATE`, preserving the `learner_rubric_version = 1.0` marker rather than triggering `evaluationStore.updateResultLearnerScores`'s default version-overwrite (Sonnet 4.5 judge for holistic). Rubric d on the paired subset reproduces the paradox essentially exactly: base **rubric** $d = -2.82$ ($\Delta = -23.3$, $n = 11/12$); recognition **rubric** $d = -3.11$ ($\Delta = -22.3$, $n = 12/12$). Same rows, holistic: base $d = -1.34$, recognition $d = -1.06$ — roughly half the rubric effect size. Within-architecture rubric holistic r on the paradox cells is 0.06–0.44 versus the corpus-wide 0.72–0.77, i.e., the two judges specifically diverge on these cells rather than agreeing as they do elsewhere. Added a fourth §8.1 paragraph reporting this cell-matched result; updated §7.9 to cross-reference. The paradox is now characterized as a **combined measurement-plus-single-prompt-mode interaction** (≈50% rubric artifact, ≈50% single-prompt-mode interaction, ≈0% capability ceiling), not a general property of internal self-critique. The ∼$2K higher-capability-model experiment previously envisioned is no longer required for the central adjudication. Extended `scripts/analyze-learner-paradox-holistic.js` with `--run-id` and `--cell-min-n` flags; regenerated `exports/learner-paradox-holistic-analysis.md` with paradox-run slice. Cost: $2–3 OpenRouter. Addresses v4 roadmap C4 at full closure.

v3.0.33 (2026-04-17) Learner paradox: convergent-validity evidence from holistic scoring (C4 tightening): Added a third §8.1 paragraph reporting a corpus-wide convergent-validity check on the learner-rubric vs holistic-learner-scoring relation ($n = 1{,}759$ paired rows across the four learner architectures: `unified` $n = 432$, `ego_superego` $n = 411$, `unified_recognition` $n = 446$, `ego_superego_recognition` $n = 470$). Within-architecture Pearson correlations are $r = 0.720$–0.773 (convergent validity holds uniformly — it does **not** break down for ego_superego learners, which would be expected under a deliberation-invisible-to-rubric account), and the pooled ego_superego-vs-unified contrast goes in the *opposite* direction from the original $d = 3.05$ paradox on both metrics (base $\Delta = +4.14$ rubric / $+7.39$ holistic, $d = 0.25$ / 0.39; recognition contrast near zero, $|d| < 0.08$). The cell-level breakdown of messages-mode cells 80–87 ($n = 179$–216 per cell) consistently shows ego_superego $\geq$ unified on both metrics. This reframes the $d = 3.05$ result from "general architecture property" to "scenario-and-mode-specific interaction between

rubric-content dimensions and deliberation-internal content generation," which is what the rubric-artifact account predicts. §7 discussion paragraph updated to cross-reference the convergent-validity result. The remaining definitive test — adding holistic learner scores to the original paradox run (eval-2026-02-20-25c78e91, zero holistic rows) — is now a modest follow-up rather than the ~$2K higher-capability-model experiment originally priced. Added `scripts/analyze-learner-paradox-holistic.js` (reproducible analysis) and `exports/learner-paradox-holistic-analysis.md` (generated report). No new runs; DB-only analysis of existing paired data. Addresses v4 roadmap C4 at the convergent-validity level; the final-mile holistic-scoring step remains the only outstanding work to fully close C4.

v3.0.32 (2026-04-16) Inter-LLM baseline landed (C2): Executed the pilot infrastructure from v3.0.30. Generated the stratified 40-item sample (`exports/human-validation-pilot-sample.csv`, `exports/human-validation-pilot-key.jsonl`), then re-classified the same 40 items with Sonnet 4.5 (vs Haiku 4.5 in the original corpus, identical taxonomy prompt, temp=0, JSON response format) via a new `scripts/human-validation-reclassify.js` helper. Ran `scripts/human-validation-analyze.js --synthetic-rater sonnet:...` and obtained **Cohen's = 0.633 (substantial agreement), n=40, observed agreement 0.675, chance agreement 0.114**. Added an "Inter-LLM Baseline Result" subsection to §5.3 reporting the , per-category F1 (4 high, 4 moderate, 1 pilot-unstable), and confusion-matrix diagnosis. FABRICATION flagged as the single pilot-unstable category (Haiku=4, Sonnet=0; Sonnet redistributes to CONTEXT_BLINDNESS/VAGUENESS/MEMORY_FAILURE). RECOGNITION_FAILURE identified as a catch-all under Sonnet's decision policy — matches codebook's boundary rule that it should only apply when intellectual-autonomy framing is explicit. This =0.63 now serves as the lower-bound reference for the upcoming human-rater pilot: human-LLM substantially above 0.63 indicates non-idiosyncratic tracking. Cost: $0.10 API. No pre-registered hypothesis change; pure methodological closure.

v3.0.31 (2026-04-16) Learner superego paradox: §7 §8.1 harmonization (C4 partial): Updated the §7 discussion paragraph on the d=3.05 paradox (previously framed via Hegelian *Bildung*-timescale and implementation-fidelity speculation) to lead with the v3.0.26 rubric-measurement artifact account as the favoured mechanism, treating the two Hegelian alternatives as live-but-secondary rather than primary. Also tightened the §8.1 opener so the paradox-introduction paragraph cross-references the rubric-artifact mechanism that immediately follows, rather than closing on "we cannot know." Fixed stale "Section 6.16 of the pilot study" reference to canonical "Paper 1.0, Section 6.16" form matching the P5/P2 disambiguation pattern. No new data or runs — purely an internal-consistency fix that aligns the §7 speculative framing with the §8.1 DB-mechanism-check result already in the paper. The stronger empirical test (higher-capability learner model, ~$2K) remains

pending and is now called out explicitly in both §7 and §8.1 as the outstanding step. Addresses v4 roadmap C4 at the textual-consistency level; closing C4 fully still requires the pending higher-capability test.

v3.0.30 (2026-04-16) Human validation pilot apparatus (C2): Operationalised the §5.3 / §8.2 human-coding validation step that previous revisions had flagged as "outstanding." Added a "Human Validation Pilot" subsection to §5.3 describing the end-to-end pipeline: stratified sampling (`scripts/human-validation-sample.js`, deterministic seed=20260416, 4 items × 10 categories = 40-item packet), per-category codebook (`docs/research/human-coding-codebook.md` v1.0 with signal phrases, boundary rules for adjacent categories, a 12-step decision flowchart, and one in-corpus exemplar per category except REGISTER_MISMATCH where n=0), analysis pipeline (`scripts/human-validation-analyze.js` computing Cohen's , Fleiss' , per-category F1, and confusion matrices with Landis-Koch interpretation bands), and inter-LLM baseline (Haiku vs Sonnet on the same sample) as a lower-bound reference. Updated §8.2 to reference the now-available apparatus rather than describing the procedure as future work. Target floor: ≥ 0.60 (substantial agreement). Pilot data collection is the only remaining step — the infrastructure is now reviewer-runnable end-to-end. Addresses v4 roadmap C2.

v3.0.29 (2026-04-16) Title-page date sync: Updated frontmatter `date: "March 2026"` → `date: "April 2026"` to match actual revision cadence (10 revisions in April: v3.0.20 – v3.0.28 plus the short paper v1.1.0). The short paper was already April 2026 after the P4 sync (v1.1.0), so the full paper title page was the sole remaining drift. No body content changed; single-line frontmatter edit.

v3.0.28 (2026-04-16) Revision-history judge-panel disambiguation (P2): Added an orientation header to Supplement V explaining that entries v1.0–v2.3.21 are pilot-study (Paper 1.0) revisions using Opus 4.5/4.6, while entries v3.0.0+ are Paper 2.0 mechanism revisions validated by the Sonnet/Gemini-Pro/GPT triple-judge panel reported in the abstract. Opus 4.6 is retained for specific side-analyses (§8.1 learner-paradox DB check, §6.4.2 isolation runs) and those uses are labelled in-place. Addresses the v4 P2 concern that revision-history mentions of Opus 4.5/4.6 could be read as conflicting with the abstract's three-judge claim. No body-text changes; purely a reader-orientation note. Closes v4 roadmap P2.

v3.0.27 (2026-04-16) Supplement II cross-reference disambiguation (P5): Added explicit "Paper 1.0" markers to six Supplement II section headers (B.6 Section 6.20, B.7 Section 6.8, B.8 Section 6.10, B.8b Section 6.21, B.8c Section 6.22, B.9 Section 6.11) and to Supplement III.2's "see Section 6.15" reference in the dimension weights paragraph. Paper 2.0's §6 has only six subsections (6.1–6.6), so the original unqualified "Section 6.X" pointers in Supplement II were stranded cross-references under a standalone read of Paper 2.0. The figure/table numbering audit otherwise confirmed that in-text "Section 6.10/6.16/6.21/6.22" mentions already

carry "of the companion study" or "(Paper 1.0, …)" qualifiers (lines 154, 644, 1814, 1890, 1898). No other figure, table, or section-reference drift was found. Addresses v4 roadmap P5.

v3.0.26 (2026-04-16) Learner superego paradox: rubric-artifact mechanism (M6/C4 partial): Added a DB-only mechanism check to §8.1 addressing the learner superego paradox (d=3.05 deficit for ego-superego learners vs unified learners). Query on cells 1/2/5/6 (Opus 4.6 judge, N=613/108/499/157) separates learner *behaviour* from learner *score*: the ego-superego architecture produces ~2.5× more output tokens (13,193–15,547 vs 5,125–6,675) and ~4× more API calls (9.0–10.4 vs 2.3) — i.e., *more* deliberation and *more* generated text — yet scores 12–18 points *lower* on the learner rubric. Per-dimension inspection of the learner_scores JSON shows ego-superego learners score well on `deliberation_depth` and `revision_signals` (dimensions that reward the internal process) but poorly on `conceptual_engagement` and `question_quality` (dimensions that reward content in the final external message). The most parsimonious account is therefore a rubric-measurement artifact: the rubric as currently constructed rewards content-rich external output, while the ego-superego architecture spends generative effort on deliberation-internal content that does not land in the output channel the rubric scores. Reframes the paradox from "internal self-critique degrades the learner" to "the rubric does not reward the kind of behaviour the architecture produces." Addresses v4 roadmap M6/C4 alternative test (regress learner score on length/complexity features) with a DB-only analysis; the stronger test (higher-capability learner model, $2K) remains pending. No new runs — purely a re-analysis of existing data.

v3.0.25 (2026-04-16) Reproduction cost statement (P3): Added a "Reproduction cost for Paper 2.0 mechanism runs" paragraph to §8.7 using recorded database figures rather than the pre-v4 rhetorical "$50K+" estimate. Actual recorded generation cost for the three core runs (DeepSeek V3.2, Gemini Flash 3.0, Haiku 4.5; 432 dialogues, 1,296 rows) is $114.58 on OpenRouter pass-through pricing. Estimated 3-judge panel cost $300–700; superego-taxonomy classification $80–100; core reproduction total $500–1,500. Including isolation runs (§6.4.2), trajectory-specific scenarios (§6.3), and prompt-lab autotuning (§7.8) roughly doubles to $1,000–3,000. Full-programme cost (207 runs, 8,274 dialogues, ~740M tokens) was $474.75 recorded generation + estimated $2,000–5,000 total including judging. The $50K figure the v4 roadmap speculated applies only to human-learner validation or human-expert rater scaling ($20K–50K range), not to LLM-judged replication. Clarifies the replication cost story: the apparatus is inexpensive to reproduce at the LLM-judge tier, and the genuine barrier is human validation, not compute. Addresses v4 roadmap P3.

v3.0.24 (2026-04-16) Judge-version confound quantification (M5): Added a per-judge effect-size decomposition to §8.3 (Model Transience). The §6.1 summary figure of $d = 1.34$–1.92 conflates three genuinely

different judge panels: Sonnet 4.6 judge-mean $d = 1.88$, Gemini 3.1 Pro judge-mean $d = 1.44$, GPT-5.4 judge-mean $d = 1.56$, computed on tutor_overall_score across the three generation runs. Within-row Δd ranges 0.47–0.58, well above the 0.1 d robustness threshold, so absolute magnitudes are judge-dependent even though effect direction is conserved in all 9 cells. Introduced the judge-adjusted pooled estimate $d \approx 1.63$ (mean across 9 cells) with Sonnet-only $d \approx 1.88$ as an upper bound and Gemini-3.1-Pro-only $d \approx 1.44$ as a lower bound. Noted the partial leniency-gradient pattern (Sonnet highest on all three runs, but Gemini-vs-GPT ordering flips between runs), suggesting judge × generation-model interaction beyond simple calibration differences. No §6.1 headline range changed — the 1.34–1.92 figure is the min–max of the new table. Addresses v4 roadmap M5 (judge-version confound acknowledged but not tabulated).

v3.0.23 (2026-04-16) Rubric PCA single-construct acknowledgment (M3): Added explicit PCA caveat to §7.8 opener — the 8 v2.2 tutor dimensions measure largely a single underlying construct (PC1 = 80.7%, 1-factor Kaiser, mean inter-dimension r = 0.776), so "dimension-targeted optimisation" uses dimensions as interpretive proxies rather than as empirically independent skills. Results below should be read as *shifts in the single construct via targeted prompt design*, not as moving one skill while others are held constant. Added a parallel "Implication for §7.8" paragraph at the end of §8.6 so readers coming from the other direction also encounter the interpretive qualification. The forced two-factor rotation (content_accuracy vs seven pedagogical dimensions) is preserved as a meaningful bifactor split; pairs of pedagogical dimensions are not. No §7.8 findings retracted — the §7.8.1/7.8.3 results (Qwen target dims 35→75, prosthesis-straitjacket transfer pattern) stand as informative prompt-engineering findings; only the *interpretation* of per-dimension shifts is qualified. Addresses v4 roadmap M3 (rubric PCA implications not addressed in §7.8).

v3.0.22 (2026-04-16) Model-dependence reframe (M2): Scoped the "general mechanism" language at four key locations (abstract paragraph 3, §6.4.3 separability intro, §6.6.5 generalizability assessment, §7 opener) to the tested capability tier — three models spanning a ~30-point baseline range, excluding both very-weak (Nemotron-class) and frontier-capability (Opus, GPT-5.4-class generator) models. Where the paper previously said "supported as general mechanisms," it now says "replicate across all three tested models and three independent judges, with magnitudes and architectural completeness varying systematically with generation quality." Added explicit scope-limitation paragraph to §6.6.5 noting that a fourth-model replication at the frontier tier would address whether the recognition effect saturates against an already-calibrated strong model. Addresses the v4 reviewer concern that claims of "general" mechanisms overreach the three-model test set. No data, analysis, or N-count changes.

v3.0.21 (2026-04-16) Lead-with-operationalization §1 rewrite (M1 par-

tial): Added a new paragraph 2 to §1 that states what the paper does in operational terms—the specific design choices tested, the models and N, and the three headline findings (prompt $d = 1.85$; 15–17% additivity deficit with model-dependent residual; null trajectory modulation in the primary analysis)—before the Hegelian motivation. Addresses the v4 reviewer concern that the three-level operational disclaimer arrives only after 1,000+ words of Hegelian framing. The Hegelian exposition (original paragraphs 2–4 and §3's theoretical framework) is preserved intact; the new paragraph reorders the rhetorical stakes so the empirical operational claim lands before the reader invests in theory. Lightly tweaked the opener of the now-third paragraph ("This paper proposes an alternative grounded in Hegel's" → "The alternative this paper tests is grounded in Hegel's") for smooth handoff. No data, analysis, or N-count changes.

v3.0.20 (2026-04-16) Mechanistic account of universal substitution (new §6.4.2.1): Added a taxonomy-level causal chain that explains the stable 15–17% additivity deficit and the model-dependent residual architecture benefit from the critique-category data already in §6.2.2 and §6.2.3, converting the numerical substitution finding into a mechanistic claim. Pre-empted categories (vagueness, elicitation, content accuracy) correspond to the ~16% overlap where calibration operationalises the behaviours the superego was enforcing through critique; persistent categories (recognition failure, context awareness) have low within-round resolution rates and survive calibration; the model-dependent residual (+12.3 on Gemini Flash 3.0; near-zero on DeepSeek/Haiku; positive in Affective Shutdown / Frustration on strong models) tracks the volume of post-calibration error headroom and the prompt's operationalisation of the relevant competence (weakest on affective attunement). No new data collection or N-count changes—purely an interpretive synthesis of §6.2 and §6.4.2 data. Addresses the v4 reviewer note that §6.4.2 documents the substitution numerically but asserts the causal chain without supporting it.

v3.0.19 (2026-04-16) M3 honest reframe + multiplicity correction: Sharpened the Mechanism 3 (adaptive responsiveness) framing across abstract, §1 motivating paragraph, §3.2 note-on-evidence, §6.3.2 interpretation, §6.3.8 connecting-to-predictions block, §6.4.3 mechanism separability table, §6.4.5 interaction predictions, §7.8.2 prompt-optimisation discussion, and §9 conclusion. Replaced "conditionally supported" / "conditional emergent property" throughout with "**null on the primary pre-registered analysis; pending boundary-condition finding requiring replication**." Added explicit Holm–Bonferroni correction statement (family-wise = 0.05 across three scenario tests; disengagement effect survives at /3 = 0.0167 while remaining tests do not), with exact Welch's statistic ($t(21.9) = 3.99$, $p \approx .0006$) and explicit replication caveats (single generation model, single judge, single scenario, n=12 per condition). Fixes the bait-and-switch risk that reviewers flagged in a pre-v4 critique: the three-mechanism headline was being rescued

post-hoc by a small scenario subset without the multiplicity correction or replication caveats the claim required. No data, analysis, or N-count changes—purely framing correctness.

v3.0.18 (2026-03-09) Superego redundancy refinement (§7.8.4): Added subsection connecting prompt lab findings to the universal substitution claim. On weak models, targeted superego strategies amplify error correction beyond the baseline residual. On strong models, adding superego scaffolding is actively destructive, not merely redundant. Sharpens the substitution gradient from "helpful" to "harmful" across the capability threshold.

v3.0.17 (2026-03-09) Dimension-targeted optimisation and cross-model transfer (new §7.8): Added three subsections covering dimension-targeted prompt optimisation on specific rubric dimensions (§7.8.1), M3 engineering through prompt intervention (§7.8.2), and the cross-model prompt transfer experiment with prosthesis-straitjacket inversion finding (§7.8.3). Added fifth future work direction (§9) on model-stratified prompt optimisation and optimisation threshold mapping. Old §7.8 (prompt density) renumbered to §7.9. Data from prompt lab sessions on Qwen 3.5 9B and Haiku 4.5 (N=3 per transfer cell, 10 autotune iterations on Qwen, 6 on Haiku).

v3.0.16 (2026-03-07) Fix rendering: Unicode (U+2264) does not render in the xelatex font (appears as blank/replacement character in PDF), while ≥ renders correctly. Replaced all 16 body-text with LaTeX math expressions: `d  0.15` → `$d \leq 0.15$`, `0` → `$\leq$ 0`, etc. Verified ≥ (Unicode) renders correctly and left unchanged.

v3.0.15 (2026-03-07) Unicode math operators: Replaced all 45 standalone `$\to$` with Unicode → (U+2192), 16 `$\geq$`/`$\leq$` with ≥/ , 9 `$\approx$` with ≈, 31 `$\times$` with ×, 12 `$\pm$` with ±, and 4 `$\sim$` with `${\sim}N$` expressions. These standalone math delimiters caused math mode to leak into surrounding text in pandoc→xelatex rendering, producing garbled italic text in tables and prose (e.g., page 31 rubric table, page 40 floor-lifting paragraph). Full `$...$` math expressions (e.g., `$2 \times 2$`, `$d = 0.52$`) remain unchanged. Unicode characters render cleanly with xelatex without delimiter matching issues. **Note**: was subsequently reverted to LaTeX math in v3.0.16 due to font glyph issue.

v3.0.14 (2026-03-07) LaTeX formatting fixes: Fixed 9 unformatted `2x2`/`2x2x2` factorial references (now `$2 \times 2$`), 11 split `$-$` sign notations (now single math expressions like `$-0.52$`), 1 LaTeX `\ref{}` macro in figure caption (replaced with prose reference), 1 split `$\sim$.30` (now `${\sim}.30$`), and 1 split `T0$-$T1` table header (now `$T0{-}T1$`). All issues would produce rendering aberrations in the PDF.

v3.0.13 (2026-03-07) Recognition term slippage pass: 10 targeted edits across §6.1, §6.4, §6.5, §6.6, §7.2, §7.8, and §9 to prevent "recognition" from sliding between design-heuristic usage (level 2–3) and philosophical-agency usage (level 1). Key changes: "recognition eliminates" → "recognition-oriented prompts eliminate"; "recognition rescues the tutor"

→ "recognition-enhanced prompts rescue the tutor's output quality"; "genuine dialogue" → "dialogue-structured responses"; "genuinely shaped" → "specifically adapted"; "empirical echo" → "functional parallel"; "The tutor recognizes the learner" → "The prompts instruct the tutor to treat the learner as an autonomous subject." Maintains the three-senses framework established in v3.0.12.

v3.0.12 (2026-03-07) Three senses of "recognition": Added terminological clarification paragraph to §1 (after pedagogy framing, before "The ablative finding") explicitly distinguishing recognition as (1) philosophical inspiration (Hegel), (2) operational design heuristic (prompt instructions and architectural choices), and (3) observable discourse effects (question-asking, variance reduction, substantive revision). States that empirical claims concern primarily the third level. Responds to external review concern about construct validity and double duty of "recognition."

v3.0.11 (2026-03-07) Clarify evidence status markers in §3: Renamed three "Missing evidence" markers (§3.2, mechanisms 1–3) to "Evidence needed beyond pilot" and added forward references (*Addressed in Section 6.X*) to prevent readers from misinterpreting them as gaps in the current paper. Verified provable discourse counts (119 claims, 18 adapters) remain accurate.

v3.0.10 (2026-03-07) Human expert coding validation: Strengthened §8.2 to identify human expert coding as the most important outstanding validation for mechanism claims, with explicit procedure (30–50 exchanges, two coders, 10-category taxonomy, Cohen's κ, human-LLM agreement). Added matching future work entry in §9 (now "Third") with the same procedural specificity. Both sections note that the trace infrastructure makes this feasible without re-running dialogues. Added cross-reference between §8.2 and §9.

v3.0.9 (2026-03-07) Unexplained phenomena in conclusion: Added paragraph to §9 noting three findings the current mechanism model does not explain: near-identical effect magnitude across structurally different models (DeepSeek $d = 1.88$, Haiku $d = 1.84$, Gemini Flash 3.0 $d = 1.92$), the failure of adaptive responsiveness as a general mechanism, and the learner superego paradox ($d = 3.05$). Suggests these anomalies may point to a deeper process the current apparatus cannot yet trace. Responds to external review concern about consolidating what recognition doesn't explain.

v3.0.8 (2026-03-07) Added prompt elaboration baseline (Paper 1.0 §6.21) as fourth evidence line in §7.8 prompt density paragraph: the 344-line base prompt is itself a long detailed pedagogical prompt, yet it doesn't approach recognition quality and actually hurts performance on Haiku ($M = 82.5$ naive vs base, recognition $M = 90.9$). Prompt length without recognition content can be inert or counterproductive.

v3.0.7 (2026-03-07) Prompt density defense: Added "Prompt density as alternative explanation" paragraph to §7.8, consolidating three existing evidence lines against the reduction of recognition to prompt elaboration: (1)

placebo control (Paper 1.0, length-matched prompts without recognition theory), (2) autotuning experiment (§7.7, gap widens under optimization), (3) mediation analysis (§6.1.7, 42.4% mediated through question-asking). Acknowledges the strongest version of the alternative remains untested: whether *any* equally specific pedagogical prompt would match recognition.

v3.0.6 (2026-03-07) Engineered vs. discovered mechanism distinction: Added paragraph to §8.4 explicitly addressing the circularity objection: the architecture was designed for observability, not to guarantee specific outcomes; three key findings were genuinely unpredicted (superego redundancy on strong models, adaptive responsiveness null, consistent 15–17% additivity deficit). Distinguishes "mechanisms of this system" from "mechanisms of AI tutoring in general."

v3.0.5 (2026-03-07) Theoretical framing and scope: Three changes responding to external review. (1) Reframed Hegelian mapping as heuristic translation: "derive mechanisms from Hegel" → "motivate mechanisms drawing on Hegel" across abstract, §1, §3.2, §3.5, and §9 (~6 passages). Changed "derivative application" → "heuristic translation" in §3.5. The mappings are productive redescriptions, not logical derivations. (2) Added scope qualification to abstract: "All findings concern tutor output quality as assessed by LLM judges interacting with synthetic learners." (3) Added cost qualifier after contributions list: architecture is a research instrument, not a deployable system (~225K tokens/dialogue). No data or analysis changes.

v3.0.4 (2026-03-07) Epistemic language audit: Systematic softening of causal language throughout, responding to external review feedback. Changed ~45 instances: "confirmed" → "supported" for mechanism claims; "disconfirmed" → "not supported as general mechanism"; "dominant mechanism" → "strongest effect" / "primary mechanism"; "confirms [prediction]" → "is consistent with" / "supports"; "The disconfirmation of" → "The null finding for". Retained "confirmed" only for factual verification ("blind scoring confirmed by code audit") and revision history entries. Changed "Mechanism isolation confirmation" → "Mechanism isolation evidence". Softened figure captions and prediction labels throughout §6.1–6.4. No data, analysis, or N-count changes—purely rhetorical recalibration to better match the evidentiary base.

v3.0.3 (2026-03-07) Final claim audit corrections: (1) Gemini Flash 3.0 stale data fixed: `tutor_overall_score` values updated from 2026-03-03 scoring to 2026-03-05 re-scoring (17.3/37.0/52.6/64.9 → 22.4/49.3/57.7/70.0), changing the interaction from −7.4 to −14.6 and additivity deficit from 10% to 17%. (2) "Substitution-to-additive transition" narrative reframed as "universal substitution (15–17% deficit across all models) with model-dependent residual architecture benefit" across ~20 passages (abstract, §3, §5, §6.4, §6.6, §7.3, §8). (3) N=570→N=432 corrected throughout (isolation runs had expanded the figure-generation query beyond the clean factorial scope). (4) §7.3 "same model, different sample" error fixed (run 45163390 is Haiku, not DeepSeek). (5) §6.3.7

dialogue quality table updated with verified DB values and Gemini Flash 3.0 rows added. Provable discourse: 87 pass, 0 warn, 0 fail.