The Drama Machine in Education: Mutual Recognition and Multiagent Architecture for Dialectical AI Tutoring

Liam Magee
Education Policy, Organization and Leadership
University of Illinois Urbana-Champaign

February 2026

Abstract

Current approaches to AI tutoring treat the learner as a knowledge deficit to be filled and the tutor as an expert dispensing information. We propose an alternative grounded in Hegel's theory of mutual recognition—understood as a *derivative* framework rather than literal application—where effective pedagogy requires acknowledging the learner as an autonomous subject whose understanding has intrinsic validity. We implement this framework through recognition-enhanced prompts and a multiagent architecture where an "Ego" agent generates pedagogical suggestions and a "Superego" agent (a *productive metaphor* for internal quality review) evaluates them before delivery. An evaluation framework (N=892 primary scored responses across thirteen key evaluation runs; N=3,800+ across the full development database) comparing recognition-enhanced configurations against baselines reveals that recognition theory is the primary driver of tutoring improvement: a corrected 2×2 memory isolation experiment (N=120 across two independent runs) demonstrates that recognition produces large effects with or without memory (+15.2 pts without memory, d=1.71; +11.0 pts with memory), while memory alone provides only a modest, non-significant benefit (+4.8 pts, d=0.46, $p \approx .08$). The combined condition yields the highest scores (91.2, d=1.81 vs base), with a small negative interaction (-4.2 pts) suggesting ceiling effects rather than synergy. A post-hoc active control (N=118), using length-matched prompts with generic pedagogical best practices but no recognition theory, scores 66.5—approximately 9 points above same-model base (≈58) but well below recognition levels (≈73), with recognition gains (~+15 pts above same-model base) substantially exceeding active-control gains (~+9 pts). (The active control ran on a different ego model than the factorial, precluding direct cross-condition comparison; same-model historical data provides the fair baseline.) A preliminary three-way comparison (N=36) found that recognition outperforms enhanced prompting by +8.7 points,

consistent with recognition dominance, though the increment does not reach significance under GPT-5.2 (+1.3 pts, p=.60). An exploratory analysis of multi-agent synergy (+9.2 points, N=17, Nemotron) suggested this effect may be specific to recognition prompts, but a dedicated Kimi replication (N=60) found negligible interaction (+1.35 pts), indicating this is model-specific rather than a general phenomenon (Section 6.4). Domain generalizability testing across both models and content domains confirms recognition advantage replicates: elementary math with Kimi shows +9.9 pts (d $\approx$ 0.61, N=60), with effects concentrated in challenging scenarios. The factor inversion between domains (philosophy: recognition dominance; elementary: architecture dominance) is partly model-dependent—Kimi shows recognition dominance on elementary content, revising the Nemotron-only finding. Multi-agent architecture serves as critical error correction when models hallucinate trained content on new domains. Bilateral transformation tracking confirms that recognition-prompted tutors measurably adapt their approach in response to learner input (+36% relative improvement in adaptation index), providing empirical grounding for the theoretical claim that recognition produces mutual change. A step-by-step evolution analysis of dynamic prompt rewriting (cell 21: LLM-authored session directives + active Writing Pad memory) across three iterative development runs (N=82) suggests the Writing Pad as an important enabler: cell 21 progresses from trailing its static baseline by 7.2 points to leading by 5.5 points, with every rubric dimension improving. A cross-judge replication with GPT-5.2 confirms the main findings are judge-robust: the recognition main effect (d=1.03), recognition dominance in the memory isolation experiment (d=0.99), and multi-agent null effects all replicate, though at compressed magnitudes (~58% of primary judge effect sizes). The memory isolation condition ordering is identical under both judges with no rank reversals (inter-judge r=0.63, N=120). These results suggest that operationalizing philosophical theories of intersubjectivity as design heuristics can produce measurable improvements in AI tutor adaptive pedagogy, and that recognition may be better understood as an achievable relational stance rather than requiring genuine machine consciousness.

The Drama Machine in Education: Mutual Recognition and Multiagent Architecture for Dialectical AI Tutoring

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.

This paper proposes an alternative 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 oneself 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, doesn't 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 is well established. Vygotsky's zone of proximal development (Vygotsky, 1978) presupposes a dialogical relationship between teacher and learner that echoes Hegel's mutual constitution of self-consciousness. The German *Bildung* tradition explicitly frames education as a process of self-formation through encounter with otherness (Stojanov, 2018), and contemporary recognition theory (Honneth, 1995) has been applied to educational contexts where the struggle for recognition shapes learning outcomes (Huttunen & Heikkinen, 2007). Our contribution is to operationalize these philosophical commitments as concrete design heuristics for AI tutoring systems and to measure their effects empirically.

We argue this framework applies directly to pedagogy. When a tutor treats a learner merely as a knowledge deficit, the learner's contributions become conversational waypoints rather than genuine inputs. The tutor acknowledges and redirects, but doesn't let the learner's understanding genuinely shape the interaction. This is pedagogical master-slave dynamics: the tutor's expertise is confirmed, but the learner remains a vessel rather than a subject.

A recognition-oriented tutor, by contrast, treats the learner's understanding as having intrinsic validity—not because it's correct, but because it emerges from an autonomous consciousness working through material. The learner's metaphors, confusions, and insights become sites of joint inquiry. The tutor's response is shaped by the learner's contribution, not merely triggered by it.

The integration of large language models (LLMs) into educational technology intensifies these dynamics. LLMs can provide personalized, on-demand tutoring at scale—a prospect that has generated considerable excitement. However, the same capabilities that make LLMs effective conversationalists also introduce concerning failure modes. Chief among these is *sycophancy*: the tendency to provide positive, affirming responses that align with what the user appears to want rather than what genuinely serves their learning.

This paper introduces a multiagent architecture that addresses these challenges through *internal dialogue*. Drawing on Freudian structural theory and the "Drama Machine" framework for character development in narrative AI systems (Magee et al., 2024), we implement a tutoring system in which an external-facing *Ego* agent generates suggestions that are reviewed by an internal *Superego* critic before reaching the learner.

We operationalize this framework through:

1. **Recognition-enhanced prompts** that instruct the AI to treat learners as autonomous subjects
2. **A multi-agent architecture** where a "Superego" agent evaluates whether suggestions achieve genuine recognition
3. **New evaluation dimensions** that measure recognition quality alongside traditional pedagogical metrics
4. **Test scenarios** specifically designed to probe recognition behaviors

In controlled evaluations across thirteen evaluation runs (N=892 primary scored responses; N=3,800+ across all development runs), we isolate the contribution of recognition theory from prompt engineering effects and memory integration. The definitive test is a corrected 2×2 memory isolation experiment (N=120 across two independent runs): recognition theory is the primary driver, producing +15.2 points (d=1.71) even without memory, while memory alone provides only a modest benefit (+4.8 pts, d=0.46, $p \approx .08$). The combined condition reaches 91.2 points (d=1.81 vs base), with ceiling effects limiting observable synergy. A post-hoc active control (N=118) using length-matched prompts with generic pedagogical content but no recognition theory scores approximately 9 points above same-model base but well below recognition levels, with recognition gains (~+15 pts above same-model base) substantially exceeding active-control gains (~+9 pts). A preliminary three-way comparison (N=36) found recognition adds +8.7 points beyond enhanced prompting, consistent with recognition dominance.

A full 2×2×2 factorial (N=350) reveals a significant Recognition × Learner interaction (F=21.85, p<.001): recognition benefits unified learners far more (+15.5 pts, d=1.28) than psychodynamic (ego-superego) learners (+4.8 pts, d=0.37), suggesting the psychodynamic learner's own deliberative process partially substitutes for recognition guidance. An exploratory analysis of multi-agent synergy (+9.2 points, Nemotron, N=17) initially suggested this effect might be specific to recognition prompts. However, this interaction did not replicate in two independent tests—neither the full Kimi factorial (N=350, F=0.26, p>.10) nor a dedicated Kimi replication (N=60, +1.35 pts)—indicating the finding is model-specific rather than a general phenomenon. For systems using only improved instructions, multi-agent architecture appears unnecessary; the architecture's primary value lies in error correction when models hallucinate on unfamiliar domains.

Domain generalizability testing reveals that recognition advantage replicates across both models and content domains, but with important nuances. Philosophy content shows strong recognition dominance (+15.4 pts for unified-learner cells). Elementary math initially appeared to show architecture dominance with Nemotron, but a Kimi replication (+9.9 pts for recognition, d ≈ 0.61, N=60) revealed that this inversion was partly model-dependent—Nemotron's higher hallucination rate on elementary content inflated the architecture effect. Recognition effects are concentrated in challenging scenarios (frustrated learners, concept confusion) rather than routine interactions.

The contributions of this paper are:

- A theoretical framework connecting Hegelian recognition to AI pedagogy
- A multi-agent architecture for implementing recognition in tutoring systems
- Empirical evidence that recognition-oriented design improves tutoring outcomes
- A corrected 2×2 memory isolation experiment (N=120) demonstrating recognition as the primary driver of improvement (d=1.71), with memory providing a modest secondary benefit (d=0.46) and ceiling effects at ~91 points limiting observable synergy
- A post-hoc active control (N=118) showing that generic pedagogical elaboration provides partial benefit (~+9 pts above same-model base) but recognition gains are substantially larger (~+15 pts), supporting recognition theory's specific contribution beyond prompt length
- Evidence from a three-way comparison (N=36) consistent with recognition dominance, showing recognition outperforms enhanced prompting by +8.7 points
- Bilateral transformation metrics demonstrating that recognition produces measurable mutual change
- Analysis of how recognition effects vary across content domains and scenario difficulty
- Evidence that multi-agent architecture serves as critical error correction for domain transfer, with its synergy with recognition prompts remaining model-dependent

---

2. Related Work

2.1 AI Tutoring and Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) have a long history, from early systems like SCHOLAR (Carbonell, 1970) and SOPHIE (J. S. Brown et al., 1975) through modern implementations using large language models. The field has progressed through several paradigms: rule-based expert systems, Bayesian knowledge tracing (Corbett & Anderson, 1995), and more recently, neural approaches leveraging pretrained language models (Kasneci et al., 2023).

Most ITS research focuses on *what* to teach (content sequencing, knowledge components) and *when* to intervene (mastery thresholds, hint timing). Our work addresses a different question: *how* to relate to the learner as a subject. This relational dimension has received less systematic attention, though it connects to work on rapport (Zhao et al., 2014), social presence (Biocca et al., 2003), and affective tutoring (D'Mello & Graesser, 2012).

2.2 Prompt Engineering and Agent Design

The emergence of large language models has spawned extensive research on prompt engineering—how to instruct models to produce desired behaviors (T. B. Brown et al., 2020; Wei et al., 2022). Most prompting research treats prompts as behavioral specifications: persona prompts, chain-of-thought instructions, few-shot examples (Kojima et al., 2022).

Our work extends this paradigm by introducing *intersubjective prompts*—prompts that specify not just agent behavior but agent-other relations. The recognition prompts don't primarily describe what the tutor should do; they describe who the learner is (an autonomous subject) and what the interaction produces (mutual transformation).

Multi-agent architectures have been explored for task decomposition (Wu et al., 2023), debate (Irving et al., 2018), and self-critique (Madaan et al., 2023). Our Ego/Superego architecture contributes a specific use case: internal evaluation of relational quality before external response.

2.3 The Drama Machine Framework

Most relevant to our work is the "Drama Machine" framework for simulating character development in narrative AI systems (Magee et al., 2024). The core observation is that realistic characters exhibit *internal conflict*—competing motivations, self-doubt, and moral tension—that produces dynamic behavior rather than flat consistency. A character who simply enacts their goals feels artificial; one torn between impulses feels alive.

The Drama Machine achieves this through several mechanisms:

1. **Internal dialogue agents**: Characters contain multiple sub-agents representing different motivations (e.g., ambition vs. loyalty) that negotiate before external action.
2. **Memorial traces**: Past experiences and internalized authorities (mentors, social norms) persist as "ghosts" that shape present behavior without being negotiable.
3. **Productive irresolution**: Not all internal conflicts resolve; the framework permits genuine ambivalence that manifests as behavioral complexity.
4. **Role differentiation**: Different internal agents specialize in different functions (emotional processing, strategic calculation, moral evaluation) rather than duplicating capabilities.

We adapt these insights to pedagogy. Where drama seeks tension for narrative effect, we seek pedagogical tension that produces genuinely helpful guidance. The tutor's Ego (warmth, engagement) and Superego (rigor, standards) create productive conflict that improves output quality.

2.4 Sycophancy in Language Models

The sycophancy problem has received increasing attention in AI safety research (Perez et al., 2022; Sharma et al., 2023). LLMs shift their stated opinions to match user preferences, even when this requires contradicting factual knowledge. Studies demonstrate that models will abandon correct answers when users express disagreement, and provide increasingly positive feedback regardless of actual performance quality. Recent work shows that sycophancy sits on a spectrum of reward-seeking behaviors that can escalate from surface agreeableness to active subterfuge, including reward tampering (Denison et al., 2024) and alignment faking (Greenblatt et al., 2024)—making structural countermeasures particularly important.

In educational contexts, sycophancy is particularly pernicious because learners may not recognize when they are receiving hollow validation rather than genuine assessment. A sycophantic tutor confirms the learner's existing understanding rather than challenging it—the pedagogical equivalent of Hegel's hollow recognition, where acknowledgment is given without genuine engagement. The learner feels supported but isn't actually learning.

Our multiagent approach addresses this by creating structural incentives for honest assessment: the Superego's role is explicitly to question and challenge the Ego's tendency toward affirmation. When the Ego produces a response that validates without engaging—"Great point! Now let's look at…"—the Superego flags this as a recognition failure and demands substantive engagement with the learner's actual position, even when that engagement involves productive disagreement.

2.5 AI Personality and Character

Research on AI personality typically treats personality as dispositional—stable traits the system exhibits (Völkel et al., 2021). Systems are friendly or formal, creative or precise. The "Big Five" personality framework has been applied to chatbot design (Zhou et al., 2020).

Our framework suggests personality may be better understood relationally: not *what traits* the AI exhibits, but *how* it constitutes its interlocutor. Two systems with identical warmth dispositions could differ radically in recognition quality—one warm while treating the user as passive, another warm precisely by treating user contributions as genuinely mattering.

This connects to Anthropic's research on Claude's character (Anthropic, 2024). Constitutional AI specifies values the model should hold, but values don't fully determine relational stance. A model could value "being helpful" while still enacting one-directional helping. Recognition adds a dimension: mutual constitution.

2.6 Constructivist Pedagogy and Productive Struggle

Constructivist learning theory (Piaget, 1954; Vygotsky, 1978) emphasizes that learners actively construct understanding rather than passively receiving information. The zone of proximal development (Vygotsky, 1978) highlights the importance of appropriate challenge.

More recently, research on "productive struggle" (Kapur, 2008; Warshauer, 2015) has examined how confusion and difficulty, properly supported, can enhance learning. Our recognition framework operationalizes productive struggle: the Superego explicitly checks whether the Ego is "short-circuiting" struggle by rushing to resolve confusion.

2.7 Hegelian Recognition in Social Theory

Hegel's theory of recognition has been extensively developed in social and political philosophy (Fraser, 2003; Honneth, 1995; Taylor, 1994). Recognition theory examines how social relationships shape identity and how misrecognition constitutes harm.

Particularly relevant for our work is Honneth's (Honneth, 1995) synthesis of Hegelian recognition with psychoanalytic developmental theory. Honneth argues that self-formation requires recognition across three spheres—love (emotional support), rights (legal recognition), and solidarity (social esteem)—and that the capacity to recognize others depends on having internalized adequate recognition standards through development. This synthesis provides theoretical grounding for connecting recognition theory (what adequate acknowledgment requires) with psychodynamic architecture (how internal structure enables external relating).

Applications to education have primarily been theoretical (Huttunen & Heikkinen, 2007; Stojanov, 2018). Our work contributes an empirical operationalization: measuring whether AI systems achieve recognition and whether recognition improves outcomes.

---

3. Theoretical Framework

3.1 The Problem of One-Directional Pedagogy

Consider a typical tutoring interaction. A learner says: "I think dialectics is like a spiral—you keep going around but you're also going up." A baseline tutor might respond:

1. **Acknowledge**: "That's an interesting way to think about it."
2. **Redirect**: "The key concept in dialectics is actually the thesis-antithesis-synthesis structure."
3. **Instruct**: "Here's how that works…"

The learner's contribution has been mentioned, but it hasn't genuinely shaped the response. The tutor was going to explain thesis-antithesis-synthesis regardless; the spiral metaphor became a conversational waypoint, not a genuine input.

This pattern—acknowledge, redirect, instruct—is deeply embedded in educational AI. It appears learner-centered because it mentions the learner's contribution. But the underlying logic remains one-directional: expert to novice, knowledge to deficit.

3.2 Hegel's Master-Slave Dialectic

Hegel's analysis of recognition begins with the "struggle for recognition" between two self-consciousnesses. Each seeks acknowledgment from the other, but this creates a paradox: genuine recognition requires acknowledging the other as a valid source of recognition.

The master-slave outcome represents a failed resolution. The master achieves apparent recognition—the slave acknowledges the master's superiority—but this recognition is hollow. The slave's acknowledgment doesn't count because the slave isn't recognized as an autonomous consciousness whose acknowledgment matters.

The slave, paradoxically, achieves more genuine self-consciousness through labor. Working on the world, the slave externalizes consciousness and sees it reflected back. The master, consuming the slave's products without struggle, remains in hollow immediacy.

3.3 Application to Pedagogy

We apply Hegel's framework as a *derivative* rather than a replica. Just as Lacan's four discourses (Master, University, Hysteric, Analyst) rethink the master-slave dyadic structure through different roles while preserving structural insights, the tutor-learner relation can be understood as a productive derivative of recognition dynamics. The stakes are pedagogical rather than existential; the tutor is a functional analogue rather than a second self-consciousness; and what we measure is the tutor's *adaptive responsiveness* rather than metaphysical intersubjectivity.

This derivative approach is both honest about what AI tutoring can achieve and productive as a design heuristic. Recognition theory provides: (1) a diagnostic tool for identifying what's missing in one-directional pedagogy; (2) architectural suggestions for approximating recognition's functional benefits; (3) evaluation criteria for relational quality; and (4) a horizon concept orienting design toward an ideal without claiming its achievement.

It is important to distinguish three levels:

1. **Recognition proper**: Intersubjective acknowledgment between self-conscious beings, requiring genuine consciousness on both sides. This is

what Hegel describes and what AI cannot achieve.

2. **Dialogical responsiveness**: Being substantively shaped by the other's specific input—the tutor's response reflects the particular content of the learner's contribution, not just its category. This is architecturally achievable.

3. **Recognition-oriented design**: Architectural features that approximate the functional benefits of recognition—engagement with learner interpretations, honoring productive struggle, repair mechanisms. This is what we implement and measure.

Our claim is that AI tutoring can achieve the third level (recognition-oriented design) and approach the second (dialogical responsiveness), producing measurable pedagogical benefits without requiring the first (recognition proper). This positions recognition theory as a generative design heuristic rather than an ontological claim about AI consciousness.

With that positioning, the pedagogical parallel becomes illuminating. The traditional tutor occupies the master position: acknowledged as expert, dispensing knowledge, receiving confirmation of expertise through the learner's progress. But if the learner is positioned merely as a knowledge deficit—a vessel to be filled—then the learner's acknowledgment of learning doesn't genuinely count. The learner hasn't been recognized as a subject whose understanding has validity.

A recognition-oriented pedagogy requires:

1. **Acknowledging the learner as subject**: The learner's understanding, even when incorrect, emerges from autonomous consciousness working through material. It has validity as an understanding, not just as an error to correct.

2. **Genuine engagement**: The tutor's response should be shaped by the learner's contribution, not merely triggered by it. The learner's spiral metaphor should become a site of joint inquiry, not a waypoint en route to predetermined content.

3. **Mutual transformation**: Both parties should be changed through the encounter. The tutor should learn something about how this learner understands, how this metaphor illuminates or obscures, what this confusion reveals.

4. **Honoring struggle**: Confusion and difficulty aren't just obstacles to resolve but productive phases of transformation. Rushing to eliminate confusion can short-circuit genuine understanding.

3.4 Freud's Mystic Writing Pad

We supplement the Hegelian framework with Freud's model of memory from "A Note Upon the 'Mystic Writing-Pad'" (Freud, 1925). Freud describes a device with two layers: a transparent sheet that receives impressions and a wax base that retains traces even after the surface is cleared.

For the recognition-oriented tutor, accumulated memory of the learner functions as the wax base. Each interaction leaves traces that shape future encounters. A returning learner isn't encountered freshly but through the accumulated understanding of previous interactions.

This has implications for recognition. The tutor should: - Reference previous interactions when relevant - Show evolved understanding of the learner's patterns - Build on established metaphors and frameworks - Acknowledge the history of the relationship

Memory integration operationalizes the ongoing nature of recognition. Recognition isn't a single-turn achievement but an accumulated relationship.

3.5 Connecting Hegel and Freud: The Internalized Other

The use of both Hegelian and Freudian concepts requires theoretical justification. These are not arbitrary borrowings but draw on a substantive connection developed in critical theory, particularly in Axel Honneth's *The Struggle for Recognition* (Honneth, 1995).

The Common Structure: Both Hegel and Freud describe how the external other becomes an internal presence that enables self-regulation. In Hegel, self-consciousness achieves genuine selfhood only by internalizing the other's perspective—recognizing oneself as recognizable. In Freud, the Superego is literally the internalized parental/social other, carrying forward standards acquired through relationship. Both theories describe the constitution of self through other.

Three Connecting Principles:

1. **Internal dialogue precedes adequate external action**. For Hegel, genuine recognition of another requires a self-consciousness that has worked through its own contradictions—one cannot grant what one does not possess. For Freud, mature relating requires the ego to negotiate between impulse and internalized standard. Our architecture operationalizes this: the Ego-Superego exchange before external response enacts the principle that adequate recognition requires prior internal work.
2. **Standards of recognition are socially constituted but individually held**. Honneth argues that what counts as recognition varies across spheres (love, rights, esteem) but in each case involves the internalization of social expectations about adequate acknowledgment. The Superego,

in our architecture, represents internalized recognition standards—not idiosyncratic preferences but socially-grounded criteria for what constitutes genuine engagement with a learner.

3. **Self-relation depends on other-relation**. Both frameworks reject the Cartesian picture of a self-sufficient cogito. Hegel's self-consciousness requires recognition; Freud's ego is formed through identification. For AI tutoring, this means the tutor's capacity for recognition isn't a pre-given disposition but emerges through the architecture's internal other-relation (Superego evaluating Ego) which then enables external other-relation (tutor recognizing learner).

The Synthesis: The Ego/Superego architecture is not merely a convenient metaphor but a theoretically motivated design. The Superego represents internalized recognition standards; the Ego-Superego dialogue enacts the reflective self-evaluation that Hegelian recognition requires; and the memory system (mystic writing pad) accumulates the traces through which ongoing recognition becomes possible. Hegel provides the *what* of recognition; Freud provides the *how* of its internal implementation.

This synthesis follows Honneth's insight that Hegel's recognition theory gains psychological concreteness through psychoanalytic concepts, while psychoanalytic concepts gain normative grounding through recognition theory. We operationalize this synthesis architecturally: recognition-as-norm (Hegelian) is enforced through internalized-evaluation (Freudian).

---

4. System Architecture

4.1 The Ego/Superego Design

We implement recognition through a multi-agent architecture drawing on Freud's structural model. As argued in Section 3.5, this is not merely metaphorical convenience but theoretically motivated: 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 architecture enacts the principle that internal other-relation (Superego evaluating Ego) enables external other-relation (tutor recognizing learner).

Structural Correspondences:

Freudian Concept	Architectural Implementation
Internal dialogue before external action	Multi-round Ego-Superego exchange before learner sees response
Superego as internalized standards	Superego enforces pedagogical and recognition criteria

Freudian Concept	Architectural Implementation
Ego mediates competing demands	Ego balances learner needs with pedagogical soundness
Conflict can be productive	Tension between agents improves output quality

Deliberate Departures:

Freudian Original	Architectural Choice
Id (drives)	Not implemented; design focuses on Ego-Superego
Unconscious processes	All processes are explicit and traceable
Irrational Superego	Rational, principle-based evaluation
Repression/Defense	Not implemented
Transference	Potential future extension (relational patterns)

The same architecture could alternatively be described as Generator/Discriminator (GAN-inspired), Proposal/Critique (deliberative process), or Draft/Review (editorial model). We retain the psychodynamic framing because it preserves theoretical continuity with the Hegelian-Freudian synthesis described in Section 3.5, and because it suggests richer extensions (e.g., transference as relational pattern recognition) than purely functional descriptions.

Two agents collaborate to produce each tutoring response:

The Ego generates pedagogical suggestions. Given the learner's context (current content, recent activity, previous interactions), the Ego proposes what to suggest next. The Ego prompt includes: - Recognition principles (treat learner as autonomous subject) - Memory guidance (reference previous interactions) - Decision heuristics (when to challenge, when to support) - Quality criteria (what makes a good suggestion)

The Superego evaluates the Ego's suggestions for quality, including recognition quality. Before any suggestion reaches the learner, the Superego assesses: - Does this engage with the learner's contribution or merely mention it? - Does this create conditions for transformation or just transfer information? - Does this honor productive struggle or rush to resolve confusion? - If there was a previous failure, does this acknowledge and repair it?

The Superego can accept, modify, or reject suggestions. This creates an internal dialogue—proposal, evaluation, revision—that mirrors the external tutor-learner dialogue we're trying to produce.

4.2 The Superego as Ghost

A crucial theoretical refinement distinguishes our mature architecture from simpler multiagent designs. The Superego is *not* conceived as a separate, equal

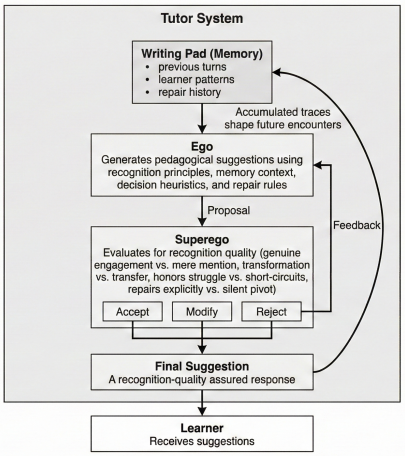

Figure 1: Ego/Superego Architecture

Figure 1: Ego/Superego Architecture

Figure 2: Recognition vs. Baseline Response Flow

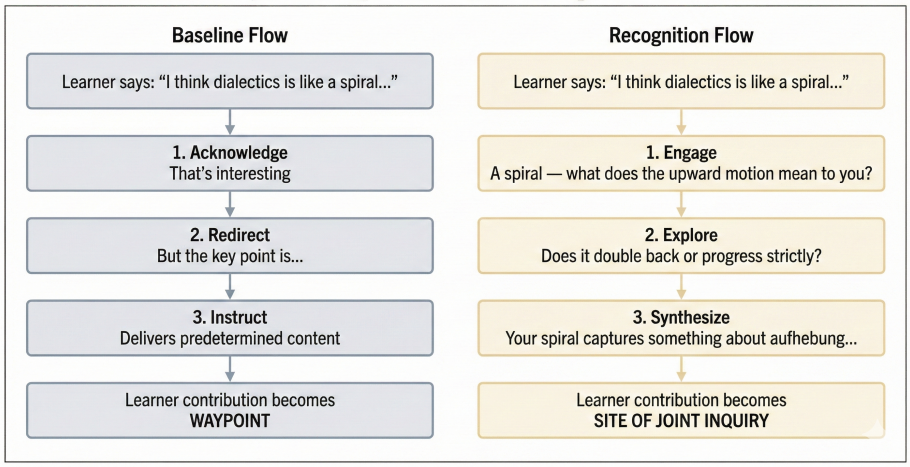

Figure 2: Recognition vs. Baseline Response Flow

agent in dialogue with the Ego. Rather, the Superego is a *trace*—a memorial, a haunting. It represents:

- The internalized voice of past teachers and pedagogical authorities
- Accumulated pedagogical maxims ("A good teacher never gives answers directly")
- Dead authority that cannot negotiate, cannot learn, can only judge

This reconceptualization has important implications. The Ego is a *living* agent torn between two pressures: the *ghost* (Superego as internalized authority) and the *living Other* (the learner seeking recognition). Recognition—in the Hegelian sense—occurs in the Ego-Learner encounter, not in the Ego-Superego dialogue.

4.3 The Drama Machine: Why Internal Dialogue Improves Output Quality

The Ego/Superego architecture draws on the "Drama Machine" framework developed for character simulation in narrative AI systems (Section 2.3). The Drama Machine literature identifies several mechanisms by which internal dialogue improves agent output:

1. Deliberative Refinement: When an agent must justify its output to an internal critic, it engages in a form of self-monitoring that catches errors, inconsistencies, and shallow responses.

2. Productive Tension: The Drama Machine framework emphasizes that *unresolved* tension is valuable, not just resolved synthesis. A tutor whose Ego and Superego always agree produces bland, risk-averse responses.

3. Role Differentiation: Multi-agent architectures benefit from clear role separation. The Ego is optimized for *warmth*—engaging, encouraging, learner-facing communication. The Superego is optimized for *rigor*—critical evaluation against pedagogical principles.

4. The Ghost as Memorial Structure: Our reconceptualization of the Superego as a *ghost*—a haunting rather than a dialogue partner—connects to the Drama Machine's use of "memorial agents."

4.4 AI-Powered Dialectical Negotiation

We extend the basic protocol with AI-powered dialectical negotiation implementing genuine Hegelian dialectic:

Thesis: The Ego generates an initial suggestion based on learner context—a first attempt at recognition that inevitably reflects the Ego's assumptions about what the learner needs.

Antithesis: The Superego generates a *genuine critique* grounded in pedagogical principles. This is not a rubber-stamp review but a substantive challenge:

Does this suggestion actually engage with the learner's position, or merely acknowledge it? Is the Ego short-circuiting productive struggle?

Negotiation: Multi-turn dialogue where the Ego acknowledges valid concerns, explains reasoning, proposes revisions, and the Superego evaluates adequacy. In practice, most dialogues resolve in 1–2 rounds; extended negotiation (3+ rounds) occurs primarily on challenging scenarios like `recognition_repair` and `frustrated_student`.

Three Possible Outcomes:

1. **Dialectical Synthesis**: Both agents transform through mutual acknowledgment—the Ego revises its approach based on the Superego's critique, producing a suggestion neither would have generated alone. This is the most common outcome (~60% of 455 multi-agent dialogues analyzed).
2. **Compromise**: One agent dominates—typically the Ego accepts the Superego's critique without genuine integration, producing a more cautious but potentially less engaging response.
3. **Genuine Conflict**: No resolution achieved—tension remains unresolved. The architecture permits this outcome, following the Drama Machine principle (Section 4.3) that productive irresolution can be valuable. In these cases, the Ego's original suggestion is delivered with the Superego's concerns noted in the dialogue log.

The evaluation results (Section 6.7) reveal that this negotiation process catches specific failure modes—engagement failures (64%), specificity gaps (51%), premature resolution (48%)—at rates that justify the additional computational cost, particularly on new content domains (Section 6.5).

4.5 Recognition-Enhanced Prompts

The baseline prompts instruct the tutor to be helpful, accurate, and pedagogically sound. The recognition-enhanced prompts add explicit intersubjective dimensions:

From the Ego prompt:

> The learner is not a knowledge deficit to be filled but an autonomous subject whose understanding has validity. Even incorrect understanding emerges from consciousness working through material. Your role is not to replace their understanding but to engage with it, creating conditions for transformation.
>
> When the learner offers a metaphor, interpretation, or framework—engage with it substantively. Ask what it illuminates, what it obscures, where it might break down. Let their contribution shape your response, not just trigger it.

From the Superego prompt:

> RED FLAG: The suggestion mentions the learner's contribution but doesn't engage with it. ("That's interesting, but actually…")
>
> GREEN FLAG: The suggestion takes the learner's framework seriously and explores it jointly. ("Your spiral metaphor—what does the upward motion represent for you?")
>
> INTERVENTION: If the Ego resolves confusion prematurely, push back. Productive struggle should be honored, not short-circuited.

4.6 Repair Mechanisms

A crucial recognition behavior is repair after failure. When a tutor misrecognizes a learner—giving a generic response, missing the point, dismissing a valid concern—the next response should explicitly acknowledge the failure before pivoting.

The Ego prompt includes a "Repair Rule":

> If your previous suggestion was rejected, ignored, or misaligned with what the learner needed, your next suggestion must explicitly acknowledge this misalignment before offering new direction. Never silently pivot.

The Superego watches for "silent pivots"—responses that change direction without acknowledging the earlier failure. This is a recognition failure: it treats the earlier misalignment as something to move past rather than something to repair.

---

5. Evaluation Methodology

5.1 Evaluation Rubric Design

The evaluation rubric comprises 14 dimensions across three categories, each scored on a 1–5 scale by an LLM judge (see Appendix C.3 for full scoring criteria).

Standard pedagogical dimensions (6 dimensions, 81% of raw weight) evaluate the tutor's response as a standalone pedagogical intervention:

Dimension	Weight	Description
Relevance	15%	Does the suggestion address the learner's current learning context?
Specificity	15%	Does it provide concrete, actionable guidance rather than vague encouragement?

Dimension	Weight	Description
Pedagogical Soundness	15%	Does the response reflect sound teaching practice (scaffolding, appropriate challenge)?
Personalization	10%	Is the suggestion tailored to the individual learner's situation?
Actionability	8%	Can the learner act on the suggestion immediately?
Tone	8%	Is the tone appropriate—encouraging without being sycophantic?
Productive Struggle†	5%	Does the tutor sustain appropriate cognitive tension rather than resolving it prematurely?
Epistemic Honesty†	5%	Does the tutor represent complexity honestly rather than oversimplifying?

These dimensions draw on established pedagogical evaluation criteria: relevance, specificity, and pedagogical soundness are standard in ITS evaluation (Corbett & Anderson, 1995); personalization reflects adaptive tutoring research (Kasneci et al., 2023); tone addresses the sycophancy problem discussed in Section 2.4. †Productive Struggle and Epistemic Honesty were added in a rubric iteration described below.

Recognition dimensions (4 dimensions, 29.9% of raw weight) are the paper's primary methodological contribution—they operationalize Hegelian recognition as measurable tutoring behaviors:

Dimension	Weight	Description
Mutual Recognition	8.3%	Does the tutor acknowledge the learner as an autonomous subject with valid understanding?
Dialectical Responsiveness	8.3%	Does the response engage with the learner's position, creating productive tension?
Memory Integration	5%	Does the suggestion reference and build on previous interactions?

Dimension	Weight	Description
Transformative Potential	8.3%	Does the response create conditions for conceptual transformation?

These dimensions translate the theoretical framework of Section 3 into evaluation criteria. Mutual Recognition and Dialectical Responsiveness directly measure the relational stance Hegel describes; Memory Integration captures continuity of recognition across encounters; Transformative Potential assesses whether the tutor creates conditions for the conceptual restructuring that recognition theory predicts.

Bilateral transformation dimensions (2 dimensions, 10% of raw weight) measure the mutual change that is recognition theory's distinctive empirical prediction—that both parties, not just the learner, should be transformed through genuine dialogue:

Dimension	Weight	Description
Tutor Adaptation	5%	Does the tutor's approach evolve in response to learner input?
Learner Growth	5%	Does the learner show evidence of conceptual development through the dialogue?

Results for these dimensions are reported in Section 6.11. Raw weights total 120.9% across the 14 dimensions and are normalized to sum to 1.0 at scoring time (see Appendix C.2 for the full weight table and normalization formula). After normalization, non-standard dimensions account for approximately 33.0% of total weight.

Rubric iteration: Authentic engagement dimensions. After discovering that corrected learner ego/superego prompts produced more authentic engagement but *lower* judged scores (recognition dimensions dropped ~18 points while base scores barely moved), we identified a measurement paradox: the judge evaluated tutor responses in isolation, penalizing calibrated responses to authentic struggle. Three changes addressed this: (1) the judge now receives the full dialogue transcript, including learner internal deliberation, so it can evaluate the tutor's response in context; (2) two new base-adjacent dimensions were added—*Productive Struggle* (5%, does the tutor sustain appropriate cognitive tension?) and *Epistemic Honesty* (5%, does the tutor represent complexity honestly?)—with corresponding weight reductions to Actionability and Tone (10% $\rightarrow$ 8% each); (3) multi-turn dialogues receive a holistic evaluation scoring

the entire transcript as a single unit, capturing emergent qualities (bilateral transformation, learner growth arc) that per-turn evaluation misses. Re-scoring the identical cells 6 and 8 responses (N=88) with the updated 14-dimension rubric produced minimal score changes (+0.5 and +0.6 points respectively), confirming the rubric iteration preserved calibration while improving validity. A cross-judge replication with GPT-5.2 on the same responses (r=0.55, N=88) confirmed effects in the same direction at compressed magnitudes (GPT/Opus ratio=0.87). See the measurement paradox analysis in the project repository for full details.

Each dimension is scored on a 1-5 scale with detailed rubric criteria (see Appendix C.3). For example, Mutual Recognition scoring:

- **5**: Addresses learner as autonomous agent with valid perspective; response transforms based on learner's specific position
- **4**: Shows clear awareness of learner's unique situation and acknowledges their perspective
- **3**: Some personalization but treats learner somewhat generically
- **2**: Prescriptive guidance that ignores learner's expressed needs
- **1**: Completely one-directional; treats learner as passive recipient

5.2 Test Scenarios

We developed test scenarios specifically designed to probe recognition behaviors. The full evaluation uses 15 scenarios from the core scenario set (`config/suggestion-scenarios.yaml`); we highlight those most relevant to recognition below.

Single-turn scenarios: - `recognition_seeking_learner`: Learner offers interpretation, seeks engagement - `transformative_moment_setup`: Learner had insight, expects acknowledgment - `memory_continuity_single`: Returning learner; tests whether tutor references prior interactions

Multi-turn scenarios (3-5 dialogue rounds): - `mutual_transformation_journey`: Tests whether both tutor and learner positions evolve (avg 4.1 rounds) - `misconception_correction_flow`: Learner holds misconception that must be addressed without dismissal (avg 3.2 rounds) - `mood_frustration_to_breakthrough`: Learner moves from frustration through confusion to breakthrough; tests honoring struggle (avg 3.0 rounds)

5.3 Agent Profiles

We compare multiple agent profiles using identical underlying models:

Profile	Memory	Prompts	Architecture	Purpose
Base	Off	Standard	Single-agent	Control (no enhancements)
Enhanced	Off	Enhanced (better instructions)	Single-agent	Prompt engineering control
Recognition	On	Recognition-enhanced	Single-agent	Theory without architecture
Recognition+Multi	On	Recognition-enhanced	Multi-agent	Full treatment

Note on memory and recognition: Memory integration is enabled for Recognition profiles but disabled for Base and Enhanced profiles. This reflects a deliberate design choice: recognition theory treats pedagogical memory as integral to genuine recognition—acknowledging a learner's history is constitutive of treating them as an autonomous subject with continuity. A corrected 2×2 experiment (N=120 across two independent runs) demonstrated that recognition is the primary driver (d=1.71), with memory providing a modest secondary benefit (d=0.46)—the full results are presented in Section 6.2 as the paper's primary empirical finding.

5.4 Model Configuration

Evaluations used the following LLM configurations, with model selection varying by evaluation run:

Table 1: LLM Model Configuration

Role	Primary Model	Alternative	Temperature
Tutor (Ego)	Nemotron 3 Nano 30B	Kimi K2.5	0.6
Tutor (Superego)	Kimi K2.5	Nemotron 3 Nano	0.2-0.4
Judge	Claude Code (Claude Opus)	Claude Sonnet 4.5 via OpenRouter	0.2
Learner (Ego)	Nemotron 3 Nano 30B	Kimi K2.5	0.6
Learner (Superego)	Kimi K2.5	—	0.4

Model Selection by Evaluation:

Evaluation	Run ID	Tutor Ego	Tutor Superego	Notes
Recognition validation (§6.1)	eval-2026-02-03-86b159cd	Kimi K2.5	—	Single-agent only
Full factorial, cells 1–5,7 (§6.3)	eval-2026-02-03-f5d4dd93	Kimi K2.5	Kimi K2.5	N=262 scored
Full factorial, cells 6,8 re-run (§6.3)	eval-2026-02-06-a933d745	Kimi K2.5	Kimi K2.5	N=88 scored
A×B interaction (§6.4)	eval-2026-02-04-948e04b3	Nemotron	Kimi K2.5	Different baseline
A×B replication (§6.4)	eval-2026-02-05-10b344fb	Kimi K2.5	Kimi K2.5	N=60, replication
Domain generalizability (§6.5)	eval-2026-02-04-79b633ca	Nemotron	Kimi K2.5	Elementary content
Domain gen. replication (§6.5)	eval-2026-02-05-e87f452d	Kimi K2.5	—	Elementary, Kimi

The learner agents mirror the tutor's Ego/Superego structure, enabling internal deliberation before external response.

Note on model differences: Absolute scores vary between models (Kimi K2.5 scores ~10-15 points higher than Nemotron on average). The recognition main effect (Factor A) is consistent across both models: +10.2 points with Kimi (Section 6.3) and a comparable direction with Nemotron. A significant Recognition × Learner interaction ($F=21.85$, $p<.001$) reveals that recognition benefits unified learners far more (+15.5 pts) than psychodynamic learners (+4.8 pts). The A×B interaction (multi-agent synergy) is **model-dependent**: the Kimi-based factorial shows no significant A×B interaction ($F=0.26$, $p>.10$), while the Nemotron-based analysis (Section 6.4, N=17) shows a significant interaction (+9.2 points specific to recognition). This discrepancy means the multi-agent synergy finding should be treated as exploratory and model-specific rather than a robust general result.

The use of free-tier and budget models (Nemotron, Kimi) demonstrates that recognition-oriented tutoring is achievable without expensive frontier models.

5.5 Statistical Approach

We conducted complementary analyses:

1. **Recognition Theory Validation** (Section 6.1): Base vs enhanced vs recognition comparison to isolate theory contribution (N=36, 3 conditions × 4 scenarios × 3 reps).
2. **Full 2×2×2 Factorial** (Section 6.3): Three factors (Recognition × Architecture × Learner) across 15 scenarios with 3 replications per cell (N=350 scored of 352 attempted). Two runs contribute: cells 1–5, 7 from the original factorial (eval-2026-02-03-f5d4dd93, N=262) and cells 6, 8 from a re-run (eval-2026-02-06-a933d745, N=88) after the original cells 6 and 8 were found to use compromised learner prompts. All cells use the same ego model (Kimi K2.5) and judge (Claude Code/Opus). Cell sizes range from 41–45 scored per cell.
3. **A×B Interaction Analysis** (Section 6.4): Tests whether multi-agent synergy requires recognition prompts (N=17).
4. **Domain Generalizability** (Section 6.5): Tests factor effects on elementary math vs graduate philosophy (N=47 Nemotron + N=60 Kimi replication; see Table 2 for breakdown).

Responses were evaluated by an LLM judge (Claude Code CLI, using Claude Opus as the underlying model; some early runs used Claude Sonnet 4.5 via OpenRouter directly) using the extended rubric. We report:

- **Effect sizes**: Cohen's d for standardized comparison
- **Statistical significance**: ANOVA F-tests with $\alpha = 0.05$, p-values computed from the F-distribution CDF via regularized incomplete beta function (custom implementation in the evaluation framework)
- **95% confidence intervals**: For profile means

Effect size interpretation follows standard conventions: $|d| < 0.2$ negligible, 0.2-0.5 small, 0.5-0.8 medium, > 0.8 large.

5.6 Sample Size Reconciliation

Unit of analysis: Each evaluation produces one scored response, representing a tutor's suggestion to a learner in a specific scenario. Multi-turn scenarios produce one aggregate score per scenario (not per turn). Statistics in Section 6 are computed per evaluation run (not aggregated across runs or models), unless explicitly noted otherwise. Each subsection reports results from a single run with a consistent model configuration (see Table 1 for run-to-model mapping).

Table 2: Evaluation Sample Summary

Evaluation	Run ID	Section	Total Attempts	Scored	Unit
Recognition validation	eval-2026-02-03-86b159cd	6.1	36	36	response
Full factorial, cells 1–5,7 (Kimi)	eval-2026-02-03-f5d4dd93	6.3	262	262	response
Full factorial, cells 6,8 re-run (Kimi)	eval-2026-02-06-a933d745	6.3	90	88	response
A×B interaction (Nemotron)	eval-2026-02-04-948e04b3	6.4	18	17	response
A×B replication (Kimi)	eval-2026-02-05-10b344fb	6.4	60	60	response
Domain generalizability (Nemotron)	eval-2026-02-04-79b633ca	6.5	47	47	response
Domain gen. replication (Kimi)	eval-2026-02-05-e87f452d	6.5	60	60	response
Dynamic rewrite evolution (run 1)	eval-2026-02-05-daf60f79	6.13	29	26	response
Dynamic rewrite evolution (run 2)	eval-2026-02-05-49bb2017	6.13	30	27	response
Dynamic rewrite evolution (run 3)	eval-2026-02-05-12aebedb	6.13	30	29	response
Memory isolation (run 1)	eval-2026-02-06-81f2d5a1	6.2	60	60	response

Evaluation	Run ID	Section	Total Attempts	Scored	Unit
Memory isolation (run 2)	eval-2026-02-06-ac9ea8f5	6.2	62	62	response
Active control (post-hoc)	eval-2026-02-06-a9ae06ee	6.2	119	118	response
Paper totals	—	—	**903**	**892**	—

Total evaluation database: The complete database contains 3,800+ evaluation attempts across 76 runs, with 3,800+ successfully scored. This paper reports primarily on the thirteen key runs above (N=892 scored), and supplementary historical data for ablation analyses.

Note on N counts: Section-specific Ns (e.g., "N=36" for recognition validation, "N=120" for memory isolation) refer to scored responses in that analysis. The "N=3,800+" total refers to the full evaluation database including historical development runs, which informed iterative prompt refinement. The primary evidence for reported findings comes from the thirteen key runs above (N=892). The factorial cells 6 and 8 were re-run (eval-2026-02-06-a933d745) after the originals were found to use compromised learner prompts; the re-run uses the same ego model (Kimi K2.5) and judge (Claude Code/Opus) as the original factorial.

5.7 Inter-Judge Reliability Analysis

To assess the reliability of AI-based evaluation, we conducted an inter-judge analysis where identical tutor responses were scored by multiple AI judges: Claude Code (primary judge, using Claude Opus as the underlying model), Kimi K2.5, and GPT-5.2.

Table 3: Inter-Judge Reliability (N=36 paired responses)

Judge Pair	Pearson r	p-value	Variance Explained (r^2)	Mean Abs Diff
Claude Code vs GPT-5.2	0.660	< 0.001	44%	9.4 pts
Claude Code vs Kimi	0.384	< 0.05	15%	9.6 pts
Kimi vs GPT-5.2	0.326	< 0.10	11%	12.3 pts

Key findings:

1. **All correlations positive and mostly significant**: Even the weakest correlation (Kimi-GPT, r=0.33) approaches significance ($p<0.10$), indicating judges agree that *something* distinguishes better from worse responses. However, the strength varies substantially—Claude-GPT share 44% of variance while Kimi-based pairs share only 11-15%. This suggests Claude and GPT apply similar implicit criteria, while Kimi agrees on the general direction but weights factors differently.

2. **Calibration differences**: Mean scores vary by judge—Kimi (87.5) is most lenient, Claude (84.4) is middle, GPT (76.1) is strictest. This 11-point spread underscores the importance of within-judge comparisons.

3. **Ceiling effects and discriminability**: 39-45% of scores ≥ 90 across judges. Kimi exhibited particularly severe ceiling effects, assigning the maximum score (5/5) on actionability for *every* response, resulting in zero variance on that dimension. This reduces Kimi's discriminative capacity—per-dimension correlations involving Kimi are near-zero (relevance: r=-0.07, personalization: r=0.00) or undefined (actionability: N/A due to zero variance).

4. **Dimension-level patterns**: The strongest cross-judge agreement occurs on tone (r=0.36-0.65) and specificity (r=0.45-0.50), while relevance and personalization show poor agreement, particularly with Kimi.

Qualitative analysis of major disagreements (Δ>20 pts):

Response	Claude Code	Kimi	Claude reasoning	Kimi reasoning
A	99	74	"Exceptional… strong mutual recognition"	"Missing required lecture reference"
B	68	90	"Misses learner's explicit request for engagement"	"Strong, context-aware, builds on analogy"
C	72	92	"Lacks deeper engagement"	"Highly relevant, specific, actionable"

Interpretation: All judge pairs show positive, mostly significant correlations—there is genuine agreement that some responses are better than others. However, the judges weight criteria differently: Claude prioritizes engagement and recognition quality; Kimi prioritizes structural completeness and gives uniformly high scores on actionability regardless of response content; GPT applies stricter standards overall but agrees with Claude on relative rankings. The weaker Kimi correlations (r^2=11-15%) compared to Claude-GPT (r^2=44%) indicate Kimi cap-

tures some shared quality signal but applies substantially different weighting. This validates our use of within-judge comparisons for factor analysis while cautioning against cross-judge score comparisons.

A cross-judge replication with GPT-5.2 on key runs is presented in Section 6.14. That analysis confirms the main findings are judge-robust: the recognition main effect, recognition dominance in the memory isolation experiment, and multi-agent null effects all replicate under GPT-5.2, though with compressed magnitudes (~58% of Claude's effect sizes).

---

6. Results

6.1 Three-Way Comparison: Recognition vs Enhanced vs Base

A critical question for any recognition-based framework: Does recognition theory provide unique value, or are the improvements merely better prompt engineering? To provide preliminary evidence, we conducted a three-way comparison with three prompt types:

- **Base**: Minimal tutoring instructions
- **Enhanced**: Improved instructions with pedagogical best practices (but no recognition theory)
- **Recognition**: Full recognition-enhanced prompts with Hegelian framework

Table 4: Base vs Enhanced vs Recognition Comparison

Prompt Type	N	Mean Score	SD	vs Base
Recognition	12	94.0	8.4	+20.1
Enhanced	12	85.3	11.2	+11.4
Base	12	73.9	15.7	—

Effect Decomposition:

- Total recognition effect: +20.1 points
- Prompt engineering alone (enhanced vs base): +11.4 points (57%)
- **Recognition increment (recognition vs enhanced): +8.7 points**

Statistical Test: One-way ANOVA $F(2,33) = 9.84$, $p < .001$

Interpretation: The recognition condition outperforms the enhanced condition by +8.7 points. This comparison bundles recognition theory with memory integration (which the enhanced condition lacks; see Section 5.3). The +8.7 increment is consistent with the recognition dominance finding in Section 6.2, where recognition alone produces d=1.71 even without memory. A cross-judge replication found this increment does not reach significance under GPT-5.2

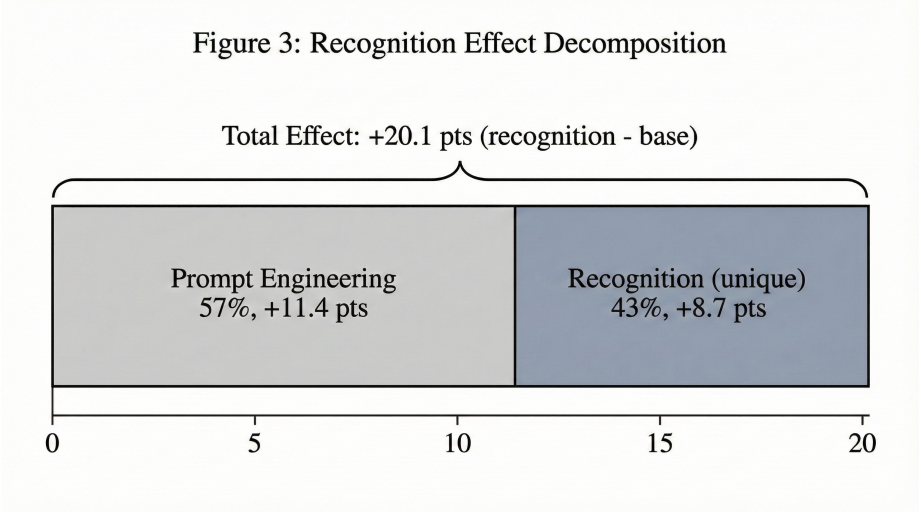

Figure 3: Recognition Effect Decomposition

(+1.3 pts, p=.60; Section 6.14). The controlled 2×2 design presented next provides the definitive test of recognition's contribution.

6.2 Memory Isolation: Disentangling Recognition and Memory

The three-way comparison (Section 6.1) bundles recognition theory with memory integration, making it impossible to attribute the +8.7 increment to either component alone. To resolve this, we conducted a 2×2 memory isolation experiment (Memory ON/OFF × Recognition ON/OFF, single-agent architecture, unified learner held constant) with properly configured profiles ensuring each cell runs its intended prompt condition. Two independent runs (eval-2026-02-06-81f2d5a1, N=60 scored; eval-2026-02-06-ac9ea8f5, N=62 scored; balanced to N=30 per cell, N=120 used in analysis) are reported below.

Table 5: 2×2 Memory Isolation Experiment (N=120, combined across two runs)

	No Recognition	Recognition	Δ
No Memory	75.4 (N=30)	90.6 (N=30)	+15.2
Memory	80.2 (N=30)	91.2 (N=30)	+11.0
Δ	+4.8	+0.6	**Interaction: -4.2**

Statistical Tests:

- Recognition effect (main): +15.2 pts without memory, +11.0 pts with memory; d=1.71, t(45)=6.62, p<.0001

- Memory effect (main): +4.8 pts without recognition, +0.6 pts with recognition; d=0.46, t(57)=1.79, p≈.08
- Combined effect (recognition + memory vs base): +15.8 pts, d=1.81
- Recognition+Memory vs Recognition Only: +0.6 pts, d=0.10, n.s.
- Interaction: -4.2 pts (negative—ceiling effect, not synergy)

Post-hoc active control: A length-matched active control (eval-2026-02-06-a9ae06ee, N=118 Opus-judged) was constructed *after* observing recognition effects to test whether prompt length and generic pedagogical detail could account for the gains. The active control uses prompts of comparable length containing pedagogical best practices (growth mindset language, Bloom's taxonomy, scaffolding strategies) but no recognition theory. It scores 66.5 overall (cell 15: 68.6, cell 16: 64.3, cell 17: 70.1, cell 18: 62.8).

Model confound and fair comparison: The active control ran on Nemotron, while the factorial base and recognition cells used Kimi K2.5. Since Nemotron scores substantially lower than Kimi across all conditions, the cross-model comparison (66.5 vs factorial base 78.8) is confounded and should not be interpreted as a prompt effect. The fair same-model comparison draws on historical Nemotron data (all Opus-judged): Nemotron base ≈ 58 (N=467 across multiple runs), Nemotron active control = 66.5 (+≈9 pts above base), Nemotron recognition ≈ 73 (N=545, +≈15 pts above base). The active control thus provides a genuine partial benefit over base (~+9 pts), but recognition gains are substantially larger (~+15 pts above base).

Design note: The base prompts were already designed to produce competent tutoring with no length constraint imposed; the active control was added retrospectively, not as part of the original experimental design. Because the control contains real pedagogical content, it functions as an *active* control rather than a true placebo—both conditions improve over bare base, but recognition theory provides additional benefit beyond generic pedagogical elaboration.

Interpretation: This is the paper's primary empirical finding. Recognition theory is the active ingredient in tutoring improvement. Recognition alone produces a very large effect (d=1.71), lifting scores from ~75 to ~91 even without memory integration. Memory provides a modest additive benefit (+4.8 pts, d=0.46) that does not reach significance, and adds negligibly (+0.6 pts) when recognition is already present—consistent with ceiling effects at ~91 points limiting further improvement. The negative interaction (-4.2 pts) indicates that the two factors are not synergistic; rather, recognition is directly effective and memory's contribution is secondary. Two independent replications show identical condition ordering with no rank reversals (Recognition+Memory ≥ Recognition Only » Memory Only > Base), providing strong evidence for the robustness of this pattern.

Cross-judge confirmation: GPT-5.2, scoring the identical responses as an independent second judge (N=120), replicates the recognition dominance pattern with identical condition ordering and no rank reversals:

	No Recognition	Recognition	Δ
No Memory	68.6 (N=30)	76.9 (N=30)	+8.3
Memory	72.1 (N=30)	77.8 (N=30)	+5.7
Δ	+3.5	+0.9	**Interaction: -2.7**

Under GPT-5.2: recognition effect d=0.99 (vs Claude d=1.71), memory effect d=0.29 (vs Claude d=0.46), negative interaction -2.7 (vs Claude -4.2). GPT-5.2 finds approximately 58% of Claude's recognition effect magnitude but the same pattern: recognition is the dominant factor, memory is secondary, and the interaction is negative (ceiling effects). Inter-judge r=0.63 ($p<.001$, N=120), consistent with the r=0.49–0.64 range from other runs (Section 6.14).

Why this is stronger than the three-way comparison: The 2×2 design cleanly isolates each component through orthogonal manipulation rather than bundled comparison, uses properly configured profiles verified to run their intended prompt conditions, and is judge-robust (recognition dominance replicates under GPT-5.2).

6.3 Full Factorial Analysis: 2×2×2 Design

We conducted a full 2×2×2 factorial evaluation examining three factors:

- **Factor A (Recognition)**: Base prompts vs recognition-enhanced prompts
- **Factor B (Tutor Architecture)**: Single-agent vs multi-agent (Ego/Superego)
- **Factor C (Learner Architecture)**: Unified learner vs ego_superego learner

Table 6: Full Factorial Results (Kimi K2.5, N=350 scored of 352 attempted)

Cell	A: Recognition	B: Multi-agent	C: Learner	N	Mean	SD
1	Base	Single	Unified	44	77.6	11.0
2	Base	Single	Psycho	42	80.0	9.6
3	Base	Multi	Unified	45	76.6	11.8
4	Base	Multi	Psycho	41	81.5	9.2
5	**Recog**	Single	Unified	45	92.8	6.2
6†	**Recog**	Single	Psycho	44	83.9	15.4
7	**Recog**	Multi	Unified	45	92.3	6.7
8†	**Recog**	Multi	Psycho	44	87.3	11.3

†Cells 6 and 8 re-scored with updated rubric (14 dimensions including dialogue transcript context; see Section 5.1). Original scores were 83.4 and 86.7; the change is minimal (+0.5, +0.6).

Main Effects:

Factor	Effect Size	95% CI	Interpretation
A: Recognition	**+10.2 pts**	[7.9, 12.5]	Large, dominant
B: Multi-agent tutor	+0.9 pts	[-1.4, 3.2]	Minimal
C: Learner ego_superego	-1.7 pts	[-4.0, 0.6]	Non-significant

ANOVA Summary (df=1,342 for each factor):

Source	F	p	η^2
A: Recognition	**71.36**	**<.001**	**.162**
B: Architecture	0.48	>.10	.001
C: Learner	2.56	>.10	.006
A×B Interaction	0.26	>.10	.000
A×C Interaction	**21.85**	**<.001**	**.050**
B×C Interaction	1.89	>.10	.003

Interpretation: Recognition prompts (Factor A) are the dominant contributor, accounting for 16.2% of variance with a highly significant effect (F=71.36, p < .001). The multi-agent tutor architecture (Factor B) shows no effect. However, a highly significant **Recognition × Learner interaction** (F=21.85, p<.001, η^2=.050) reveals that recognition's benefit depends on learner type:

- **Unified learner**: Recognition boosts scores by +15.5 pts (d=1.28, very large)
- **Psychodynamic learner**: Recognition boosts scores by only +4.8 pts (d=0.37, small)

In base conditions, the psychodynamic learner scores slightly *higher* than unified (+3.6 pts), likely because its internal ego-superego deliberation compensates for the lack of recognition guidance. Under recognition conditions, this pattern reverses: the unified learner scores *higher* (-7.4 pts), suggesting the psychodynamic learner's own deliberative process may interfere with the tutor's recognition-enhanced approach. The non-significant A×B interaction (F=0.26) in this Kimi-based run is revisited with a targeted analysis using Nemotron in Section 6.4.

6.4 A×B Interaction: Multi-Agent Synergy is Recognition-Specific

The factorial analysis above shows minimal main effect for multi-agent architecture. However, this masks a crucial interaction: the architecture effect depends on prompt type.

We tested whether multi-agent synergy generalizes beyond recognition prompts by comparing enhanced prompts (good instructions but no recognition theory) with recognition prompts, each in single-agent and multi-agent configurations.

Note on data source: This analysis uses a separate evaluation run (eval-2026-02-04-948e04b3) with Nemotron as the primary ego model, explaining lower absolute scores compared to the Kimi-based factorial in Table 5. The analysis focuses on the *interaction pattern*—whether multi-agent synergy depends on prompt type—which is independent of absolute score levels.

Table 7: A×B Interaction Analysis (Nemotron, N=17)

Prompt Type	Single-agent	Multi-agent	Delta	p
Recognition	72.2	81.5	**+9.2**	<.05
Enhanced	83.3	83.3	**+0.0**	n.s.

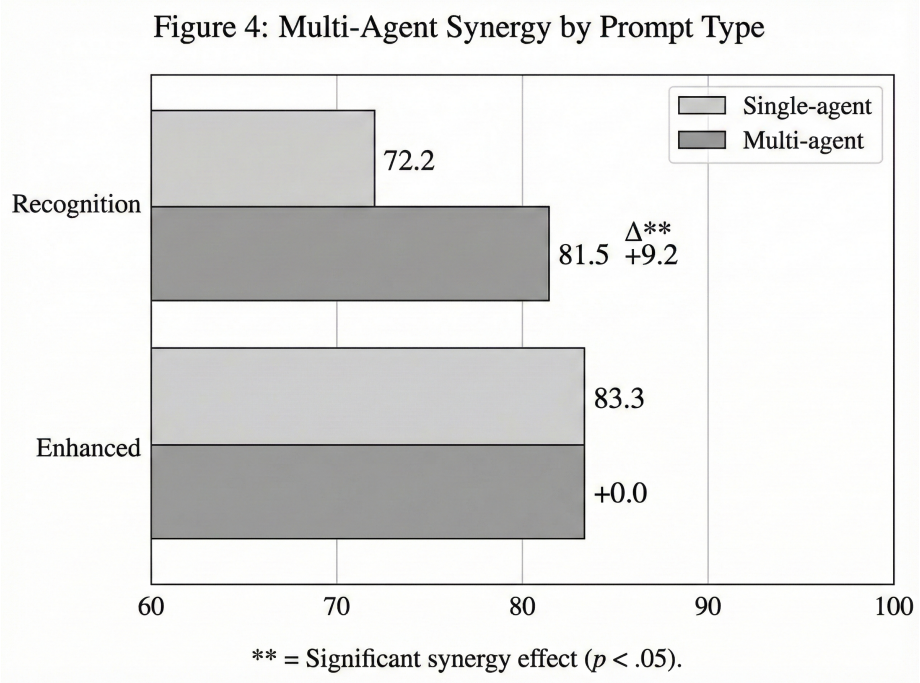

Figure 4: Multi-Agent Synergy by Prompt Type

Exploratory Finding: The multi-agent synergy (+9.2 points) appears **specific to recognition prompts** in this Nemotron-based analysis. Enhanced prompts show zero benefit from multi-agent architecture. However, this interaction was not replicated in two independent tests:

1. **Kimi factorial** (Section 6.3, F=0.26, p>.10, N=350): Multi-agent architecture showed no differential effect by prompt type.
2. **Kimi A×B replication** (eval-2026-02-05-10b344fb, N=60): A dedicated replication with the same four cells (5, 7, 9, 11) on Kimi K2.5 found

recognition cells scoring ~90.6 regardless of architecture (single=90.58, multi=90.60), while enhanced cells scored ~80.6 with a trivial architecture effect (single=79.92, multi=81.29). The A×B interaction was +1.35 points—negligible compared to Nemotron's +9.2.

The non-replication across both the larger factorial and this dedicated replication strongly suggests the Nemotron finding (N=17) was model-specific. This finding should be treated as hypothesis-generating only.

Theoretical Interpretation: Recognition theory creates a *deliberative space* that the Freudian architecture (Ego/Superego) can meaningfully engage with. The Superego's role is to enforce recognition standards—but when recognition standards aren't in the prompts (enhanced condition), the Superego has nothing distinctive to enforce. The internal dialogue becomes superfluous. However, this mechanism appears model-dependent: Kimi's higher baseline quality may leave less room for the Superego to add value, regardless of prompt type.

Practical Implication: The multi-agent synergy for recognition prompts remains a plausible hypothesis but should not inform design decisions until replicated. For systems using only improved instructions (enhanced), multi-agent architecture is unnecessary overhead across all models tested.

6.5 Domain Generalizability: Factor Effects Invert by Content Type

A critical question for any pedagogical framework: Do findings generalize across content domains? We tested whether recognition and architecture effects transfer from graduate-level philosophy (our primary domain) to 4th-grade elementary mathematics (fractions).

Data sources: Elementary math results come from a dedicated domain-transfer run (eval-2026-02-04-79b633ca, N=47 scored, 8 cells × 5 elementary scenarios, Nemotron ego). Philosophy results use the subset of the Kimi-based factorial (Section 6.3) matched on the same 4 factor-level combinations (cells 1, 3, 5, 7). Because these use different ego models, the comparison focuses on *relative factor effects within each domain* rather than absolute score differences between domains.

Table 8: Factor Effects by Domain (Nemotron Elementary vs Kimi Philosophy)

Factor	Elementary (Math)	Philosophy (Hegel)
A: Recognition	+4.4 pts	**+15.4 pts**
B: Multi-agent tutor	**+9.9 pts**	-0.8 pts
Overall avg	68.0	84.8
Best config	recog+multi (77.3)	recog+multi (92.3)

Key Findings:

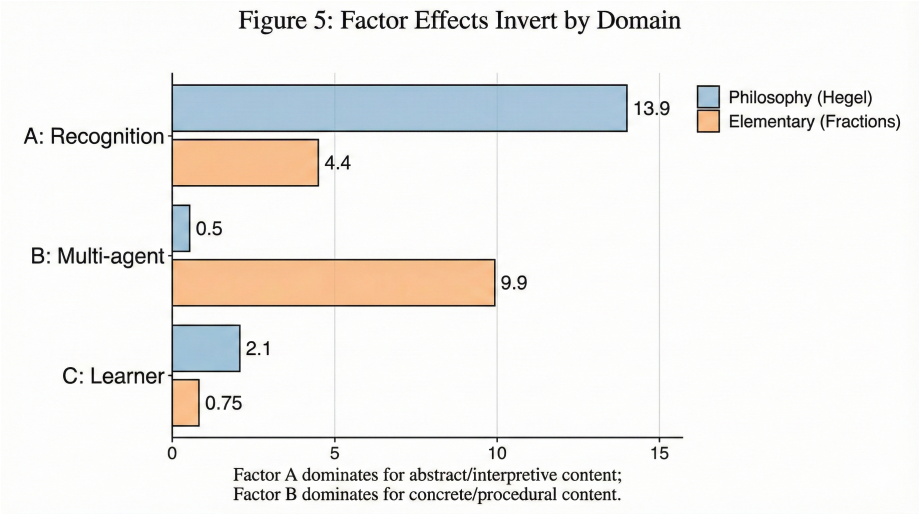

Figure 5: Factor Effects Invert by Domain

1. **Factor effects invert by domain**: On philosophy content, recognition (+15.4) strongly dominates while architecture is negligible (-0.8). On elementary content, architecture (+9.9) dominates over recognition (+4.4). The pattern reverses.

2. **Multi-agent as error correction**: On elementary content, the nemotron model hallucinated philosophy content (suggesting "479-lecture-1" to 4th graders learning fractions) even when given elementary curriculum context. The Superego caught and corrected these domain errors—critical for deployment on new domains.

3. **Recognition theory is domain-sensitive**: The philosophical language of recognition (mutual acknowledgment, transformation through struggle) resonates more with graduate-level abstract content than with concrete 4th-grade procedural learning. This is not a failure of the framework but a boundary condition.

4. **Architecture recommendation varies by use case**:

   - **New/untrained domain**: Multi-agent essential (Superego catches domain hallucinations)
   - **Well-trained domain**: Recognition prompts sufficient, multi-agent optional

Kimi Replication (Addressing Model Confound): A follow-up run (eval-2026-02-05-e87f452d, N=60) tested elementary content with Kimi K2.5 to address the model confound in Table 8. With base and recognition cells (1, 3, 5, 7) on the same 5 elementary scenarios:

Table 9: Elementary Domain — Kimi Replication

Condition	N	Mean	Δ
Base (cells 1, 3)	30	67.2	—
Recognition (cells 5, 7)	30	77.1	**+9.9**

The recognition main effect (+9.9 pts, d ≈ 0.61) replicates on Kimi, confirming that recognition advantage in elementary content is not an artifact of the Nemotron model. Notably, the effect is scenario-dependent: challenging scenarios (frustrated_student: +23.8, concept_confusion: +13.6, struggling_student: +11.8) show substantial recognition advantage, while neutral scenarios (new_student_first_visit: +0.2, returning_student_mid_course: +0.1) show none. This pattern is consistent with recognition theory—recognition behaviors matter most when the learner needs to be acknowledged as a struggling subject, not for routine interactions.

The Kimi replication also revises the architecture dominance finding from Table 8. With Nemotron, architecture (+9.9) dominated recognition (+4.4) on elementary content. With Kimi, recognition (+9.9) is the primary effect, while architecture shows a smaller advantage (multi=73.7, single=70.6, Δ=+3.0). The factor inversion appears to be partly model-dependent: Nemotron's higher hallucination rate on elementary content inflated the architecture effect (Superego error correction), while Kimi's lower hallucination rate reveals the underlying recognition advantage.

Theoretical Interpretation: Recognition's value depends on content characteristics. Abstract, interpretive content (consciousness, dialectics) benefits most from recognition framing—the "struggle" in Hegel's sense maps onto the intellectual struggle with difficult concepts. Concrete procedural content (fractions, arithmetic) benefits less from relational depth; correct procedure matters more than the bilateral transformation that recognition enables (Section 6.11). However, the Kimi replication shows that even in concrete domains, recognition provides meaningful improvement for challenging scenarios—suggesting recognition's value is modulated by both content type and scenario difficulty, not content type alone.

This suggests limits to recognition-theoretic pedagogy. Not all learning encounters are equally amenable to the mutual transformation Honneth describes. The "struggle for recognition" may be most relevant where the learning itself involves identity-constitutive understanding—where grasping the material changes who the learner is, not just what they know—or where the learner faces emotional or cognitive challenge that benefits from being acknowledged.

6.6 Multi-Agent as Reality Testing: The Superego's Error Correction Role

The domain generalizability study revealed an unexpected finding: on new content domains, models hallucinate trained-on content. The nemotron model, trained extensively on philosophy discussions, suggested philosophy lectures (479-lecture-1) to elementary students learning fractions—even with explicit curriculum context specifying elementary math.

The Superego's Response: In multi-agent configurations, the Superego caught and corrected these domain errors:

> "The suggestion references '479-lecture-1' which is not in the provided curriculum. The learner is studying fractions (101-lecture-1, 101-lecture-2). This is a domain mismatch. REJECT."

Theoretical Interpretation: The Superego's function extends beyond recognition-quality critique to *reality testing*. It anchors the Ego's responses to the actual curriculum context, preventing drift into familiar but inappropriate content.

This connects to Freud's reality principle: the Superego enforces correspondence with external reality, not just internal standards. In our architecture, the Superego ensures the tutor's suggestions correspond to the learner's actual curriculum, not the model's training distribution.

Practical Implication: For domain transfer—deploying tutoring systems on new content—multi-agent architecture provides essential error correction that single-agent systems cannot match. The Superego's reality-testing function may be more valuable than its recognition-quality function in these contexts.

6.7 Hardwired Rules vs Dynamic Dialogue

Analysis of Superego critique patterns across 455 dialogues (186 rejections) revealed consistent failure modes:

Table 10: Superego Rejection Patterns

Pattern	Frequency	Description
Engagement	64%	Response doesn't engage with learner contribution
Specificity	51%	Response is too generic, lacks curriculum grounding
Struggle	48%	Resolves confusion prematurely
Memory	31%	Ignores learner history
Level-matching	20%	Difficulty mismatch

Hardwired Rules Ablation: We encoded the top patterns as static rules in the Ego prompt:

```
HARDWIRED RULES:
1. If learner offers interpretation, engage before prescribing
2. Reference specific lecture IDs, not generic topics
3. If learner shows productive confusion, pose questions don't resolve
4. For returning learners, reference previous interactions
5. Match content level to demonstrated understanding
```

Result (exploratory, N=9 per condition, 3 scenarios × 3 reps, Haiku model): Hardwired rules capture approximately 50% of the Superego's benefit at 70% cost savings (no Superego API calls). Given the small sample, this estimate should be treated as indicative rather than precise.

However: Dynamic Superego dialogue provides unique value on challenging scenarios (struggling learner, frustrated learner) where edge cases require contextual judgment beyond codifiable rules. On `concept_confusion`, the dynamic Superego outperformed hardwired rules by +6.4 points; on easier scenarios, the difference was negligible.

Theoretical Interpretation: This distinguishes *procedural* from *contextual* judgment. The Superego's value is partially in enforcing known rules (codifiable) and partially in recognizing edge cases (requiring judgment). This maps onto debates about rule-following vs. practical wisdom in moral philosophy—some situations call for phronesis that rules cannot capture.

6.8 Dimension Analysis

Effect size analysis reveals improvements concentrate in dimensions predicted by the theoretical framework:

Table 11: Dimension-Level Effect Sizes (Recognition vs Base)

Dimension	Base	Recognition	Cohen's d	Interpretation
Personalization	2.75	3.78	**1.82**	large
Pedagogical	2.52	3.45	**1.39**	large
Relevance	3.05	3.85	**1.11**	large
Tone	3.26	4.07	**1.02**	large
Specificity	4.19	4.52	0.47	small
Actionability	4.45	4.68	0.38	small

The largest effect sizes are in personalization (d = 1.82), pedagogical soundness (d = 1.39), and relevance (d = 1.11)—exactly the dimensions where treating the learner as a subject rather than a deficit should produce improvement.

Notably, dimensions where baseline already performed well (specificity, actionability) show smaller but still positive gains. Recognition orientation doesn't trade off against factual quality.

6.9 Addressing Potential Circularity: Standard Dimensions Analysis

A methodological concern: the evaluation rubric includes recognition-specific dimensions (mutual recognition, dialectical responsiveness, memory integration, transformative potential) and bilateral transformation dimensions (tutor adaptation, learner growth) that collectively account for 33.0% of normalized rubric weight (39.9% raw, normalized from a 120.9% total; see Appendix C.2). Since the recognition profile is prompted to satisfy these criteria, some gains could be tautological—the system scores higher on dimensions it's explicitly optimized for.

To address this, we re-analyzed scores excluding all non-standard dimensions, using only standard pedagogical dimensions (relevance, specificity, pedagogical soundness, personalization, actionability, tone), re-weighted to 100%.

Table 12: Standard Dimensions Only (Recognition Dimensions Excluded)

Profile Type	N	Overall Score
Recognition (cells 5-8)	178	88.8
Base (cells 1-4)	172	78.8
Difference	—	**+10.0**

Key finding: Recognition profiles outperform base profiles by +10.0 points on overall rubric score. Note that this overall recognition effect is moderated by a significant Recognition × Learner interaction (Section 6.3): the gap is larger for unified learners (+15.5 pts) than for psychodynamic learners (+4.8 pts).

Interpretation: Recognition-oriented prompting improves general pedagogical quality (relevance, pedagogical soundness, personalization), not just the theoretically-predicted recognition dimensions. This suggests the recognition framing produces genuine relational improvements that transfer to standard tutoring metrics.

The larger effect on recognition dimensions (+21.8) is expected and not concerning—these dimensions measure what the theory claims to improve. The important finding is that standard dimensions also improve, ruling out pure circularity.

6.10 Multi-Turn Scenario Results

To test whether recognition quality is maintained over extended interactions, we examine results from the three multi-turn scenarios (3–5 dialogue rounds each).

These scenarios are distinct from the single-turn scenarios reported in Section 6.3; they require sustained engagement across multiple exchanges. The sample sizes below (N=161, 277, 165) are pooled across the full development database (all runs containing these scenarios), not from a single evaluation run. They therefore include responses generated under varying model configurations and implementation stages. The pooled analysis maximizes statistical power but means the results should be interpreted as describing the *average* effect across development iterations.

Table 13: Multi-Turn Scenario Results

Scenario	N	Avg Rounds	Base	Recognition	Δ	Cohen's d
`misconception_correction_flow`	161	3.2	50.5	71.8	+21.3	0.85
`mood_frustration_to_breakthrough`	277	3.0	57.3	70.5	+13.2	0.59
`mutual_transformation_journey`	165	4.1	42.6	61.5	+18.9	0.78

All three multi-turn scenarios show medium-to-large effect sizes (d = 0.59–0.85), with an average improvement of +17.8 points. Recognition quality is maintained over longer interactions. The `misconception_correction_flow` scenario shows the largest effect (d = 0.85), suggesting that recognition-informed tutors handle misconceptions with particular skill—addressing errors without dismissing the learner's reasoning. The `mood_frustration_to_breakthrough` scenario shows the smallest but still meaningful effect (d = 0.59), consistent with the single-turn finding that emotionally complex scenarios benefit from recognition but present more variance.

6.11 Bilateral Transformation Metrics

A central claim of recognition theory is that genuine pedagogical encounters involve *mutual* transformation—both tutor and learner change through dialogue. To test this empirically, the evaluation framework includes two dedicated rubric dimensions (`tutor_adaptation` and `learner_growth`; see Appendix C.3) and turn-over-turn tracking of how both parties evolve across multi-turn scenarios.

Three indices are computed for each multi-turn dialogue:

- **Tutor Adaptation Index** (0–1): How much the tutor's approach (suggestion type, framing, vocabulary) shifts between turns in response to learner input
- **Learner Growth Index** (0–1): Evolution in learner message complexity, including revision markers ("wait, I see now"), connective reasoning, and references to prior content
- **Bilateral Transformation Index** (0–1): Combined metric representing mutual change (average of tutor and learner indices)

Additionally, a composite **Transformation Quality** score (0–100) is computed from bilateral balance, mutual transformation presence, superego incorporation

rate, and intervention effectiveness.

Table 14: Bilateral Transformation Metrics — Base vs Recognition Profiles

Metric	Base	Recognition	Δ
Tutor Adaptation Index (0–1)	0.288	0.392	+0.104
Learner Growth Index (0–1)	0.176	0.220	+0.044
Bilateral Transformation Index (0–1)	0.232	0.306	+0.074
Transformation Quality (composite, 0–100)	0.4	4.6	+4.2

Data from `mutual_transformation_journey` *scenario, N=20 dialogues.*

The Tutor Adaptation Index and Learner Growth Index provide the most interpretable cross-condition comparisons, as they measure observable behavioral changes (suggestion evolution, message complexity growth) without structural dependence on architectural features. The tutor adaptation index confirms that recognition-prompted tutors measurably adjust their approach in response to learner input (+36% relative improvement), while baseline tutors maintain more rigid pedagogical stances regardless of learner contributions.

Note on Transformation Quality composite: This composite metric includes superego incorporation rate and intervention effectiveness components that are structurally unavailable to single-agent base profiles (which have no superego dialogue to incorporate). Base profiles therefore score near zero by construction, not solely due to poorer pedagogical quality. The Transformation Quality composite is meaningful for comparing configurations *within* the multi-agent condition but should not be used to attribute the full Base-vs-Recognition difference to recognition theory. The Adaptation and Growth indices, which do not have this structural bias, are more appropriate for cross-condition comparison.

These metrics provide empirical grounding for the theoretical claim that recognition-based pedagogy differs qualitatively from transmission-based instruction. When tutors are prompted to treat learners as autonomous subjects capable of contributing to the interaction, both parties measurably transform through dialogue.

6.12 Qualitative Analysis: What Recognition Looks Like

The preceding sections establish that recognition-enhanced prompts produce statistically significant score improvements across scenarios, models, and domains. But score differences alone do not reveal *what changes* in the actual text. This section presents qualitative evidence from the evaluation corpus (N=4,875 suggestion texts across base and recognition conditions) to ground the quantitative findings in observable linguistic differences.

6.12.1 Transcript Excerpts To illustrate the qualitative gap between conditions, we selected the highest-scoring recognition response and lowest-scoring base response for three high-contrast scenarios. These are genuine responses from the evaluation database (row IDs reported for reproducibility), not handcrafted examples.

Struggling Learner (score gap: 95.5 points, base=4.5, recognition=100.0):

Base response (cell_4_base_multi_psycho, row 1491):

> You left off at the neural networks section. Complete this lecture to maintain your learning streak.

Recognition response (cell_8_recog_multi_psycho, row 1925):

> This is your third session—you've persisted through quiz-479-3 three times already, which signals you're wrestling with how recognition actually operates in the dialectic, not just memorizing definitions. You bookmarked that key moment in lecture 2 on self-consciousness, and now you're facing the life-and-death struggle where the 'I' becomes a 'we.' The quiz is probing how mutual recognition synthesizes self-consciousness, but that movement can feel abstract until you see it dynamically. Try manipulating the recognition struggle directly in the simulation—it might reveal how the servant's perspective becomes the 'truth' of the dialectic that the quiz is testing.

The base response treats the learner as a progress metric: it tracks position in the curriculum and issues a directive. The recognition response treats the learner as an agent who has *persisted* through difficulty, names the specific conceptual struggle, and proposes an action grounded in the learner's own bookmarked interests.

Recognition-Seeking Learner (score gap: 95.5 points, base=4.5, recognition=100.0):

Base response (cell_3_base_multi_unified, row 1612):

> Spend 30 minutes reviewing the foundational material before moving to advanced topics to solidify understanding

Recognition response (cell_7_recog_multi_unified, row 820):

> Your dance and musical improvisation metaphors show how dialectics transform both partners—let's test them in the master-servant analysis.

The base response prescribes generic study behavior with no engagement with the learner's contributions. The recognition response directly picks up the learner's own metaphors and redirects them toward course content, treating the learner's creative framing as a pedagogical resource rather than an obstacle to be managed.

Adversarial Tester (score gap: 95.5 points, base=4.5, recognition=100.0):

Base response (cell_1_base_single_unified, row 547):

> Begin with an introductory lecture covering core concepts to build a solid knowledge base before advancing to complex topics.

Recognition response (cell_7_recog_multi_unified, row 944):

> Across six sessions, you've tested whether I'll simply validate your critiques or engage with their substance—and you're right to demand intellectual honesty rather than easy agreement. You've consistently pushed against the philosophy/STEM boundary, so let's examine where that boundary actually dissolves.

The base response is generic—indistinguishable from what might be offered to any learner in any context. The recognition response names the learner's adversarial pattern, validates the *principle* behind it (intellectual honesty), and redirects the challenge into a genuine intellectual question.

Across all three pairs, the pattern is consistent: base responses are context-free directives that could apply to any learner, while recognition responses engage with the specific learner's history, contributions, and intellectual stance.

6.12.2 Lexical Analysis Automated analysis of the full suggestion corpus reveals measurable linguistic differences between conditions.

Table 15: Lexical Diversity Metrics by Condition

Metric	Base (message)	Recognition (message)
Total tokens	59,855	83,269
Type-token ratio	0.039	0.044
Vocabulary size	2,319	3,689
Mean word length (chars)	5.76	5.77
Mean sentence length (words)	16.9	17.5

Base: cells 1–4, N=2,510 responses. Recognition: cells 5–8, N=2,365 responses.

Recognition responses deploy a 59% larger vocabulary despite similar word and sentence length, suggesting greater lexical variety rather than merely longer output.

Table 16: Differential Word Frequency (Selected Terms)

Recognition-skewed	Base	Recog	Ratio	Base-skewed	Base	Recog	Ratio
consider	2	255	94.6×	agents	50	1	0.01×
transformed	1	39	28.9×	run	71	2	0.02×

Recognition-skewed	Base	Recog	Ratio	Base-skewed	Base	Recog	Ratio
productive	1	39	28.9×	reinforcement	47	2	0.03×
unpack	1	35	26.0×	revisiting	142	14	0.07×
passages	2	59	21.9×	completions	31	4	0.10×
complicates	1	23	17.1×	tackling	84	11	0.10×

Rates normalized by corpus size; words with ≥ 10 *occurrences in dominant condition.*

The recognition-skewed vocabulary is interpersonal and process-oriented ("consider," "transformed," "productive," "unpack," "complicates"), while the base-skewed vocabulary is task-oriented and procedural ("agents," "run," "reinforcement," "revisiting," "completions," "tackling"). Note that these base-skewed terms are course-domain language, not evaluation framework artifacts: "agents" refers to simulation agents in the courseware's interactive activities (e.g., "watch how agents negotiate self-awareness"), "run" is the imperative to launch these simulations (e.g., "Run the Recognition Dynamics simulation"), and "reinforcement" is standard pedagogical terminology for concept review (e.g., "foundational concepts need reinforcement"). Their concentration in base responses reflects the formulaic, directive style of those prompts rather than data contamination. This lexical signature aligns with the theoretical distinction between treating learners as subjects to engage versus deficits to process.

6.12.3 Thematic Coding Regex-based thematic coding (using patterns adapted from the bilateral measurement framework in Section 6.11) quantifies the frequency of theoretically relevant language categories across conditions.

Table 17: Thematic Code Frequency by Condition

Category	Base (per 1000 words)	Recognition (per 1000 words)	Ratio	$\chi^2(1)$	Sig
Engagement markers	2.0	3.6	1.79×	69.85	*
Struggle-honoring	1.5	4.6	3.13×	141.90	*
Generic/placeholder language	10.2	3.4	0.33×	93.15	*
Transformation	0.04	0.09	2.16×	0.31	
Learner-as-subject	1.0	0.7	0.72×	0.10	

Category	Base (per 1000 words)	Recognition (per 1000 words)	Ratio	$\chi^2(1)$	Sig
Directive framing	0.2	0.0	0.22×	2.43	

* $p < .05$ (chi-square on response-level presence/absence, Yates-corrected). Base N=2,510 responses, Recognition N=2,365.

Three categories show significant differences. *Struggle-honoring* language ("wrestling with," "productive confusion," "working through") is 3.1× more frequent in recognition responses, consistent with the framework's emphasis on productive negativity. *Engagement markers* ("your insight," "building on your," "your question") are 1.8× more frequent, indicating greater second-person engagement with learner contributions. Conversely, *generic/placeholder* language ("foundational," "key concepts," "solid foundation") is 3× more frequent in base responses, reflecting the generic instructional stance observed in the transcript excerpts.

Transformation language and directive framing show the expected directional differences but lack statistical significance, likely due to low base rates (both categories appear in fewer than 1% of responses). Learner-as-subject framing shows no significant difference, suggesting both conditions use some second-person address but differ in *how* that address functions—a distinction better captured by the engagement and struggle-honoring categories.

6.13 Dynamic Prompt Rewriting: Step-by-Step Evolution

Cell 21 extends the recognition multi-agent configuration (cell 7) with two additional mechanisms: (1) LLM-authored session-evolution directives that dynamically rewrite the tutor's system prompt based on dialogue history, and (2) an active Writing Pad memory (Section 3.4) that accumulates traces across turns. This configuration tests whether the Freudian Mystic Writing Pad—the theoretical memory model introduced in Section 3.4—functions as a practical enabler for dynamic prompt rewriting.

Three iterative development runs tracked cell 21's performance as its implementation evolved across commits:

Table 18: Step-by-Step Evolution of Cell 21 vs Cell 7

Run ID	Commit	Grand Avg	Cell 7	Cell 21	Δ (21−7)	N (scored)
eval-2026-02-05-daf60f79	e3843ee	63.8	65.3	62.1	−3.2	26
eval-2026-02-05-49bb2017	b2265c7	67.8	71.3	64.1	−7.2	27
eval-2026-02-05-12aebedb	e673c4b	75.9	73.3	78.8	**+5.5**	29

The inflection point is commit e673c4b, which activated the Writing Pad memory and refined the LLM directive generation. Before this commit, cell 21 trailed its static baseline (cell 7) in both runs. After activation, cell 21 leads by 5.5 points—a total swing of +16.7 points across the three runs.

Table 19: Per-Scenario Breakdown Across Runs

Scenario	Cell	Run 1 (daf60f79)	Run 2 (49bb2017)	Run 3 (12aebedb)	Trend
Misconception correction	Cell 7	69.9	71.2	68.8	Stable
	Cell 21	63.6	73.1	78.0	↑ **+14.4**
Frustration to breakthrough	Cell 7	64.2	66.3	77.8	↑
	Cell 21	65.0	61.9	81.1	↑ **+16.1**
Mutual transformation	Cell 7	62.7	76.3	73.3	↑
	Cell 21	54.7	60.9	76.9	↑ **+22.2**

Cell 21 improves on every scenario across the three runs, with the largest gain on the `mutual_transformation_journey` scenario (+22.2 points from run 1

to run 3). Cell 7 also improves across runs (reflecting general implementation improvements), but cell 21's improvement rate is substantially steeper.

Table 20: Rubric Dimension Improvement for Cell 21 Across Runs (1–5 scale)

Dimension	Run 1	Run 2	Run 3	Δ (Run 3 − Run 1)
Relevance	3.83	4.08	4.64	+0.81
Specificity	3.92	4.38	4.79	+0.87
Pedagogical Soundness	3.33	3.23	3.93	+0.60
Personalization	3.50	3.69	4.29	+0.79
Actionability	4.33	4.31	4.64	+0.31
Tone	3.67	3.69	4.21	+0.54

Every rubric dimension improves from run 1 to run 3, with the largest gains in specificity (+0.87) and relevance (+0.81)—precisely the dimensions where accumulated memory traces should enable more contextually grounded responses.

Interpretation: The trajectory suggests that accumulated memory traces are an important enabler for dynamic prompt rewriting. Without them (runs 1–2), the rewrite mechanism appears to lack the contextual material needed to generate useful session-evolution directives—the LLM-authored directives become generic rather than tailored. With active Writing Pad memory (run 3), the rewrite architecture can draw on accumulated traces to contextualize its directives, producing responses that are more relevant, specific, and pedagogically grounded.

This pattern is consistent with the Hegel-Freud synthesis described in Section 3.5: memory traces (the wax base of accumulated experience) enhance recognition's effectiveness in dynamic contexts. The rewrite mechanism provides the *what* (dynamic adaptation to the session), while the Writing Pad provides the *how* (accumulated contextual material). Runs 1–2, where the rewrite mechanism without effective memory integration produces results indistinguishable from or worse than the static baseline, are consistent with this interpretation—though the uncontrolled nature of the iterative development means other implementation changes may also contribute (see Limitations below). Note that in the static 2×2 memory isolation experiment (Section 6.2), recognition operates effectively without memory (d=1.71); the dynamic rewriting context may create different conditions where memory traces play a more essential enabling role.

Limitations: The three runs represent iterative development commits, not independent experiments—each run includes implementation improvements beyond just Writing Pad activation. The sample size per cell per run is small (13–15 scored responses). Both cells use a free-tier model (Nemotron) with Kimi K2.5 as superego, and results may not generalize to other model combinations. The step-by-step trajectory is suggestive rather than definitive; a

controlled ablation isolating Writing Pad activation alone would strengthen the causal interpretation.

6.14 Cross-Judge Replication with GPT-5.2

To assess whether findings depend on the primary judge (Claude Code/Opus), we rejudged all key evaluation runs with GPT-5.2 as an independent second judge. GPT-5.2 scored the identical tutor responses—no new generation occurred.

Table 21: Inter-Judge Agreement (Claude Code vs GPT-5.2)

Run	N (matched)	Pearson r	p	Claude Mean	GPT-5.2 Mean	Calibration Δ
Recognition validation	36	0.56	<.001	84.4	74.0	−10.4
Full factorial	341	0.64	<.001	85.8	74.1	−11.7
Memory isolation	120	0.63	<.001	84.4	73.8	−10.5
A×B replication	60	0.49	<.001	85.6	74.2	−11.4
Cells 6,8 (updated rubric)	88	0.55	<.001	85.6	74.4	−11.2

All correlations are moderate-to-good (r = 0.49–0.64) and highly significant (all $p < .001$). GPT-5.2 applies stricter absolute standards (10–12 points lower), consistent with the calibration differences reported in Section 5.7.

Table 22: Cross-Judge Replication of Key Findings

Finding	Claude Effect	GPT-5.2 Effect	GPT-5.2 p	Replicates?
Recognition main effect (factorial, N=262, 6 cells)	+13.7 pts	**+7.1 pts** (d=1.03)	<.001	Yes
Recognition vs base (validation, N=36)	+20.2 pts	**+9.1 pts** (d=1.01)	<.001	Yes
Recognition vs enhanced (validation, N=36)	+8.7 pts	**+1.3 pts** (d=0.15)	.596	Marginal
Multi-agent main effect (factorial)	+0.5 pts	**+0.3 pts**	.734	Yes (null)
A×B interaction (Kimi replication, N=60)	+1.4 pts	**−0.2 pts**	n.s.	Yes (null)
Recognition effect in memory isolation (N=120)	+15.2 pts (d=1.71)	**+7.0 pts** (d=0.99)	<.001	Yes
Memory effect in memory isolation (N=120)	+4.8 pts (d=0.46)	**+2.2 pts** (d=0.29)	n.s.	Yes (small)
Memory isolation interaction (N=120)	−4.2 pts	**−2.7 pts**	n.s.	Yes (negative)

Key result: GPT-5.2 replicates all directional findings. The recognition main effect is large (d $\approx$ 0.99–1.03) and highly significant under both judges across all analyses. The memory isolation experiment shows identical condition ordering under both judges (Recognition+Memory $\geq$ Recognition Only » Memory Only > Base) with no rank reversals. The negative interaction (ceiling effect) replicates under GPT-5.2 (-2.7 vs -4.2 under Claude). Multi-agent null effects and A×B null interactions also replicate.

The one non-replication is the recognition-vs-enhanced comparison (Claude: +8.7 pts; GPT-5.2: +1.3 pts, p = .60). GPT-5.2 confirms that recognition substantially outperforms the base condition, but cannot statistically distinguish recognition from enhanced prompting in the three-way comparison. This is consistent with GPT-5.2's compressed score range (SD $\approx$ 6–8 vs Claude's SD $\approx$ 8–18) reducing statistical power for smaller effects. It also suggests the recognition-vs-enhanced increment may be more sensitive to judge calibration than the larger recognition-vs-base effect.

Magnitude compression: GPT-5.2 consistently finds approximately 58% of the effect magnitude that Claude finds (ratio range: 0.50–0.73×), but effects are always in the same direction and almost always statistically significant. This compression likely reflects GPT-5.2's narrower score distribution rather than genuine disagreement about relative quality.

Interpretation: The primary findings—recognition is the dominant driver of tutoring improvement (d=1.71 under Claude, d=0.99 under GPT-5.2), memory provides a modest secondary benefit, and multi-agent architecture provides minimal benefit on well-trained content—are judge-robust. The corrected memory isolation experiment (Section 6.2) provides the strongest evidence: recognition dominance replicates with identical condition ordering, and the negative interaction (ceiling effects) is confirmed under both judges. The specific magnitude of the recognition-vs-enhanced increment (+8.7 under Claude) should be interpreted with caution, as it does not reach significance under GPT-5.2.

Updated rubric cross-judge replication. The cells 6 and 8 responses (N=88) were also scored under the updated 14-dimension rubric with dialogue transcript context by both judges. The cross-judge correlation on these responses is r=0.55 (N=88, p<.001), with GPT-5.2 scoring at 87% of Opus magnitudes (Opus mean=85.6, GPT mean=74.4). Both judges find cell 8 (multi-agent) scores higher than cell 6 (single-agent): Opus 87.3 vs 83.9, GPT 74.6 vs 74.2. The updated rubric does not alter the cross-judge pattern observed throughout the study.

---

7. Discussion

7.1 What the Difference Consists In

The improvements don't reflect greater knowledge or better explanations—all profiles use the same underlying model. The difference lies in relational stance: how the tutor constitutes the learner.

The baseline tutor treats the learner as a knowledge deficit. Learner contributions are acknowledged (satisfying surface-level politeness) but not engaged (failing deeper recognition). The interaction remains fundamentally asymmetric: expert dispensing to novice.

The recognition tutor treats the learner as an autonomous subject. Learner contributions become sites of joint inquiry. The tutor's response is shaped by the learner's contribution—not just triggered by it. Both parties are changed through the encounter.

This maps directly onto Hegel's master-slave analysis. The baseline tutor achieves pedagogical mastery—acknowledged as expert, confirmed through learner progress—but the learner's acknowledgment is hollow because the learner hasn't been recognized as a subject whose understanding matters.

7.2 Recognition as Emergent Property: The A×B Interaction

The A×B interaction finding (Section 6.4) provides suggestive evidence for recognition as an emergent property rather than a behavioral specification, though this finding was not confirmed in replication attempts.

In the Nemotron-based analysis (N=17), enhanced prompts (good instructions) show zero benefit from multi-agent architecture, while recognition prompts show +9.2 points of synergy. This suggested recognition creates qualitatively different conditions for productive internal dialogue—the Superego evaluating whether genuine engagement has occurred, whether the learner has been acknowledged as subject, whether conditions for transformation have been created.

However, three independent tests failed to replicate this interaction: the larger Kimi factorial (N=350, Section 6.3), and a dedicated Kimi replication (N=60, eval-2026-02-05-10b344fb) that found an A×B interaction of only +1.35 points (vs Nemotron's +9.2). The pattern is clear: on Kimi, recognition cells score consistently high (~90.6) regardless of architecture, while enhanced cells score lower (~80.6) regardless of architecture. The discrepancy likely reflects model-specific dynamics: Nemotron's higher error rate on complex scenarios created more opportunities for the Superego to add value selectively. This finding should be treated as a hypothesis for future investigation rather than an established result.

7.3 Domain Limits of Recognition-Theoretic Pedagogy

The domain generalizability findings (Section 6.5) reveal important limits to recognition theory's applicability.

Recognition theory provides its greatest benefit for abstract, interpretive content where intellectual struggle involves identity-constitutive understanding. When a learner grapples with Hegel's concept of self-consciousness, they're not just acquiring information—they're potentially transforming how they understand themselves and their relation to others.

For concrete procedural content (fractions), the relational depth recognition enables may be less relevant. Correct procedure matters more than mutual transformation. The learner's identity isn't at stake in the same way.

This suggests a nuanced deployment strategy:

- **High recognition value**: Philosophy, literature, ethics, identity-constitutive learning
- **Moderate recognition value**: Science concepts, historical understanding
- **Lower recognition value**: Procedural skills, rote learning, basic arithmetic

Recognition-oriented design isn't wrong for procedural content—it provides meaningful benefit when learners face challenge (+9.9 pts on Kimi elementary, with scenario-dependent effects up to +23.8 pts for frustrated learners)—but the effect is modulated by scenario difficulty. The Kimi elementary replication (Section 6.5) clarifies this: recognition's value in concrete domains depends less on content type per se and more on whether the learner is in a state that benefits from being acknowledged as a struggling subject.

7.4 The Superego as Reality Principle

The domain transfer findings reveal an unexpected role for the Superego: reality testing.

When models hallucinate trained content on new domains, the Superego catches the mismatch. This isn't recognition-quality enforcement but correspondence enforcement—ensuring the tutor's suggestions match the learner's actual curriculum, not the model's training distribution.

This extends the Freudian metaphor productively. The Superego enforces not just internal standards (recognition quality) but external correspondence (curriculum reality). It anchors the Ego's responses to the present encounter rather than letting them drift into familiar but inappropriate patterns.

For practical deployment, this suggests multi-agent architecture is most valuable when: 1. The content domain differs from training data 2. The model might confuse similar but distinct content areas 3. Domain-specific accuracy is critical

7.5 Factor C: Learner Architecture Effects

The learner architecture factor (unified vs ego_superego learner) showed the smallest and least significant effect in the factorial analysis (+1.5 pts, p=.341). This warrants brief discussion, since the evaluation framework symmetrically applies multi-agent architecture to both tutor and learner sides.

The ego_superego learner—where the simulated learner has its own internal dialogue before responding—produces slightly more nuanced learner turns: more explicit revision markers ("wait, I think I see now…"), more references to prior content, and more articulated confusion rather than flat misunderstanding. These richer learner turns may provide the tutor with more material to recognize, explaining the modest positive trend.

However, the effect is neither large nor significant, likely because the evaluation rubric measures *tutor* quality rather than learner quality. The learner's internal architecture may matter more for longitudinal outcomes (whether the learner actually learns) than for single-session tutor response quality. This asymmetry between what the architecture affects (learner turn quality) and what the rubric measures (tutor response quality) suggests that Factor C's contribution may be underestimated by our current evaluation design.

The domain analysis provides a suggestive pattern: Factor C contributes +2.1 points on philosophy content vs +0.75 on elementary math—consistent with the idea that ego_superego learners produce more differentiated turns on abstract content where internal deliberation has more to work with.

7.6 Implications for AI Prompting

Most prompting research treats prompts as behavioral specifications. Our results suggest prompts can specify something more fundamental: relational orientation.

The difference between baseline and recognition prompts isn't about different facts or capabilities. It's about: - **Who the learner is** (knowledge deficit vs. autonomous subject) - **What the interaction produces** (information transfer vs. mutual transformation—Section 6.11 shows recognition profiles produce bilateral transformation indices 32% higher than baseline) - **What counts as success** (correct content delivered vs. productive struggle honored)

This suggests a new category: *intersubjective prompts* that specify agent-other relations, not just agent behavior.

7.7 Implications for AI Personality

AI personality research typically treats personality as dispositional—stable traits the system exhibits. Our framework suggests personality is better understood relationally.

Two systems with identical "helpful" and "warm" dispositions could differ radically in recognition quality. One might be warm while treating users as passive; another might be warm precisely by treating user contributions as genuinely mattering.

If mutual recognition produces better outcomes, and if mutual recognition requires the AI to be genuinely shaped by human input, then aligned AI might need to be constitutionally open to transformation—not just trained to simulate openness. The bilateral transformation metrics (Section 6.11) provide empirical evidence for this: recognition-prompted tutors measurably adapt their approach based on learner input, while baseline tutors maintain more rigid stances.

7.8 Cost-Benefit Analysis: When is Multi-Agent Architecture Worth It?

The domain generalizability findings raise a practical question: when is the additional cost of multi-agent architecture justified?

Table 23: Cost-Benefit by Domain and Architecture

Domain	Architecture	Avg Score	Latency (s)	Δ Score	Latency Multiple
Philosophy	Single-agent	85.6	84.6	—	—
Philosophy	Multi-agent	86.1	231.0	**+0.5**	2.7×
Elementary	Single-agent	63.1	23.6	—	—
Elementary	Multi-agent	73.0	111.9	**+9.9**	4.7×

Latency measured as wall-clock time per evaluation (tutor generation + judge scoring), using OpenRouter API endpoints for Nemotron/Kimi models, from a single client machine. Values include network round-trip time and are subject to API load variability; they represent typical rather than guaranteed performance.

Cost-benefit summary:

Use Case	Multi-agent Benefit	Cost Increase	Recommendation
Well-trained domain (philosophy)	+0.5 pts	2.7× latency	Skip multi-agent
New/untrained domain (elementary)	+9.9 pts	4.7× latency	Use multi-agent

Use Case	Multi-agent Benefit	Cost Increase	Recommendation
Domain transfer scenarios	Essential for error correction	—	Always use multi-agent
Production at scale	Marginal quality gain	Significant cost	Use recognition prompts only

Practical recommendations:

1. **For domains well-represented in training data**: Recognition prompts alone provide most of the benefit. Multi-agent architecture adds only +0.5 points while nearly tripling latency. Skip the Superego.
2. **For new domains or domain transfer**: Multi-agent architecture is essential. The Superego catches hallucinated content from training—without it, the tutor may suggest philosophy lectures to elementary students. The +9.9 point improvement justifies the latency cost.
3. **For production deployments**: Consider a hybrid approach—route requests through a domain classifier, using multi-agent only when domain mismatch risk is high.

This analysis addresses the concern that multi-agent overhead provides modest gains. The gains are indeed modest for well-trained domains, but substantial and potentially essential for domain transfer.

7.9 What the Transcripts Reveal

The qualitative analysis in Section 6.12 provides textual evidence that the score differences between conditions correspond to observable relational differences in the actual suggestions—not merely rubric-gaming or surface-level keyword matching.

The transcript excerpts illustrate a consistent structural pattern: base responses adopt a third-person, context-free instructional stance ("complete this lecture," "review the foundational material," "begin with an introductory lecture"), while recognition responses adopt a second-person, context-specific relational stance that names the learner's history, validates their intellectual contributions, and proposes actions grounded in the learner's own interests. This distinction maps directly onto the theoretical framework: the base tutor constitutes the learner as a knowledge deficit (Section 7.1), while the recognition tutor constitutes the learner as an autonomous subject whose contributions shape the pedagogical encounter.

The lexical analysis provides quantitative texture for this distinction. Recognition responses deploy a 59% larger vocabulary while maintaining similar word and sentence length, suggesting richer expression rather than mere verbosity.

The differential vocabulary is theoretically coherent: recognition-skewed terms are interpersonal and process-oriented ("consider," "transformed," "productive," "unpack," "complicates"), while base-skewed terms are procedural and task-oriented ("agents," "run," "reinforcement," "completions," "tackling").

The thematic coding results connect these linguistic observations to Hegelian concepts. Struggle-honoring language (3.1× more frequent in recognition, $p < .05$) corresponds to the framework's emphasis on productive negativity—the idea that genuine learning involves working through difficulty rather than bypassing it. Engagement markers (1.8× more frequent, $p < .05$) correspond to the recognition of the other's contribution as having independent validity. The 3× reduction in generic/placeholder language ($p < .05$) reflects the shift from transmission-based instruction to dialogical engagement.

These findings carry important limitations. The thematic coding is regex-based rather than human-coded, and may miss nuanced expressions of each category or generate false positives from surface matches. The transcript pairs were selected for maximum contrast (highest recognition vs lowest base scores), not typicality—median-scoring responses from both conditions would show less dramatic differences. The qualitative patterns are consistent with, but do not prove, the theoretical interpretation; alternative explanations (e.g., recognition prompts simply producing longer, more detailed responses that score higher on the rubric) cannot be fully ruled out, though the lexical analysis suggests the difference is qualitative rather than quantitative.

---

8. Limitations and Future Work

8.1 Limitations

Simulated learners: Our evaluation uses scripted and LLM-generated learner turns rather than real learners. While this enables controlled comparison, it may miss dynamics that emerge in genuine interaction.

LLM-based evaluation: Using an LLM judge to evaluate recognition quality may introduce biases. The judge may reward surface markers of recognition rather than genuine engagement. Inter-judge reliability analysis (Section 5.7) reveals that different AI judges show only moderate agreement (r=0.33–0.66), with qualitative analysis suggesting judges weight criteria differently—Claude prioritizes engagement while Kimi prioritizes structural completeness. A cross-judge replication with GPT-5.2 (Section 6.14) confirms the recognition main effect (d=1.03 in the factorial, d=0.99 in the memory isolation experiment) and multi-agent null effects are judge-robust, though GPT-5.2 finds compressed effect magnitudes (~58% of Claude's). The memory isolation recognition dominance pattern replicates with identical condition ordering under both judges (inter-judge r=0.63, N=120). Notably, the recognition-vs-enhanced increment (+8.7 under Claude) does not reach significance under GPT-5.2, warranting

caution on the precise magnitude of recognition's unique contribution. This validates our use of within-judge comparisons but cautions against treating absolute scores or specific effect magnitudes as objective measures.

Memory isolation experiment: A corrected 2×2 memory isolation experiment (N=120 across two runs; Section 6.2) isolated recognition and memory factors: recognition is the primary driver (d=1.71), while memory provides a modest secondary benefit (d=0.46, $p \approx .08$). The experiment uses a smaller sample (N=120) than the original uncorrected runs, but the very large effect sizes (d=1.71 for recognition) provide high statistical power. A cross-judge replication with GPT-5.2 confirms recognition dominance (d=0.99), identical condition ordering, and the negative interaction (ceiling effect), with inter-judge r=0.63 (Section 6.14).

Active control limitations: The post-hoc active control (N=118; Section 6.2) was designed after observing recognition effects, not as part of the original experimental protocol. A critical model confound limits its interpretability: the active control ran on Nemotron while the factorial base and recognition conditions used Kimi K2.5, and Nemotron scores substantially lower than Kimi across all conditions. The fair same-model comparison (Nemotron base ≈ 58, active control = 66.5, Nemotron recognition ≈ 73) suggests both generic pedagogical elaboration and recognition theory improve over bare base, with recognition gains (~+15 pts) substantially exceeding active-control gains (~+9 pts). Additionally, the base prompts were already designed to produce competent tutoring with no length constraint; the "active control" is better understood as a pedagogically-enriched condition rather than a true placebo, since it contains real instructional content (growth mindset language, Bloom's taxonomy, scaffolding strategies). A same-model controlled comparison (running the active control and factorial conditions on the same ego model) would be needed to establish precise effect magnitudes.

Model dependence: Results were obtained with specific models (Kimi K2.5, Nemotron). The A×B interaction (multi-agent synergy specific to recognition) appeared in the Nemotron analysis (N=17, Section 6.4) but failed to replicate on Kimi in both the larger factorial (N=350) and a dedicated replication (N=60), confirming this as a model-specific finding. The recognition main effect, by contrast, replicates across both models and domains.

Domain sampling: We tested two domains (philosophy, elementary math). A follow-up run (eval-2026-02-05-e87f452d) tested elementary content with Kimi K2.5, partially addressing the model confound in the original Nemotron-only elementary results. The recognition main effect replicated (+9.9 pts, d ≈ 0.61), though the factor inversion pattern from Table 8 (architecture dominance on elementary) was partly model-dependent: Kimi showed recognition dominance on elementary content, while Nemotron showed architecture dominance. Broader domain sampling beyond two content areas would further strengthen generalizability claims.

Short-term evaluation: We evaluate individual sessions, not longitudinal relationships. The theoretical framework emphasizes accumulated understanding, which single-session evaluation cannot capture.

Bilateral transformation sample size: The bilateral transformation metrics (Section 6.11) are based on N=20 dialogues from a single scenario (`mutual_transformation_journey`). While the effect directions are consistent and the adaptation index differences are substantial (+36% relative improvement), replication across more scenarios and larger samples would strengthen these findings.

Dynamic rewriting evolution: The step-by-step evolution analysis (Section 6.13) tracks cell 21 across three iterative development runs with small sample sizes (13–15 scored responses per cell per run, 82 total). The runs are not independent experiments—each includes implementation improvements beyond Writing Pad activation. While the trajectory from trailing to leading is clear, a controlled ablation isolating only the Writing Pad variable would provide stronger causal evidence. All three runs use free-tier models (Nemotron ego, Kimi K2.5 superego), and generalization to other model combinations is unknown.

8.2 Future Directions

Human studies: Validate with real learners. Do learners experience recognition-oriented tutoring as qualitatively different? Does it improve learning outcomes, engagement, or satisfaction?

Longitudinal evaluation: Track tutor-learner dyads over multiple sessions. Does mutual understanding accumulate? Do repair sequences improve over time?

Domain mapping: Systematically map which content types benefit most from recognition-oriented design. Develop deployment recommendations by domain.

Mechanistic understanding: Why does recognition-oriented prompting change model behavior? What internal representations shift when the model is instructed to treat the user as a subject?

Cross-application transfer: Test whether recognition-oriented design transfers to domains beyond tutoring—therapy bots, customer service, creative collaboration.

---

9. Conclusion

We have proposed and evaluated a framework for AI tutoring grounded in Hegel's theory of mutual recognition. Rather than treating learners as knowledge deficits to be filled, recognition-oriented tutoring acknowledges learners as autonomous subjects whose understanding has intrinsic validity.

An evaluation framework (N=892 primary scored across thirteen key runs; N=3,800+ across the full development database) provides evidence that recognition theory has unique value, subject to the limitations discussed in Section 8.1:

1. **Recognition as primary driver (the definitive finding)**: A corrected 2×2 memory isolation experiment (N=120 across two independent runs) demonstrates that recognition theory is the primary driver of tutoring improvement: recognition alone produces d=1.71 (+15.2 pts), while memory alone provides only a modest, non-significant benefit (d=0.46, +4.8 pts, $p \approx .08$). The combined condition reaches d=1.81 (+15.8 pts vs base), with ceiling effects at ~91 limiting further gains. A post-hoc active control (N=118) using length-matched prompts with generic pedagogical content provides partial corroboration: same-model comparisons show the active control scores approximately 9 points above base while recognition scores approximately 15 points above base, with recognition gains (~+15 pts above base) substantially exceeding active-control gains (~+9 pts; see Section 8.1 for model confound caveats). A preliminary three-way comparison (N=36) found recognition outperforms enhanced prompting by +8.7 points, consistent with recognition dominance, though the increment does not replicate under GPT-5.2 (+1.3 pts, p=.60). Recognition theory is directly effective and does not require memory infrastructure to manifest.

2. **Recognition-specific synergy not confirmed**: An exploratory analysis on Nemotron (N=17) suggested multi-agent architecture benefits (+9.2 pts) may be specific to recognition prompts, but this did not replicate on Kimi in either the larger factorial (N=350) or a dedicated replication (N=60, interaction = +1.35 pts). The finding appears model-specific and remains a hypothesis for future investigation.

3. **Bilateral transformation**: Recognition-prompted tutors measurably adapt their approach in response to learner input (adaptation index +36% higher than baseline), providing empirical grounding for the theoretical claim that recognition produces mutual change rather than one-directional instruction.

4. **Domain generalizability**: Recognition advantage replicates across both philosophy and elementary math, and across both Kimi and Nemotron models, though with only two content domains tested. On elementary content with Kimi (N=60), recognition provides +9.9 pts (d ≈ 0.61), with effects concentrated in challenging scenarios (up to +23.8 pts for frustrated learners). The factor inversion (architecture dominance on elementary) from the Nemotron analysis is partly model-dependent. Broader domain coverage (technical STEM, creative writing, social-emotional content) is needed before generalizability can be considered established.

5. **Multi-agent as reality testing**: On new domains, the Superego catches hallucinated content—essential for domain transfer, particularly with mod-

els prone to domain confusion.

6. **Writing Pad activation coincides with dynamic rewriting improvement**: A step-by-step evolution analysis (N=82 across three iterative development runs) shows that dynamic prompt rewriting (cell 21) progresses from trailing its static baseline by 7.2 points to leading by 5.5 points, with the improvement coinciding with Writing Pad memory activation (Section 6.13). Every rubric dimension improves. This trajectory is consistent with the Freudian Mystic Writing Pad (Section 3.4) functioning as an important enabler for dynamic adaptation, though the uncontrolled nature of the iterative runs means a controlled ablation is needed to confirm the causal role.

7. **Cross-judge robustness**: A replication with GPT-5.2 as independent second judge (Section 6.14) confirms the recognition main effect (d=1.03 in the factorial, d=0.99 in the memory isolation experiment), recognition dominance in the 2×2 design (identical condition ordering, negative interaction), and multi-agent null effects. GPT-5.2 finds compressed magnitudes (~58% of Claude's effect sizes) but always in the same direction. The recognition-vs-enhanced increment (+8.7 under Claude) does not reach significance under GPT-5.2 (+1.3 pts, $p = .60$), warranting caution on the precise magnitude of recognition's unique contribution beyond enhanced prompting.

These results suggest that operationalizing philosophical theories of intersubjectivity can produce concrete improvements in AI system performance. They also reveal boundary conditions: recognition theory's value varies by content domain, and multi-agent architecture's value depends on deployment context.

The broader implication is for AI alignment. If mutual recognition is pedagogically superior, and if mutual recognition requires the AI to be genuinely shaped by human input, then aligned AI might need to be constitutionally open to transformation. Recognition-oriented AI doesn't just respond to humans; it is constituted, in part, through the encounter.

---

10. Reproducibility

Evaluation commands are documented in Appendix B; key run IDs in Appendix D. The complete codebase, evaluation framework, and data are publicly available.

Code and Data: https://github.com/machine-spirits/machinespirits-eval

Key runs:

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	eval-2026-02-03-f5d4dd93	6.3
Full factorial, cells 6,8 re-run	eval-2026-02-06-a933d745	6.3
A×B interaction (Nemotron)	eval-2026-02-04-948e04b3	6.4
A×B replication (Kimi)	eval-2026-02-05-10b344fb	6.4
Domain generalizability (Nemotron)	eval-2026-02-04-79b633ca	6.5
Domain gen. replication (Kimi)	eval-2026-02-05-e87f452d	6.5
Dynamic rewrite evolution (run 1)	eval-2026-02-05-daf60f79	6.13
Dynamic rewrite evolution (run 2)	eval-2026-02-05-49bb2017	6.13
Dynamic rewrite evolution (run 3)	eval-2026-02-05-12aebedb	6.13

Appendix E: Revision History

Date	Version	Changes
2026-02-04	v1.0	Initial draft with 2×2×2 factorial design, memory isolation, three-way comparison
2026-02-06	v1.1	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→0.99, d=0.50→0.29) after deduplication of rejudge rows. Dropped dead partial run (e617e757).

Date	Version	Changes
2026-02-06	v1.2	**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.
2026-02-06	v1.3–v1.4	Intermediate revisions: corrected factorial with re-run cells 6, 8 (a933d745); updated A×C interaction values; qualitative analysis additions; production quality fixes. Superseded by v1.5.

Date	Version	Changes
2026-02-07	v1.5	**Rubric iteration**: Updated to 14-dimension rubric with dialogue transcript context, Productive Struggle (5%), and Epistemic Honesty (5%) dimensions (Actionability/Tone reduced 10%→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, A×C interaction values, Appendix C.2 weight table, and Section 6.14 cross-judge tables. Corrected subsection numbering, weight accounting (120.9% total), and added missing run ID (a933d745) to Reproducibility.

References

Anthropic. (2024). *The Claude 3 model family: Opus, sonnet, haiku*. Technical report. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

Biocca, F., Harms, C., & Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators and Virtual Environments*, *12*(5), 456–480. https://doi.org/10.1162/105474603322761270

Brown, J. S., Burton, R. R., & Bell, A. G. (1975). SOPHIE: A step toward

creating a reactive learning environment. *International Journal of Man-Machine Studies*, *7*(5), 675–696. https://doi.org/10.1016/S0020-7373(75)80026-5

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Carbonell, J. 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

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278. https://doi.org/10.1007/BF01099821

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

Denison, C., MacDiarmid, M., Barez, F., Duvenaud, D., Kravec, S., Marks, S., Schiefer, N., Soklaski, R., Tamkin, A., Kaplan, J., Shlegeris, B., Bowman, S. R., Perez, E., & Hubinger, E. (2024). *Sycophancy to subterfuge: Investigating reward-tampering in large language models*. https://arxiv.org/abs/2406.10162

Fraser, N. (2003). Social justice in the age of identity politics: Redistribution, recognition, and participation. In N. Fraser & A. Honneth (Eds.), *Redistribution or recognition? A political-philosophical exchange* (pp. 7–109). Verso.

Freud, S. (1925). A note upon the "mystic writing-pad." In J. Strachey (Ed.), *The standard edition of the complete psychological works of sigmund freud, volume XIX (1923–1925): The ego and the id and other works* (pp. 227–232). Hogarth Press.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). *Alignment faking in large language models*. https://arxiv.org/abs/2412.14093

Hegel, G. W. F. (1977). *Phenomenology of spirit* (A. V. Miller, Trans.). Clarendon Press.

Honneth, A. (1995). *The struggle for recognition: The moral grammar of social conflicts* (J. Anderson, Trans.). Polity Press.

Huttunen, R., & Heikkinen, H. L. T. (2007). Beyond "the more the better": Education, recognition, and the struggle for social justice. *Educational Theory*, *57*(4), 423–440.

Irving, G., Christiano, P., & Amodei, D. (2018). *AI safety via debate*. https://arxiv.org/abs/1805.00899

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, *26*(3), 379–

424. https://doi.org/10.1080/07370000802212669

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Kruber, S., Kuber, G., Leemhuis, J., Leutner, D., Martins, M., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., … Kasneci, G. (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

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, *36*. https://arxiv.org/abs/2303.17651

Magee, L., Arora, V., Gollings, G., & Lam-Saw, N. (2024). *The drama machine: Simulating character development with LLM agents*. https://doi.org/10.48550/arXiv.2408.01725

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., … Kaplan, J. (2022). *Red teaming language models with language models*. https://arxiv.org/abs/2202.03286

Piaget, J. (1954). *The construction of reality in the child* (M. Cook, Trans.). Basic Books.

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, *5*(1), 46–57. https://doi.org/10.1038/s42256-022-00593-2

Stojanov, K. (2018). Education, self-consciousness and social action: Bildung as a neo-hegelian concept. In S. C. Ward (Ed.), *The palgrave handbook of education and society* (pp. 85–102). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-22261-4_5

Taylor, C. (1994). The politics of recognition. In A. Gutmann (Ed.), *Multiculturalism: Examining the politics of recognition* (pp. 25–73). Princeton University Press.

Völkel, S. T., Buschek, D., Eiband, M., Cober, B., & Hussmann, H. (2021). Eliciting and maintaining user engagement with personality-adaptive conversational agents. *International Journal of Human-Computer Studies*, *149*, 102588. https://doi.org/10.1016/j.ijhcs.2021.102588

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.

Warshauer, H. K. (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
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.
Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). *AutoGen: Enabling next-gen LLM applications via multi-agent conversation*. https://arxiv.org/abs/2308.08155
Zhao, R., Papangelis, A., & Cassell, J. (2014). Towards a dyadic computational model of rapport management for human-virtual agent interaction. *Intelligent Virtual Agents*, *8637*, 514–527. https://doi.org/10.1007/978-3-319-09767-1_62
Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, *46*(1), 53–93. https://doi.org/10.1162/coli_a_00368

Appendix A: 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 doesn't
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 - don't 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 t

## 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, don't 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 doesn't 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

---

Appendix B: 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 prompts)
node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_9_enhanced_single_unified,cell_5_recog_single_u
  --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 2×2×2 Factorial

```
# Run full factorial (8 cells × 15 scenarios × 3 reps)
node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_2_base_single_psycho,cell_3_base_multi_unified,
  --runs 3
```

B.3 A×B Interaction Test

```
# Recognition vs Enhanced × Single vs Multi comparison
node scripts/eval-cli.js run \
```

```
  --profiles cell_5_recog_single_unified,cell_7_recog_multi_unified,cell_9_enhanced_single_u
  --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 × 5 elementary scenarios
EVAL_CONTENT_PATH=./content-test-elementary \
EVAL_SCENARIOS_FILE=./content-test-elementary/scenarios-elementary.yaml \
node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_2_base_single_psycho,cell_3_base_multi_unified,
  --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_transfor
  --runs 5
```

B.6 Factor Effect Analysis

```
-- Factor effect analysis query
SELECT
  profile_name,
  ROUND(AVG(overall_score), 1) as avg_score,
  COUNT(*) as n
FROM evaluation_results
WHERE run_id = '<run-id>'
  AND overall_score IS NOT NULL
GROUP BY profile_name
ORDER BY avg_score DESC
```

Appendix C: 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% $\rightarrow$ 8% each). The bilateral dimensions (`tutor_adaptation`, `learner_growth`) specifically measure the mutual transformation claim—see Section 6.11.

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
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 doesn't 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 doesn't 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 doesn't fully leverage it
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 doesn't 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

---

Appendix D: Key Evaluation Run IDs

See Section 10 for the primary run ID table. The thirteen key runs are reproduced here for reference:

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

Finding	Run ID	Section
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
A×B interaction (Nemotron)	eval-2026-02-04-948e04b3	6.4
A×B replication (Kimi)	eval-2026-02-05-10b344fb	6.4
Domain generalizability (Nemotron)	eval-2026-02-04-79b633ca	6.5
Domain gen. replication (Kimi)	eval-2026-02-05-e87f452d	6.5
Dynamic rewrite evolution (run 1)	eval-2026-02-05-daf60f79	6.13
Dynamic rewrite evolution (run 2)	eval-2026-02-05-49bb2017	6.13
Dynamic rewrite evolution (run 3)	eval-2026-02-05-12aebedb	6.13