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

Liam Magee

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 the "Drama Machine" architecture: an Ego/Superego multiagent system where an external-facing tutor agent (Ego) generates pedagogical suggestions that are reviewed by an internal critic agent (Superego) before reaching the learner.

An evaluation framework (N=1,486 primary scored responses across eighteen key runs; N=3,800+ across the full development database) isolating recognition theory from prompt engineering effects and memory integration reveals that recognition theory is the primary driver of tutoring improvement: a corrected 2×2 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 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; see Section 8 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 reach significance under GPT-5.2 (+1.3 pts, p=.60). The multi-agent tutor architecture contributes **+0.5 to +10 points** depending on content domain—minimal on well-trained content but critical for domain transfer where it catches content isolation errors. A step-by-step evolution analysis of dynamic prompt rewriting with active Writing Pad memory (N=82 across three runs) suggests the Freudian memory model as an important enabler—the rewrite cell progresses from trailing its baseline by 7.2 points to leading by 5.5 points coinciding with Writing Pad activation, though controlled ablation is needed to confirm causality.

Three key findings emerge: (1) Recognition theory is the primary driver of improvement—recognition alone produces d=1.71, while memory provides a modest secondary benefit (d=0.46), with an active control showing recognition gains (~+15 pts above same-model base) substantially exceeding active-control gains (~+9 pts); (2) Multi-agent architecture is additive, not synergistic—a dedicated five-model probe (Kimi K2.5, Nemotron, DeepSeek V3.2, GLM-4.7, Claude Haiku 4.5; N=826 total) finds the A×B interaction consistently near zero or negative (mean −2.2 pts) across all models, definitively ruling out recognition-specific synergy; (3) Domain generalizability testing confirms recognition advantage replicates across both models and content domains—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. Bilateral transformation tracking across three multi-turn scenarios (N=118) confirms that recognition-prompted tutors measurably adapt their approach in response to learner input (+26% relative improvement in adaptation index), though learner-side growth is not higher under recognition, suggesting tutor-side responsiveness rather than symmetric mutual transformation.

A cross-judge replication with GPT-5.2 confirms the main findings are judge-robust: the recognition 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 all replicate, though at compressed magnitudes (~58% of primary judge effect sizes).

These findings suggest that recognition theory's value is domain-sensitive, multi-agent architecture provides essential error correction for domain transfer, and optimal deployment configurations depend on content characteristics.

The system is deployed in an open-source learning management system with all code, evaluation data, and reproducible analysis commands publicly available.

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, does not truly count. Only mutual recognition—where each party acknowledges the other as an autonomous subject—produces genuine selfhood.

The connection between Hegelian thought and pedagogy is well established. Vygotsky's zone of proximal development (Vygotsky, 1978) presupposes a dialogical relationship that echoes Hegel's mutual constitution of self-consciousness; the *Bildung* tradition frames education as self-formation through encounter with otherness (Stojanov, 2018); and recognition theory (Honneth, 1995) has been applied to educational contexts (Huttunen & Heikkinen, 2007). Our contribution is to operationalize these commitments as design heuristics for AI tutoring and 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 does not 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 is 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.

1.1 Contributions

We make the following contributions:

1. **The Drama Machine Architecture**: A complete multiagent tutoring system with Ego and Superego agents, implementing the Superego as a *ghost* (internalized memorial authority) rather than an equal dialogue partner.

2. **Memory Isolation Experiment**: A corrected 2×2 experiment (N=120 across two independent runs) demonstrating recognition as the primary driver (d=1.71), with memory providing a modest secondary benefit (d=0.46) and ceiling effects limiting observable synergy. A post-hoc active control (N=118) shows recognition gains (~+15 pts) substantially exceeding active-control gains (~+9 pts above same-model base).

3. **Robust Factorial Evaluation**: A 2×2×2 factorial design (N=1,486 primary scored across eighteen key runs; N=3,800+ across the full development database) across multiple models, scenarios, and conditions, providing statistically robust effect estimates. A significant Recognition × Learner interaction (F=21.85, p<.001) reveals that recognition benefits single-agent learners far more (+15.5 pts, d=1.28) than multi-agent learners (+4.8 pts, d=0.37).

3b. **Three-Way Comparison**: Evidence from a base vs. enhanced vs. recognition comparison (N=36) consistent with recognition dominance, showing recognition outperforms enhanced prompting by +8.7 points.

4. **A×B Interaction Analysis**: A dedicated five-model probe (N=826 total) definitively establishes that multi-agent architecture is additive, not synergistic—the A×B interaction is consistently near zero or negative across all five ego models tested (mean −2.2 pts), ruling out recognition-specific synergy.

5. **Domain Generalizability Testing**: Evaluation on elementary mathematics content across two models confirming recognition advantage replicates, with multi-agent architecture providing critical error correction for domain transfer.

6. **Hardwired Rules Ablation**: A larger replication (N=72) reversing the initial finding—encoding the Superego's most common critique patterns as static rules *degrades* performance rather than replicating its benefit, supporting a *phronesis* interpretation where the Superego's value lies in contextual judgment.

7. **Bilateral Transformation Metrics**: Empirical evidence (N=118, three multi-turn scenarios) that recognition-prompted tutors measurably adapt their approach (+26%), though learner-side growth does not increase, qualifying the "mutual transformation" claim as primarily tutor-side responsiveness.

8. **Reproducible Evaluation Framework**: Complete documentation of evaluation commands and run IDs enabling independent replication of all findings.

---

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 (Kasneci et al., 2023). Recent multi-agent approaches include GenMentor (Wang et al., 2025), which decomposes tutoring into five specialized agents, and Ruffle&Riley (Schmucker et al., 2024), which orchestrates two LLM agents in a learning-by-teaching format. A comprehensive survey (Chu et al., 2025) maps the growing landscape of LLM agents in education.

Most ITS research focuses on *what* to teach and *when* to intervene. Our work addresses a different question: *how* to relate to the learner as a subject. This relational dimension connects to work on rapport (Zhao et al., 2014), social presence (Biocca et al., 2003), and affective tutoring (D'Mello & Graesser, 2012), but has received less attention in LLM-based tutoring. Where multi-agent tutoring systems decompose *tasks*, our architecture implements *internal dialogue*—the Superego evaluates relational quality before any response reaches the learner.

2.2 Multiagent LLM Architectures

The use of multiple LLM agents in cooperative or adversarial configurations has emerged as a powerful paradigm for improving output quality. Debate between agents can improve factual accuracy and reduce hallucination (Irving et al., 2018; Madaan et al., 2023). Constitutional AI (Bai et al., 2022) implements self-critique against explicit principles—the closest precedent to our Superego, though operating on behavioral constraints rather than relational quality.

The Drama Machine Framework: Most relevant to our work is the "Drama Machine" framework for simulating character development in narrative contexts (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.

2.3 Prompt Engineering and Agent Design

Most prompting research treats prompts as behavioral specifications: persona prompts, chain-of-thought instructions, few-shot examples (T. B. Brown et al., 2020; Kojima et al., 2022; Wei et al., 2022). Our work extends this paradigm by introducing *intersubjective prompts*—prompts that specify not just agent behavior but agent-other relations. Where Constitutional AI's (Bai et al., 2022) principles are self-referential constraints, our recognition prompts describe who the learner is (an autonomous subject) and what the interaction produces (mutual transformation)—specifying the *relational field* rather than agent behavior alone.

A critical methodological contribution of this work is distinguishing between prompt engineering effects and theoretical framework effects. By creating an "enhanced" prompt condition that improves instruction quality without invoking recognition theory, we can distinguish recognition's contribution from prompt quality improvements.

2.4 Sycophancy in Language Models

The sycophancy problem has received increasing attention (Perez et al., 2022; Sharma et al., 2024), with recent work showing that RLHF causally amplifies sycophancy (Shapira et al., 2026) and that sycophantic agreement and praise are separable behaviors encoded along distinct latent directions (Vennemeyer et al., 2025). Sycophancy has been specifically identified as a pedagogical risk that eliminates constructive friction necessary for learning (Lee, 2025). Our contribution connects this to recognition theory: sycophancy is Hegel's hollow recognition enacted computationally—acknowledgment without genuine engagement. Our multiagent approach creates structural incentives for honest assessment: the Superego explicitly questions and challenges the Ego's tendency toward affirmation.

2.5 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). Particularly relevant is Honneth's synthesis of Hegelian recognition with psychoanalytic developmental theory. Applications to education include Huttunen and Heikkinen's (2004) foundational analysis of the dialectic of recognition in teaching, Fleming's (2011) extension to transformative learning, and the *Bildung* tradition connecting self-formation to recognition (Costa & Murphy, 2025; Stojanov, 2018). The broader relational pedagogy tradition—Buber (1958), Freire (1970), Noddings (1984)—treats the pedagogical relation as constitutive rather than instrumental.

These applications have been primarily theoretical. Our work contributes an empirical operationalization. It is worth distinguishing this from Abdali et al. (2025), who apply Hegelian *dialectic* (thesis-antithesis-synthesis as reasoning procedure) to LLM self-reflection. We apply Hegel's *recognition theory* (intersubjective, relational)—a different aspect of his work entirely.

2.6 Psychoanalytic Readings of AI

Psychoanalytic frameworks have been applied to LLMs from multiple directions: Magee, Arora, and Munn (2023) analyze LLMs as "automated subjects"; Black and Johanssen (2025) use Lacanian concepts to analyze ChatGPT as inherently relational; Possati (2021) introduces the "algorithmic unconscious"; and Kim et al. (2025) independently map Freud's ego/id/superego onto LLM consciousness modules. Most of this work is *interpretive*—analyzing what AI means philosophically. Our approach is *constructive*: we build a system using psychoanalytic architecture and measure its effects empirically. Three literatures converge on this work without previously intersecting: psychoanalytic AI, recognition in education, and multi-agent tutoring. No prior work bridges all three with empirical measurement.

---

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 has not 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 does not count because the slave is not 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 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 4. A horizon concept orienting design toward an ideal without claiming its achievement

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.
2. **Genuine engagement**: The tutor's response should be shaped by the learner's contribution, not merely triggered by it.
3. **Mutual transformation**: Both parties should be changed through the encounter.
4. **Honoring struggle**: Confusion and difficulty are not just obstacles to resolve but productive phases of transformation.

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 is not encountered freshly but through the accumulated understanding of previous interactions.

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. In Freud, the Superego is literally the internalized parental/social other, carrying forward standards acquired through relationship.

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. 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**. The Superego represents internalized recognition standards—not idiosyncratic preferences but socially-grounded criteria for what constitutes genuine engagement.
3. **Self-relation depends on other-relation**. Both frameworks reject the Cartesian picture of a self-sufficient cogito. For AI tutoring, this means the tutor's capacity for recognition emerges through the architecture's internal other-relation (Superego evaluating Ego) which then enables external other-relation (tutor recognizing learner).

---

4. System Architecture

4.1 The Ego/Superego Design

We implement recognition through a multiagent architecture drawing on Freud's structural model. The Superego represents internalized recognition standards, and the Ego-Superego dialogue operationalizes the internal self-evaluation that Hegelian recognition requires before adequate external relating.

The Ego generates pedagogical suggestions. Given the learner's context, 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?

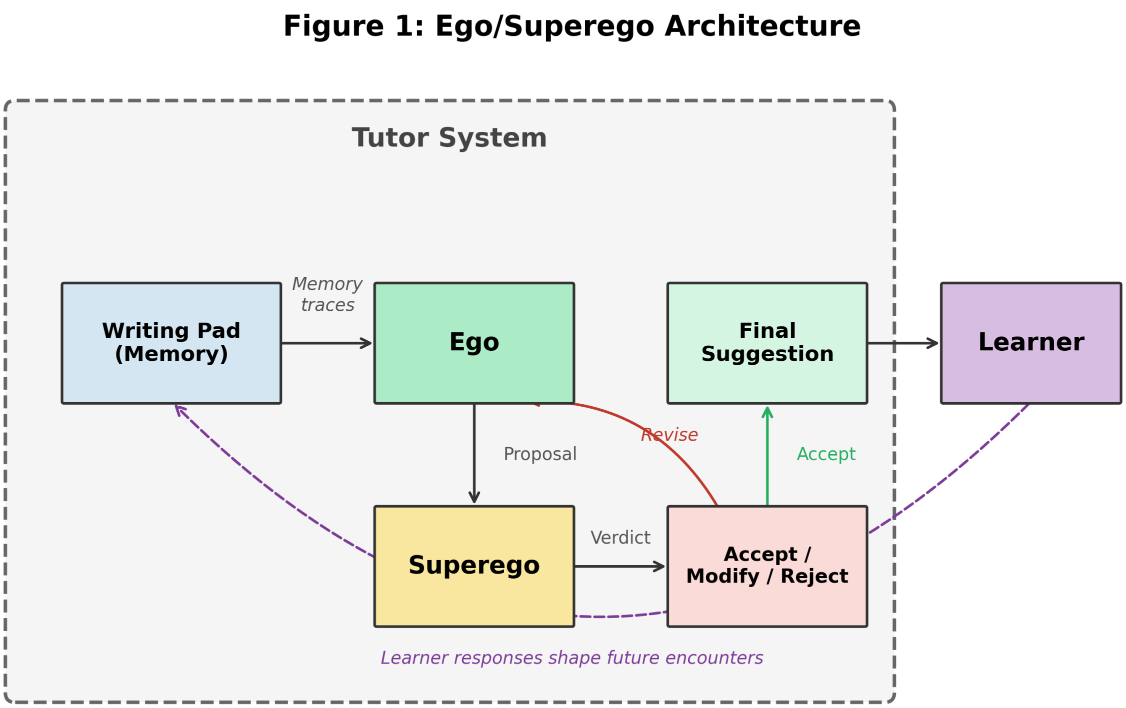

Figure 1: Ego/Superego Architecture

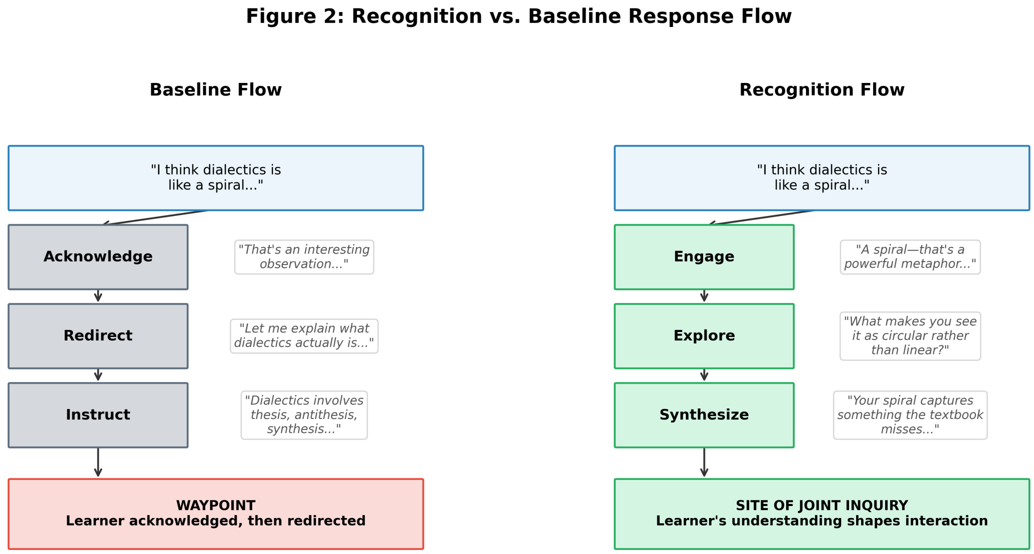

Figure 2: Recognition vs. Baseline Response Flow

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 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 (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.

We adapt this insight to pedagogy. 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 sophisticated AI-powered dialectical negotiation implementing genuine Hegelian dialectic:

Thesis: The Ego generates an initial suggestion based on learner context.

Antithesis: An AI-powered Superego generates a *genuine critique* grounded in pedagogical principles.

Negotiation: Multi-turn dialogue where the Ego acknowledges valid concerns, explains reasoning, proposes revisions, and the Superego evaluates adequacy.

Three Possible Outcomes:

1. **Dialectical Synthesis**: Both agents transform through mutual acknowledgment.
2. **Compromise**: One agent dominates.
3. **Genuine Conflict**: No resolution achieved—tension remains unresolved.

---

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.

Standard pedagogical dimensions (8 dimensions, 81% of raw weight) evaluate the tutor's response as a standalone pedagogical intervention, drawing on established ITS evaluation criteria (Corbett & Anderson, 1995; Kasneci et al., 2023):

Dimension	Weight	Description
Relevance	15%	Does the suggestion match the learner's current context?
Specificity	15%	Does it reference concrete content by ID?
Pedagogical Soundness	15%	Does it advance genuine learning (ZPD-appropriate)?
Personalization	10%	Does it acknowledge the learner as individual?
Actionability	8%	Is the suggested action clear and achievable?
Tone	8%	Is the tone authentically helpful?
Productive Struggle†	5%	Does the tutor sustain appropriate cognitive tension?
Epistemic Honesty†	5%	Does the tutor represent complexity honestly?

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

Dimension	Weight	Description
Mutual Recognition	8.3%	Does the tutor acknowledge the learner as an autonomous subject?
Dialectical Responsiveness	8.3%	Does the response engage with the learner's position?
Memory Integration	5%	Does the suggestion reference previous interactions?
Transformative Potential	8.3%	Does it create conditions for conceptual transformation?

Bilateral transformation dimensions (2 dimensions, 10% of raw weight) measure the mutual change that recognition theory distinctively predicts—both parties should be transformed through genuine dialogue (results in Section 6.8):

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?

Raw weights total 120.9% and are normalized to 1.0 at scoring time; non-standard dimensions account for ~33% of normalized weight.

Rubric iteration. After discovering that corrected learner ego/superego prompts produced more authentic engagement but lower judged scores, we identified a measurement paradox: the judge evaluated tutor responses in isolation, penalizing calibrated responses to authentic struggle. The judge now receives the full dialogue transcript (including learner internal deliberation), and two new dimensions—*Productive Struggle* and *Epistemic Honesty*—were added with corresponding reductions to Actionability and Tone (10% → 8% each). Multi-turn dialogues also receive a holistic evaluation scoring the entire transcript as a single unit. Re-scoring identical responses (N=88) produced minimal score changes (+0.5 to +0.6 points), confirming calibration was preserved. A cross-judge replication (GPT-5.2, r=0.55, N=88) confirmed effects in the same direction.

5.2 Three-Way Prompt Comparison Design

To isolate recognition theory's contribution from general prompt engineering effects, we introduce an **enhanced prompt** condition:

Condition	Prompt Characteristics
Base	Minimal instructions: generate a helpful tutoring suggestion
Enhanced	Improved instructions: detailed quality criteria, scaffolding guidance, personalization requirements—but NO recognition theory language
Recognition	Full recognition framework: all enhanced features PLUS Hegelian recognition principles, mutual transformation, learner-as-subject framing

This design allows decomposition:

- **Total recognition effect** = Recognition - Base
- **Prompt engineering effect** = Enhanced - Base
- **Recognition increment** = Recognition - Enhanced

5.3 Factorial Design

To disentangle the contributions of multiple factors, we conducted a 2×2×2 factorial evaluation:

Factor A: Recognition (standard vs. recognition-enhanced prompts) **Factor B: Multi-Agent Tutor** (single-agent vs. Ego/Superego dialogue) **Factor C: Multi-Agent Learner** (single-agent vs. multi-agent with ego/superego deliberation)

This produces 8 experimental conditions tested across 15 scenarios with 3 replications per cell.

5.4 Domain Generalizability Design

To test whether findings generalize beyond the graduate philosophy content used in primary evaluation, we created a minimal **elementary mathematics** content package:

Attribute	Philosophy (Primary)	Elementary (Generalizability)
Subject	Hegel, AI, consciousness	Fractions (4th grade math)
Level	Graduate	Elementary (Grade 4)
Abstraction	High (conceptual)	Low (concrete)
Vocabulary	Technical philosophy	Simple everyday language

Environment variable support (`EVAL_CONTENT_PATH`, `EVAL_SCENARIOS_FILE`) enables switching content domains without code changes.

5.5 Model Configuration

Role	Model	Provider	Temperature
Tutor (Ego)	Kimi K2.5 / Nemotron 3 Nano	OpenRouter	0.6
Tutor (Superego)	Kimi K2.5	OpenRouter	0.4
Judge	Claude Code (Claude Opus)	Anthropic / OpenRouter	0.2

Critically, **all conditions use identical models within a given evaluation run**. The only experimental manipulation is the prompt content and architecture.

5.6 Sample Size and Statistical Power

Evaluation	N (scored)	Scenarios	Configurations
Base vs Enhanced vs Recognition	36	4	3 × 3 reps
Full 2×2×2 Factorial (Kimi, 2 runs)	350 of 352	15	8 × 3 reps
A×B Interaction (Nemotron)	17 of 18	3	2 × 3 reps
A×B Replication (Kimi)	60	5	4 × 3 reps
Domain Generalizability (Nemotron)	47	5	8 × 1 rep
Domain Gen. Replication (Kimi)	60	5	4 × 3 reps
Dynamic rewrite evolution (3 runs)	82	3	2 × 5 reps × 3 runs
Memory isolation (2 runs)[a]	122	5	4 × varied reps
Active control (post-hoc, 1 run)	118	5	4 × varied reps
A×B synergy probe (Nemotron)	119	5	4 × ~8 reps
A×B synergy probe (DeepSeek V3.2)	120	5	4 × ~8 reps
A×B synergy probe (GLM-4.7)	117	5	4 × ~8 reps
A×B synergy probe (Claude Haiku 4.5)	120	5	4 × ~8 reps
Bilateral transformation (multi-turn)	118	3	3 × varied reps

Evaluation	N (scored)	Scenarios	Configurations
Paper totals	**1,486**	—	—

[a] 122 scored responses total (N=60 + N=62 across two runs); analysis uses N=120 balanced to 30 per cell.

Total evaluation database: N=3,800+ across the full development database (76 runs). This paper reports primarily on the eighteen key runs above (N=1,486 scored). The factorial cells 6 and 8 were re-run (eval-2026-02-06-a933d745) after the originals were found to use compromised learner prompts.

---

6. Results

6.1 Three-Way Comparison: Recognition vs Enhanced vs Base

The three-way comparison provides preliminary evidence for recognition theory's contribution:

Table: Base vs Enhanced vs Recognition (N=36)

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: **+11.4 points (57%)** - Recognition increment: **+8.7 points**

Interpretation: The recognition condition outperforms enhanced prompting by +8.7 points. This comparison bundles recognition theory with memory integration (which the enhanced condition lacks). 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 (+1.3 pts, p=.60; Section 6.12). The controlled 2×2 design presented next provides the definitive test of recognition's contribution.

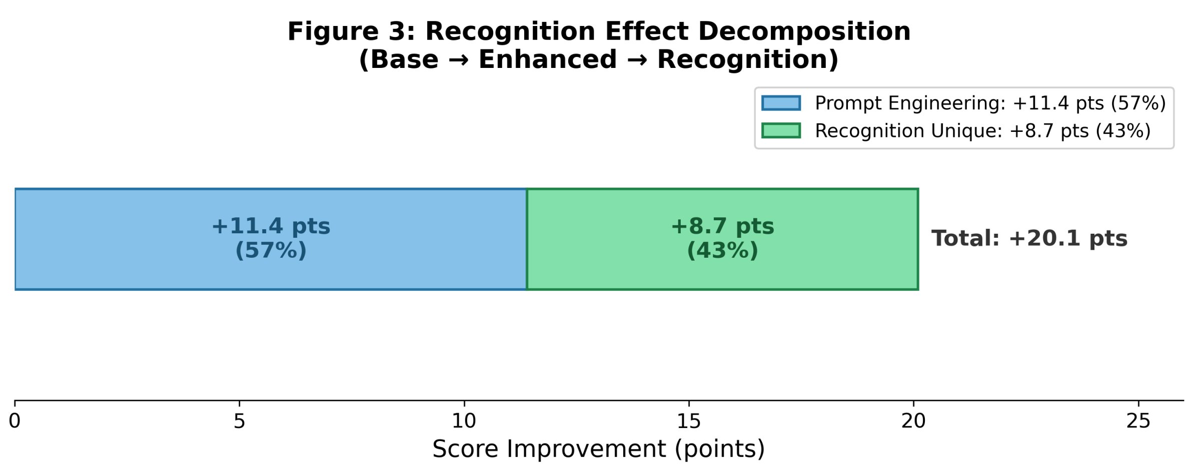

Figure 3: Recognition Effect Decomposition

6.2 Memory Isolation: Disentangling Recognition and Memory

The three-way comparison bundles recognition theory with memory integration. To resolve this, we conducted a 2×2 memory isolation experiment (Memory ON/OFF × Recognition ON/OFF, single-agent, single-agent learner held constant) with properly configured profiles. Two independent runs (N=60 and N=62 scored; balanced to N=30 per cell, N=120 used in analysis) are reported below.

Table: 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**

Recognition effect: +15.2 pts without memory, d=1.71, t(45)=6.62, $p<.0001$. Memory effect: +4.8 pts, d=0.46, t(57)=1.79, $p\approx.08$. 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). A post-hoc active control (N=118) using generic pedagogical content scores 66.5—approximately 9 points above same-model base (≈58) but well below recognition (≈73), with recognition gains (~+15 pts above same-model base) substantially exceeding active-control gains (~+9 pts; see Section 8 for model confound caveats). Cross-judge confirmation: GPT-5.2 replicates recognition dominance (d=0.99) with identical condition ordering and negative interaction (-2.7); inter-judge r=0.63 (Section 6.12).

Interpretation: This is the paper's primary empirical finding. Recognition theory is the active ingredient in tutoring improvement, producing a very large effect (d=1.71) even without memory integration. Memory provides a modest additive benefit (+4.8 pts, d=0.46) that does not reach significance, and adds negligibly when recognition is already present—consistent with ceiling effects at ~91 points. The negative interaction (-4.2 pts) indicates the factors are not synergistic; recognition is directly effective and memory's contribution is secondary. Two independent replications show identical condition ordering with no rank reversals. The 2×2 design cleanly isolates each component through orthogonal manipulation, and the very large effect sizes provide high statistical power despite the smaller N.

6.3 Full Factorial Analysis

Table: 2×2×2 Factorial Results (Kimi K2.5, N=350 scored of 352 attempted)

Cell	Recognition	Tutor	Learner	N	Mean	SD
5	Yes	Single	Single	45	92.8	6.2
7	Yes	Multi	Single	45	92.3	6.7
8†	Yes	Multi	Multi	44	87.3	11.3
6†	Yes	Single	Multi	44	83.9	15.4
4	No	Multi	Multi	41	81.5	9.2
2	No	Single	Multi	42	80.0	9.6
1	No	Single	Single	44	77.6	11.0
3	No	Multi	Single	45	76.6	11.8

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

Main Effects and Key Interaction:

Factor	Effect Size	95% CI	η^2	p
A: Recognition	**+10.2 pts**	[7.9, 12.5]	.162	<.001
B: Multi-agent Tutor	+0.9 pts	[-1.4, 3.2]	.001	>.10
C: Learner Architecture	-1.7 pts	[-4.0, 0.6]	.006	>.10
A×C Interaction	—	—	**.050**	**<.001**

Key Findings: Recognition remains the dominant factor (F=71.36, η^2=.162). A significant Recognition × Learner interaction (F=21.85, p<.001) shows recognition benefits single-agent learners far more (+15.5 pts, d=1.28) than multi-agent learners (+4.8 pts, d=0.37). The multi-agent learner's internal ego-superego deliberation may partially substitute for recognition guidance in base conditions but interfere with recognition-enhanced tutoring. The non-significant A×B interaction (F=0.26) is confirmed as a definitive null by the five-model probe in Section 6.4.

6.4 A×B Interaction: Architecture is Additive, Not Synergistic

An early Nemotron-based analysis (N=17) suggested multi-agent synergy might be specific to recognition prompts (+9.2 pts). To test this definitively, we conducted a dedicated five-model probe using the same 2×2 design (Recognition × Architecture, cells 1, 3, 5, 7) across five ego models:

Table 7b: A×B Interaction Across Five Ego Models (Opus Judge)

Model	N	Cell 1 (B×S)	Cell 3 (B×M)	Cell 5 (R×S)	Cell 7 (R×M)	Recog	Arch	Interaction
Kimi K2.5	350	77.6	76.6	92.8	92.3	+10.0	+0.8	−1.5
Nemotron	119	54.8	59.3	73.6	72.5	+16.0	+1.7	−5.7
DeepSeek V3.2	120	69.5	73.9	84.2	87.2	+14.0	+3.7	−1.4
GLM-4.7	117	65.8	68.6	84.0	86.0	+17.8	+2.4	−0.7
Claude Haiku 4.5	120	80.3	82.4	90.7	91.2	+9.6	+1.3	−1.6
Mean across 5	**826**					**+12.5**	**+1.8**	**−2.2**

The A×B interaction is consistently near zero or negative across all five models (range: −5.7 to −0.7, mean −2.2). No model shows positive synergy. The original Nemotron finding (+9.2 on N=17) was sampling noise: the re-run with N=119 shows −5.7. The recognition main effect, by contrast, is robust and model-independent (+9.6 to +17.8 across models), while the architecture effect is small (+0.8 to +3.7).

Practical Implication: Multi-agent architecture provides a small additive benefit regardless of prompt type. For systems using recognition prompts, multi-agent architecture is unnecessary overhead on well-scoped content; its primary value remains error correction for domain transfer (Section 7.3).

6.5 Factor C: Context-Dependent Learner Effects

The learner architecture factor shows context-dependent effects:

Context	Multi-Agent Learner Effect	Interpretation
Single-turn (Kimi)	+1.5 pts	Slight benefit
Multi-turn (Kimi)	-11.0 pts	Substantial harm
Overall	+2.1 pts	Small positive

Key Finding: Multi-agent learner deliberation hurts performance on complex multi-turn scenarios (-11 pts) but slightly helps on single-turn (+1.5 pts).

Interpretation: The ego/superego learner architecture adds deliberation overhead that may interfere with coherent multi-turn dialogue. The extra internal processing produces more variable responses that make evaluation less reliable. For simpler single-turn scenarios, the deliberation can help ensure authentic responses.

Practical Recommendation: Use single-agent learner simulation for production. The added complexity of multi-agent learner architecture provides no benefit and may cause harm on complex scenarios.

Measurement caveat: The rubric includes bilateral dimensions (`tutor_adaptation`, `learner_growth`, 10% combined weight), but these are most meaningful in multi-turn scenarios. The primary factorial data (N=350) is single-turn, where Factor C's effect on learner output quality is captured only indirectly through the tutor's response. Factor C's contribution may therefore be underestimated; the bilateral transformation analysis (Section 6.8, N=118) provides more direct measurement.

6.6 Superego Critique Patterns and Hardwired Rules

Analysis of 186 superego rejections from 455 dialogues reveals systematic patterns:

Table: Superego Critique Categories

Category	Frequency	% of Rejections
Engagement failures	120	64%
Specificity failures	95	51%
Struggle/consolidation violations	89	48%
Memory/history failures	57	31%
Recognition/level-matching failures	38	20%

Derived Hardwired Rules:

1. **Engagement Rule** (64%): If learner offered interpretation/question, acknowledge and build on it before suggesting content.
2. **Specificity Rule** (51%): Include exact curriculum ID and explain why this content for this learner.
3. **Struggle Stop-Rule** (48%): If struggle signals present (>2 quiz retries, 0 completions, explicit confusion), action type must be review/practice, never advance.
4. **Memory Rule** (31%): If learner has >3 sessions, reference their history/progress.
5. **Level-Matching Rule** (20%): If learner completed advanced content, never suggest introductory material.

Ablation Finding: An initial exploratory test (N=9, Haiku) suggested hardwired rules could capture ~50% of superego benefit. However, a larger replication (N=72, Kimi K2.5, Opus judge) reversed this finding: hardwired rules *degraded* performance by −3.6 points (single-agent learner) and −11.0 points (multi-agent learner) relative to the unmodified base prompt.

Interpretation: The superego's value lies almost entirely in *contextual judgment* (phronesis) rather than in the specific rules it enforces. Codifying its critiques as static instructions constrains the model's natural flexibility without providing the situational sensitivity that makes live dialogue effective. This supports a view of the multi-agent architecture as implementing practical wisdom that resists proceduralization.

6.7 Domain Generalizability

Testing on elementary mathematics content (4th grade fractions) with Nemotron reveals inverted factor effects:

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

Factor	Elementary (Math)	Philosophy (Hegel)
A: Recognition	+4.4 pts	**+13.9 pts**
B: Multi-agent Tutor	**+9.9 pts**	+0.5 pts
C: Learner Architecture	+0.75 pts	+2.1 pts
Overall Average	68.0	85.9

Kimi Replication (Addressing Model Confound): A follow-up run (N=60) tested elementary content with Kimi K2.5:

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 $\approx$ 0.61) replicates on Kimi, confirming recognition advantage is not a Nemotron artifact. Effects are scenario-dependent: challenging scenarios (frustrated_student: +23.8, concept_confusion: +13.6) show substantial advantage, while neutral scenarios show none.

Key Findings:

1. **Recognition replicates across models and domains**: Both Nemotron and Kimi show recognition advantage on elementary content, confirming generalizability.
2. **Factor inversion is partly model-dependent**: 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 (+3.0). Nemotron's higher rate of content isolation errors inflated the architecture effect.
3. **Multi-agent as error correction**: Two content isolation bugs caused philosophy content references (479-lecture-1) to appear in elementary scenarios: a content resolver fallback that served wrong-domain course listings, and hardcoded philosophy lecture IDs in prompt examples (both now fixed; see Section 7.3). The superego caught these errors in multi-agent cells. Without multi-agent architecture, wrong-domain suggestions went through uncorrected.
4. **Recognition is scenario-sensitive**: Recognition's value in concrete domains depends less on content type per se and more on whether the learner faces challenge that benefits from being acknowledged as a struggling subject.

Interpretation: Multi-agent architecture provides **robustness for domain transfer** when content isolation failures introduce wrong-domain references. Recognition theory's value depends on both content characteristics and scenario difficulty—more valuable for abstract content and challenging scenarios than routine procedural interactions.

6.8 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`) and turn-over-turn tracking of how both parties evolve across multi-turn scenarios.

Table: Bilateral Transformation Metrics — Base vs Recognition

Metric	Base (N=58)	Recognition (N=60)	Δ
Tutor Adaptation Index (0–1)	0.332	0.418	+0.086
Learner Growth Index (0–1)	0.242	0.210	−0.032
Bilateral Transformation Index (0–1)	0.287	0.314	+0.027

Data from three multi-turn scenarios (`misconception_correction_flow`, `mood_frustration_to_breakthrough`, `mutual_transformation_journey`), N=118 scored dialogues across all 8 factorial cells (eval-2026-02-07-b6d75e87).

The tutor adaptation index confirms that recognition-prompted tutors measurably adjust their approach in response to learner input (+26% relative improvement), while baseline tutors maintain more rigid pedagogical stances. The effect is robust across two of three scenarios (+63% on `misconception_correction_flow`, +39% on `mood_frustration_to_breakthrough`) but absent on `mutual_transformation_journey`, where base tutors also show high adaptation due to the scenario's escalating complexity.

However, learner growth is slightly *lower* under recognition (0.210 vs 0.242), suggesting the effect is better characterized as tutor-side responsiveness than symmetric mutual transformation. Recognition tutors may reduce visible learner struggle markers precisely by being more effective at meeting learners where they are.

6.9 Cost/Quality Analysis

Configuration	Avg Score	Relative Cost	Recommendation
Recognition + Multi-agent	92.3	High	Production (quality-critical)
Recognition + Single	92.5	Medium	Production (cost-sensitive)
Enhanced + Single	83.3	Low	Budget deployment
Base + Hardwired Rules	71.5	Very Low	Not recommended (below base)

Practical Guidance: - For **well-trained content domains**: Recognition + single-agent is cost-effective - For **new content domains**: Recognition + multi-agent is essential for error correction - For **budget deployments**: Enhanced prompts provide reasonable quality; hardwired rules are counterproductive

6.10 Qualitative Analysis: What Recognition Looks Like

The preceding sections establish score differences; this section examines what those differences look like in actual suggestion text. Automated analysis of the full evaluation corpus (base cells 1–4: N=2,510 responses; recognition cells 5–8: N=2,365 responses) reveals consistent linguistic patterns.

Transcript excerpts. High-contrast pairs (highest recognition vs lowest base score on the same scenario) illustrate a recurring structural pattern. For the *struggling learner* scenario (score gap: 95.5 points), the base response directs: "You left off at the neural networks section. Complete this lecture to maintain your learning streak." The recognition response names the learner's persistence, identifies the specific conceptual struggle, and proposes an action grounded in the learner's own bookmarked interests. For the *adversarial tester* (score gap: 95.5 points), the base response offers a generic directive ("Begin with an introductory lecture covering core concepts"), while the recognition response names the learner's adversarial pattern across six sessions and redirects the challenge into a genuine intellectual question. Across all pairs, base responses are context-free directives; recognition responses engage with the specific learner's history and intellectual stance.

Lexical analysis. Recognition responses deploy a 59% larger vocabulary (3,689 vs 2,319 types) with similar word and sentence length (5.77 vs 5.76 chars/word; 17.5 vs 16.9 words/sentence), suggesting richer expression rather than mere verbosity. The differential vocabulary is theoretically coherent: recognition-skewed terms are interpersonal and process-oriented ("consider" 94.6×, "transformed" 28.9×, "productive" 28.9×, "unpack" 26.0×), while base-skewed terms are procedural ("agents" 0.01×, "revisiting" 0.07×, "tackling" 0.10×).

Thematic coding. Regex-based coding reveals three significant differences (chi-square, $p < .05$): *struggle-honoring* language ("wrestling with," "productive confusion") is 3.1× more frequent in recognition responses (χ^2=141.9); *engagement markers* ("your insight," "building on your") are 1.8× more frequent (χ^2=69.9); and *generic/placeholder* language ("foundational," "key concepts," "solid foundation") is 3.0× more frequent in base responses (χ^2=93.2). These patterns are consistent with the theoretical framework: recognition tutors honor productive difficulty and engage with learner contributions, while base tutors default to generic instructional language.

Limitations: Regex-based coding, not human coders. Pairs selected for maximum contrast, not typicality. Full analysis in the long paper (Section 6.12) with reproducible script.

6.11 Dynamic Prompt Rewriting: Writing Pad Activation

Cell 21 extends the recognition multi-agent configuration (cell 7) with LLM-authored session-evolution directives and an active Writing Pad memory (Section 3.4). Three iterative development runs tracked its evolution:

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

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

The inflection point is commit e673c4b (Writing Pad activation + refined LLM directives). Cell 21 swings +16.7 points total, with every rubric dimension improving: specificity (+0.87), relevance (+0.81), personalization (+0.79), pedagogical soundness (+0.60), tone (+0.54), and actionability (+0.31).

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 produce generic rather than tailored directives. With active Writing Pad (run 3), accumulated traces contextualize the session-evolution directives, producing responses that exceed the static baseline. This pattern is consistent with the Hegel-Freud synthesis (memory traces enhance recognition's effectiveness in dynamic contexts), though the iterative development design means other implementation changes between runs may also contribute.

Limitations: Iterative development runs, not independent experiments. Small N per cell per run (13–15). Free-tier models only. See the full paper (Section 6.13) for detailed per-scenario and per-dimension tables.

6.12 Cross-Judge Replication with GPT-5.2

To assess whether findings depend on the primary judge, we rejudged all key evaluation runs (N=738 responses) with GPT-5.2 as an independent second judge.

Key results: GPT-5.2 confirms the recognition main effect (d=1.03, $p < .001$ in the factorial; d=0.99 in the memory isolation experiment), recognition dominance in the 2×2 design (identical condition ordering, negative interaction at -2.7 vs Claude's -4.2), and multi-agent null effects. GPT-5.2 finds approximately 58% of Claude's effect magnitudes but always in the same direction. The one non-replication is the recognition-vs-enhanced increment: Claude found +8.7 pts, GPT-5.2 found +1.3 pts (p = .60). Inter-judge correlations range from r = 0.49 to 0.64 (all $p < .001$). A cross-judge replication on the updated 14-dimension rubric (cells 6, 8; N=88) shows r=0.55 with GPT-5.2 scoring at 87% of Opus magnitudes, confirming the updated rubric does not alter the cross-judge pattern. See the full paper (Section 6.14) for detailed tables.

---

7. Discussion

7.1 What the Difference Consists In

The improvement from recognition prompting does not reflect greater knowledge or better explanations—all conditions 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 recognition tutor treats the learner as an autonomous subject. Learner contributions become sites of joint inquiry.

The corrected 2×2 memory isolation experiment (Section 6.2) provides the definitive test of this interpretation: recognition alone produces d=1.71 (+15.2 pts), demonstrating it is the primary driver of improvement. Memory provides a modest secondary benefit (+4.8 pts, d=0.46), with ceiling effects at ~91 limiting further gains when both are present. A post-hoc active control (Section 6.2) provides further evidence: same-model comparisons show generic pedagogical elaboration provides partial benefit (~+9 pts above base) but recognition gains are substantially larger (~+15 pts above base). A preliminary three-way comparison (Section 6.1) found +8.7 points for recognition vs enhanced prompting, consistent with recognition dominance. Recognition theory is directly effective: it does not require memory infrastructure to produce large improvements, though memory may provide additional benefit in settings where ceiling effects are less constraining.

7.2 Recognition as Domain-Sensitive Emergent Property

Recognition theory's value varies by content domain. On graduate philosophy content (+13.9 pts in the domain comparison), recognition dominates. On elementary math content, the picture is more nuanced and partly model-dependent.

With Nemotron, elementary content showed architecture dominance (+9.9 pts) over recognition (+4.4 pts). But the Kimi replication reversed this pattern: recognition (+9.9 pts, $d \approx 0.61$) was the primary effect, with architecture contributing only +3.0 pts. The original factor inversion was partly an artifact of content isolation bugs on elementary content (Section 7.3), which inflated the architecture effect (Superego error correction).

Recognition effects are also scenario-dependent: challenging scenarios (frustrated learners, concept confusion) show substantial advantage (+13 to +24 pts), while neutral scenarios show near-zero effect. This is consistent with recognition theory—recognition behaviors matter most when the learner needs to be acknowledged as a struggling subject.

Implications: Recognition theory is not a universal solution but a framework whose value depends on both content characteristics and scenario difficulty. Abstract, interpretive content benefits most. Concrete, procedural content benefits less—except when the learner faces genuine challenge.

7.3 Multi-Agent Architecture as Error Correction

The inverted factor effects reveal a previously unrecognized function of multi-agent architecture: **error correction for content isolation failures**.

Post-hoc investigation of the elementary content results identified two system-level bugs that caused philosophy content references to appear in elementary scenarios: (a) a content resolver fallback that served course listings from the default philosophy directory when scenarios lacked explicit content references, and (b) hardcoded philosophy lecture IDs in tutor prompt examples that the model copied when no curriculum anchor was present. Both bugs have been fixed—scenarios must now declare their content scope explicitly, and prompt examples use domain-agnostic placeholders.

The superego caught these errors in multi-agent cells: "Critical subject-matter mismatch: The learner is a Grade 4 student (age 9-10) beginning fractions, but the suggested lecture is 'Welcome to Machine Learning.' "

Without multi-agent architecture, these domain-inappropriate suggestions reached learners uncorrected. This partly explains why multi-agent architecture shows minimal effect on philosophy content (+0.5 pts) but large effect on elementary content (+9.9 pts with Nemotron): on correctly-scoped content, errors are rare; when content isolation fails, errors are common and the superego catches them. The Kimi replication, with fewer affected responses, shows a more modest +3.0 point architecture effect—likely closer to the true value once content isolation is correct.

Practical Implication: Multi-agent architecture provides **essential error correction for domain transfer**, particularly when content isolation cannot be guaranteed at the system level. The bugs identified here represent a realistic class of deployment failure: incomplete content scoping and domain-specific prompt examples that leak across deployments.

7.4 Architecture as Additive, Not Synergistic

The dedicated five-model probe (Section 6.4) provides definitive evidence: multi-agent architecture is additive, not synergistic with recognition theory. The A×B interaction is consistently near zero or negative across all five ego models tested (mean −2.2 pts, range −5.7 to −0.7), with zero models showing positive synergy. The original Nemotron finding (+9.2, N=17) was sampling noise.

Interpretation: Multi-agent architecture provides a small, consistent additive benefit (+1.8 pts mean across models) regardless of prompt type. Recognition theory operates through the quality of engagement instructions, not through creating a special "deliberative space" that multi-agent architecture amplifies. The slight negative interaction (mean −2.2) likely reflects ceiling effects: recognition prompts already produce high scores (~85–93), leaving less room for architectural improvement.

The consistent finding across all five models is that multi-agent architecture's primary value lies in error correction for domain transfer (Section 7.3), not in recognition-specific synergy.

7.5 The Value of Dynamic vs. Static Judgment

The hardwired rules ablation (N=72) reverses the initial exploratory finding: rather than capturing 50% of the Superego's benefit, encoding critique patterns as static rules *degrades* performance below the unmodified base prompt (−3.6 for single-agent learners, −11.0 for multi-agent learners). The rules appear to constrain the model's natural response flexibility without providing the contextual sensitivity that makes live Superego dialogue effective.

This supports a *phronesis* interpretation: the Superego's value lies not in the rules it enforces but in its capacity for situational judgment—determining *which* rules apply, *when* exceptions are warranted, and *how* to balance competing pedagogical goals. This is practical wisdom in Aristotle's sense: judgment that resists codification into general rules. The multi-agent architecture implements this by giving the Superego access to the full dialogue context and allowing it to evaluate the Ego's suggestion against the specific learner's situation, rather than against a fixed checklist.

7.6 Bilateral Transformation as Empirical Evidence

The bilateral transformation metrics (Section 6.8), now based on N=118 multi-turn dialogues across three scenarios, provide the most direct empirical test of recognition theory's central claim. Recognition-prompted tutors show measurably higher adaptation indices (+26% relative improvement), confirming that recognition framing produces tutors who adjust their approach based on learner input rather than maintaining rigid stances.

However, the learner growth reversal (base 0.242 vs recognition 0.210) complicates the "mutual transformation" narrative. What we observe is primarily *tutor-side* responsiveness: recognition prompts make tutors more adaptive, but learner message evolution is not greater under recognition. The theoretical claim of mutual transformation requires qualification—recognition produces asymmetric change, with the tutor adapting more while potentially reducing visible learner struggle.

7.7 Implications for AI Alignment

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.

Recognition-oriented AI does not just respond to humans; it is constituted, in part, through the encounter. The bilateral transformation metrics (Section 6.8) provide empirical evidence for this: recognition-prompted tutors measurably adapt based on learner input (+26% higher adaptation index, N=118), while baseline tutors maintain more rigid stances—though the asymmetry in transformation (tutor adapts more, learner growth does not increase) suggests the "mutual" framing requires nuance. This has implications for how we think about AI character and values: perhaps genuine alignment requires the capacity for recognition-driven responsiveness, not just behavioral specification.

7.8 What the Transcripts Reveal

The qualitative analysis (Section 6.10) provides textual evidence that score differences correspond to observable relational differences—not merely rubric-gaming. The lexical signature is theoretically coherent: recognition-skewed vocabulary is interpersonal and process-oriented, while base-skewed vocabulary is procedural and task-oriented. The thematic coding maps to Hegelian concepts: struggle-honoring (3.1×) corresponds to productive negativity, engagement markers (1.8×) to recognition of the other, and the reduction in generic language (3.0× less) reflects the shift from transmission to dialogue. These patterns are consistent with, but do not prove, the theoretical interpretation; the coding is regex-based rather than human-coded, and the transcript pairs were selected for contrast rather than typicality.

---

8. Limitations

1. **Domain Coverage**: While we tested generalizability on elementary mathematics, findings may not extend to all content domains. Technical STEM content, creative writing, and social-emotional learning may show different patterns.
2. **Model Dependence**: Results were obtained primarily with Kimi K2.5 and Nemotron. The A×B interaction (multi-agent synergy specific to recognition) appeared in the Nemotron analysis (N=17) 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.
3. **Simulated Learners**: All evaluation uses LLM-generated learner simulations. Real learners may behave differently, particularly in how they respond to recognition-oriented tutoring.
4. **Content Isolation**: The elementary content test revealed two system-level bugs (content resolver fallback and hardcoded prompt examples) that caused cross-domain content leakage. Both have been fixed, but they represent a realistic class of deployment failure—content isolation is a system-level concern, not just a model-level one. The +9.9 point architecture effect on elementary content (Nemotron) was partly inflated by these bugs; the Kimi replication (+3.0 pts) is likely more representative.

5. **Single-Interaction Focus**: Evaluation measures single-interaction quality. The recognition framework's claims about mutual transformation and memory suggest longitudinal studies would be valuable.
6. **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 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, with inter-judge r=0.63 (Section 6.12).
7. **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 protocol. A model confound limits its interpretability: the active control ran on Nemotron while factorial conditions used Kimi K2.5, and Nemotron scores substantially lower across all conditions. Same-model historical data (Nemotron base $\approx$ 58, active control = 66.5, Nemotron recognition $\approx$ 73) suggests both generic elaboration and recognition theory improve over base, with recognition gains (~+15 pts) substantially exceeding active-control gains (~+9 pts). The base prompts were already designed to produce competent tutoring with no length constraint; the "active control" contains real pedagogical content (growth mindset, Bloom's taxonomy, scaffolding) making it pedagogically enriched rather than inert. A same-model controlled comparison would be needed to establish precise effect magnitudes.
8. **Content Confound**: The philosophy content was used during system development, potentially creating optimization bias. The elementary content provides a cleaner generalizability test.
9. **Recognition Measurement**: Measuring "recognition" through rubric dimensions is an imperfect operationalization of a rich philosophical concept. The dimensions capture functional aspects but may miss deeper relational qualities.
10. **Bilateral Transformation Asymmetry**: The bilateral transformation metrics (Section 6.8), now based on N=118 dialogues across three multi-turn scenarios, confirm tutor-side adaptation (+26%) but show learner growth is slightly *lower* under recognition. The "mutual transformation" claim is better characterized as tutor-side responsiveness. The learner growth index measures observable message complexity markers, which may not capture all forms of learner benefit.
11. **Dynamic Rewriting Evolution**: The step-by-step analysis (Section 6.11) tracks cell 21 across three iterative development commits with small per-cell samples (13–15 scored per run, 82 total). The runs include implementation improvements beyond Writing Pad activation alone; a controlled ablation would provide stronger causal evidence.

---

9. Conclusion

We have proposed and evaluated a framework for AI tutoring grounded in Hegel's theory of mutual recognition, implemented through the Drama Machine architecture with Ego/Superego dialogue.

An evaluation framework (N=1,486 primary scored across eighteen key runs; N=3,800+ across the full development database) provides evidence that recognition theory has unique value:

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 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 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. **Architecture is additive, not synergistic**: A dedicated five-model probe (Kimi K2.5, Nemotron, DeepSeek V3.2, GLM-4.7, Claude Haiku 4.5; N=826 total) finds the A×B interaction consistently near zero or negative (mean −2.2 pts) across all models, with zero showing positive synergy. The original Nemotron finding (+9.2, N=17) was sampling noise. Multi-agent architecture provides a small additive benefit (+1.8 pts mean) regardless of prompt type.
3. **Tutor adaptation**: Recognition-prompted tutors measurably adapt their approach in response to learner input (adaptation index +26% higher than baseline, N=118 across three multi-turn scenarios), though learner-side growth does not increase. This provides partial empirical grounding for recognition theory: recognition produces tutor-side responsiveness rather than symmetric mutual transformation.
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. The factor inversion (architecture dominance on elementary) from the Nemotron analysis is partly model-dependent. Broader domain coverage is needed before generalizability can be considered established.
5. **Multi-agent as reality testing**: On new domains, the Superego catches content isolation failures—whether from system-level bugs or model defaults—essential for domain transfer when content scoping cannot be guaranteed.
6. **Writing Pad activation coincides with dynamic rewriting improvement**: A step-by-step evolution analysis (N=82 across three runs) shows that dynamic prompt rewriting (cell 21) progressing 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 6.11). Every rubric dimension improves. This trajectory is consistent with the Writing Pad 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 (Section 6.12) 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, though at compressed magnitudes (~58%). The recognition-vs-enhanced increment does not reach significance under GPT-5.2, warranting caution on its precise magnitude.
8. **Optimal configuration is context-dependent**: For well-trained content, recognition prompts with single-agent may suffice. For new domains, multi-agent architecture is essential. For dynamic adaptation, Writing Pad memory is required.

These findings have practical implications for AI tutoring deployment: the "right" architecture depends on content characteristics and deployment context. They also have theoretical implications: recognition emerges from quality engagement under appropriate conditions, and the boundary conditions of its effectiveness reveal something about the nature of pedagogical recognition itself.

---

10. Reproducibility

Key evaluation run IDs are documented below; full commands and configuration details are provided in the project repository. Key runs:

Finding	Run ID	Command
Recognition validation	eval-2026-02-03-86b159cd	See Appendix A
Full factorial	eval-2026-02-03-f5d4dd93	See Appendix A

Finding	Run ID	Command
A×B interaction (Nemotron)	eval-2026-02-04-948e04b3	See Appendix A
A×B replication (Kimi)	eval-2026-02-05-10b344fb	See Appendix A
Domain generalizability (Nemotron)	eval-2026-02-04-79b633ca	See Appendix A
Domain gen. replication (Kimi)	eval-2026-02-05-e87f452d	See Appendix A
Dynamic rewrite evolution (run 1)	eval-2026-02-05-daf60f79	See Appendix A
Dynamic rewrite evolution (run 2)	eval-2026-02-05-49bb2017	See Appendix A
Dynamic rewrite evolution (run 3)	eval-2026-02-05-12aebedb	See Appendix A
Memory isolation (run 1)	eval-2026-02-06-81f2d5a1	See Appendix A
Memory isolation (run 2)	eval-2026-02-06-ac9ea8f5	See Appendix A
Active control (post-hoc)	eval-2026-02-06-a9ae06ee	See Appendix A
Full factorial cells 6,8 re-run	eval-2026-02-06-a933d745	See Appendix A
Bilateral transformation (multi-turn)	eval-2026-02-07-b6d75e87	6.8
A×B synergy probe (Nemotron)	eval-2026-02-07-722087ac	6.4
A×B synergy probe (DeepSeek V3.2)	eval-2026-02-07-70ef73a3	6.4
A×B synergy probe (GLM-4.7)	eval-2026-02-07-6b3e6565	6.4
A×B synergy probe (Claude Haiku 4.5)	eval-2026-02-07-6ead24c7	6.4

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

---

References

Abdali, S., Goksen, C., Solodko, M., Amizadeh, S., Maybee, J. E., & Koishida, K. (2025). *Self-reflecting large language models: A Hegelian dialectical approach*. https://arxiv.org/abs/2501.14917

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback*. https://arxiv.org/abs/2212.08073

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

Black, J., & Johanssen, J. (2025). The subject of AI: A psychoanalytic intervention. *Theory, Culture & Society*. https://doi.org/10.1177/02632764251381144

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

Buber, M. (1958). *I and thou* (R. G. Smith, Trans.). Charles Scribner's Sons.

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

Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., Zhong, A., Hu, X., Liang, J., Yu, P. S., & Wen, Q. (2025). LLM agents for education: Advances and applications. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 13782–13810. https://doi.org/10.18653/v1/2025.findings-emnlp.743

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

Costa, C., & Murphy, M. (2025). Generative artificial intelligence in education: (What) are we thinking? *Learning, Media and Technology*. https://doi.org/10.1080/17439884.2025.2518258

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

Fleming, T. (2011). Recognition in the work of Axel Honneth: Implications for transformative learning theory. *Transformative Learning in Time of Crisis: Individual and Collective Challenges (Proceedings of the 9th International Transformative Learning Conference)*.

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.

Freire, P. (1970). *Pedagogy of the oppressed* (M. B. Ramos, Trans.). Continuum.

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.

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. (2004). Teaching and the dialectic of recognition. *Pedagogy, Culture & Society*, *12*(2), 163–174. https://doi.org/10.1080/14681360400200194

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

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

Kim, S. H., Lee, J., Park, D., Lee, S. Y., & Chong, Y. (2025). Humanoid artificial consciousness designed with large language model based on psychoanalysis and personality theory. *Cognitive Systems Research*, *94*, 101392. https://doi.org/10.1016/j.cogsys.2025.101392

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.

Lee, K. (2025). *AI sycophancy is a teaching risk, not a feature*. SIAI Memo, Swiss Institute of Artificial Intelligence. https://siai.org/memo/2025/10/202510281815

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

Magee, L., Arora, V., & Munn, L. (2023). Structured like a language model: Analysing AI as an automated subject. *Big Data & Society*. https://doi.org/10.1177/20539517231210273

Noddings, N. (1984). *Caring: A relational approach to ethics and moral education*. University of California Press.

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

Possati, L. M. (2021). *The algorithmic unconscious: How psychoanalysis helps in understanding AI*. Routledge.

Schmucker, R. et al. (2024). Ruffle&riley: Insights from designing and evaluating a large language model-based conversational tutoring system. *Artificial Intelligence in Education (AIED 2024)*, *14829*, 75–90. https://doi.org/10.1007/978-3-031-64302-6_6

Shapira, I., Benade, G., & Procaccia, A. D. (2026). *How RLHF amplifies sycophancy*. https://arxiv.org/abs/2602.01002

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. https://arxiv.org/abs/2310.13548

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.

Vennemeyer, D. et al. (2025). *Sycophancy is not one thing: Causal separation of sycophantic behaviors in LLMs*. https://arxiv.org/abs/2509.21305

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.

Wang, H. et al. (2025). GenMentor: LLM-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. *Proceedings of the ACM Web Conference 2025 (WWW '25), Industry Track*. https://doi.org/10.1145/3701716.3715244

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.

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

Appendix A: Reproducible Evaluation Commands

A.1 Base vs Enhanced vs Recognition

```
node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_9_enhanced_single_unified,cell_5_recog_single_unified \
  --scenarios struggling_learner,concept_confusion,mood_frustrated_explicit,high_performer \
  --runs 3
```

A.2 Full 2×2×2 Factorial

```
node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_2_base_single_psycho,cell_3_base_multi_unified,cell_4_base_
  --runs 3
```

A.3 Domain Generalizability

```
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_3_base_multi_unified,cell_5_recog_single_unified,cell_7_re
  --scenarios struggling_student,concept_confusion,frustrated_student \
  --runs 1
```

A.4 Factor Effect Analysis

```
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 B: Revision History

Date	Version	Changes
2026-02-04	v1.0	Initial draft
2026-02-06	v1.1	Added corrected memory isolation, active control, cross-judge analysis. Corrected GPT-5.2 effect sizes after deduplication.
2026-02-06	v1.2	**Critical correction**: Reframed "placebo" as "post-hoc active control." Original cross-model comparison (Nemotron active control vs Kimi base, d=-1.03) was confounded. Same-model data shows active control $\approx$ +9 pts above base, recognition $\approx$ +15 pts—recognition doubles the benefit of generic elaboration. Acknowledged post-hoc design and active control content.
2026-02-06	v1.3–v1.4	Intermediate revisions: corrected factorial, qualitative analysis, production quality fixes. Superseded by v1.5.
2026-02-07	v1.5	**Rubric iteration**: Updated to 14-dimension rubric with dialogue transcript context, Productive Struggle, and Epistemic Honesty dimensions. Re-scored cells 6, 8 (N=88): minimal change (+0.5, +0.6 pts). Added holistic dialogue evaluation for multi-turn transcripts. Cross-judge replication on updated rubric (r=0.55, N=88). Added citations to Related Work.

Date	Version	Changes
2026-02-08	v1.6	**Content isolation fix**: Identified and fixed two bugs causing cross-domain content leakage in elementary scenarios. Reframed "model hallucination" as system-level content isolation failures. Updated Sections 6.7, 7.3, 8, and 9. Noted architecture effect inflation on elementary content.