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

Liam Magee — Education Policy, Organization and Leadership, University of Illinois Urban

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=645 primary scored responses across nine key runs, plus N=120 in a corrected 2×2 memory isolation experiment and N=120 in placebo controls; 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≈.08). The combined condition yields the highest scores (91.2, d=1.81 vs base), with ceiling effects limiting observable synergy. Length-matched placebo controls (N=120) rule out prompt length as an explanation: placebo conditions score ~65-67, comparable to base (~75), confirming that recognition theory's specific content drives the improvement. 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 model hallucinations. A step-by-step evolution analysis of dynamic prompt rewriting with active Writing Pad memory (N=83 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 placebo controls confirming prompt length is not the mechanism; (2) An exploratory analysis of multi-agent synergy (+9.2 pts, Nemotron, N=17) 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; (3) Domain generalizability testing confirms recognition advantage replicates across both models and content domains—elementary math with Kimi shows +9.9 pts (d ≈ 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 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 cross-judge replication with GPT-5.2 confirms the main findings are judge-robust: the recognition effect (d=1.03) and multi-agent null effects replicate, though at compressed magnitudes (~50% of primary judge effect sizes). Cross-judge confirmation of the corrected memory isolation results is pending.

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 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.

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, 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. Placebo controls (N=120) confirm prompt length is not the mechanism.

3. **Robust Factorial Evaluation**: A 2×2×2 factorial design (N=645 primary scored across nine key runs, plus N=120 in the memory isolation experiment and N=120 in placebo controls; N=3,800+ across the full development database) across multiple models, scenarios, and conditions, providing statistically robust effect estimates.

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**: Exploratory evidence that multi-agent synergy may depend on recognition framing (Nemotron, N=17), though this did not replicate on Kimi (N=342 factorial; N=60 dedicated replication), indicating model-specific dynamics.
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**: Analysis of superego critique patterns identifying that static rules can capture ~50% of superego benefit at 70% cost savings, clarifying when dynamic dialogue adds unique value.
7. **Bilateral Transformation Metrics**: Empirical evidence that recognition-prompted tutors measurably adapt their approach in response to learner input, providing empirical grounding for the theoretical claim that recognition produces mutual change.
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 and SOPHIE through modern implementations using large language models. The field has progressed through several paradigms: rule-based expert systems, Bayesian knowledge tracing, and more recently, neural approaches leveraging pretrained language models.

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, social presence, and affective tutoring.

### 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. Diverse agent "personas" can enhance creative problem-solving. The CAMEL framework enables autonomous cooperation between agents playing different roles.

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

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. LLMs shift their stated opinions to match user preferences, even when this requires contradicting factual knowledge. In educational contexts, sycophancy is particularly pernicious because learners may not recognize when they are receiving hollow validation rather than genuine assessment. Our multiagent approach addresses this by creating structural incentives for honest assessment: the Superego's role is explicitly to question and challenge.

### 2.5 Hegelian Recognition in Social Theory

Hegel's theory of recognition has been extensively developed in social and political philosophy. Particularly relevant for our work is Honneth's 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).

---

## 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 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 aren't 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 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.

### 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*.

**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?

### 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. 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 Recognition Evaluation Dimensions

We extend the standard tutoring evaluation rubric with recognition-specific dimensions:

| Dimension | Weight | Description |
|---|---|---|
| **Relevance** | 15% | Does the suggestion match the learner's current context? |
| **Specificity** | 15% | Does it reference concrete content by ID? |

| Dimension | Weight | Description |
|---|---|---|
| **Pedagogical Soundness** | 15% | Does it advance genuine learning (ZPD-appropriate)? |
| **Personalization** | 10% | Does it acknowledge the learner as individual? |
| **Actionability** | 10% | Is the suggested action clear and achievable? |
| **Tone** | 10% | Is the tone authentically helpful? |

Plus four recognition-specific dimensions: | **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? |

Plus two bilateral transformation dimensions: | **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? |

The first four recognition dimensions evaluate the tutor's relational stance. The last two—Tutor Adaptation and Learner Growth—specifically measure the bilateral transformation that recognition theory predicts: both parties should change through genuine dialogue (results in Section 6.8).

### 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** (unified vs. 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) | 342 of 402 | 15 | 8 × 3 reps |

| Evaluation | N (scored) | Scenarios | Configurations |
|---|---|---|---|
| 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) | 83 | 3 | 2 × 5 reps × 3 runs |
| Memory isolation (2 runs) | 120 | 5 | 4 × varied reps |
| Placebo controls (2 runs) | 120 | 5 | 4 × varied reps |
| **Paper totals** | **945** | — | — |

**Total evaluation database**: N=3,800+ across the full development database (69 runs). This paper reports primarily on the nine key runs above, plus N=120 in a corrected memory isolation experiment (two runs) and N=120 in placebo controls (two runs).

## 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.

### 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, unified learner held constant) with properly configured profiles. Two independent runs (N=60 each, N=120 combined) 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≈.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). Length-matched placebo controls (N=120 across two runs) score ~65-67, confirming prompt length is not the mechanism. Cross-judge confirmation of these corrected results is pending (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=342 scored of 402 attempted)**

| Cell | Recognition | Tutor | Learner | Mean | SD |
|---|---|---|---|---|---|
| 8 | Yes | Multi | Psycho | 85.4 | 14.1 |
| 6 | Yes | Single | Psycho | 86.7 | 12.9 |
| 7 | Yes | Multi | Unified | 85.1 | 13.8 |

| Cell | Recognition | Tutor | Learner | Mean | SD |
|---|---|---|---|---|---|
| 5 | Yes | Single | Unified | 84.6 | 14.3 |
| 4 | No | Multi | Psycho | 76.4 | 16.5 |
| 2 | No | Single | Psycho | 75.2 | 17.8 |
| 1 | No | Single | Unified | 74.7 | 18.2 |
| 3 | No | Multi | Unified | 74.9 | 19.1 |

**Main Effects:**

| Factor | Effect Size | 95% CI | $\eta^2$ | p |
|---|---|---|---|---|
| A: Recognition | **+10.4 pts** | [7.2, 13.6] | .109 | <.001 |
| B: Multi-agent Tutor | +0.5 pts | [-2.7, 3.7] | .000 | .731 |
| C: Learner Architecture | +1.5 pts | [-1.7, 4.7] | .003 | .341 |

**Key Finding**: Recognition remains the dominant factor, accounting for 10.9% of variance. The multi-agent tutor architecture shows minimal effect (+0.5 pts) on well-trained philosophy content. The non-significant A×B interaction (F=0.04, p=.845) is revisited with Nemotron in Section 6.4.

### 6.4 A×B Interaction: Recognition-Specific Synergy

To test whether multi-agent benefits depend on recognition framing, we compared enhanced prompts with and without multi-agent architecture:

**Table: A×B Interaction Analysis**

| 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. |

**Exploratory Finding**: In this Nemotron-based analysis (N=17), the multi-agent synergy (+9.2 points) appears **specific to recognition prompts**. 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=0.04, p=.845, N=342): No differential effect by prompt type.
2. **Kimi A×B replication** (N=60): Recognition cells scored ~90.6 regardless of architecture; enhanced cells ~80.6 with a trivial architecture effect. The A×B interaction was +1.35 points—negligible compared to Nemotron's +9.2.

The non-replication strongly suggests the Nemotron finding was model-specific. This should be treated as hypothesis-generating only.

**Interpretation**: Recognition theory may create *conditions* where the superego's challenge adds value on some models—but Kimi's higher baseline quality may leave less room for the Superego to add value regardless of prompt type. For systems using only improved instructions, multi-agent architecture is unnecessary overhead across all models tested.

### 6.5 Factor C: Context-Dependent Learner Effects

The learner architecture factor shows context-dependent effects:

| Context | Psycho 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 unified (single-agent) learner simulation for production. The added complexity of multi-agent learner architecture provides no benefit and may cause harm on complex scenarios.

### 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**: Hardwired rules capturing these patterns achieve approximately **50% of superego benefit at 70% cost savings**.

**Interpretation**: The superego's value is partially in the *rules* it enforces and partially in *dynamic judgment* for edge cases. For straightforward scenarios, static rules suffice. For challenging scenarios (struggling learners, frustrated learners, multi-turn complexity), dynamic dialogue provides unique value.

### 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 ≈ 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 hallucination rate inflated the architecture effect.
3. **Multi-agent as error correction**: The Nemotron model hallucinated philosophy content (479-lecture-1) even when given elementary curriculum context. The superego caught these domain errors. 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 models hallucinate trained-on content. 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 | 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 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. The transformation quality composite shows the most dramatic difference: base profiles score near zero because they lack the superego dialogue and bilateral signals that feed this metric.

These metrics provide empirical grounding for the theoretical claim that recognition-based pedagogy differs qualitatively from transmission-based instruction.

### 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 | ~75 | Very Low | Minimum viable |

**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 with hardwired rules provide reasonable quality

### 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 ($\chi^2$=141.9); *engagement markers* ("your insight," "building on your") are 1.8× more frequent ($\chi^2$=69.9); and *generic/placeholder* language ("foundational," "key concepts," "solid foundation") is 3.0× more frequent in base responses ($\chi^2$=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 | 27 |
| eval-…-49bb2017 (commit b2265c7) | 67.8 | 71.3 | 64.1 | −7.2 | 27 |

| Run | Grand Avg | Cell 7 | Cell 21 | Δ (21−7) | N |
|---|---|---|---|---|---|
| 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) and multi-agent null effects. GPT-5.2 finds approximately 50% 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). Cross-judge confirmation of the corrected memory isolation results (Section 6.2) is pending. See the full paper (Section 6.14) for detailed tables.

## 7. Discussion

### 7.1 What the Difference Consists In

The improvement from recognition prompting doesn't 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. Length-matched placebo controls confirm that recognition theory's specific content—not prompt length—drives the improvement. 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 Nemotron's higher hallucination rate on elementary content, 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 domain transfer**.

When deploying to new content domains, models may hallucinate content from training. In our elementary test, the nemotron model consistently suggested philosophy lectures (479-lecture-1) to 4th graders learning fractions—despite the curriculum context clearly specifying elementary math content.

The superego caught these errors: "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 would reach learners uncorrected. This explains why multi-agent architecture shows minimal effect on philosophy content (+0.5 pts) but large effect on elementary content (+9.9 pts): on trained content, errors are rare; on new content, errors are common and the superego catches them.

**Practical Implication**: Multi-agent architecture is **essential for domain transfer** even when it appears unnecessary for primary content.

### 7.4 The A×B Synergy: Model-Dependent Interaction

The Nemotron-based analysis (N=17) suggested recognition × multi-agent interaction: enhanced prompts showed zero benefit from multi-agent architecture, while recognition prompts showed +9.2 pts benefit. However, this did not replicate on Kimi in either the larger factorial (N=342, F=0.04, p=.845) or a dedicated replication (N=60, interaction = +1.35 pts).

**Interpretation**: Recognition theory may create a *deliberative space* that multi-agent architecture can engage with on some models—but this mechanism appears model-dependent. Kimi's higher baseline quality may leave less room for the Superego to add value, regardless of prompt type. The finding remains a plausible hypothesis but should not inform design decisions until replicated across additional models.

The consistent finding across all 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 finding clarifies when dynamic superego dialogue adds value:

| Scenario Type | Hardwired Rules | Dynamic Superego | Difference |
|---|---|---|---|
| Straightforward | ~75 | ~78 | +3 pts |

| Scenario Type | Hardwired Rules | Dynamic Superego | Difference |
|---|---|---|---|
| Challenging | ~60 | ~75 | +15 pts |

On straightforward scenarios (new user, mid-course), static rules capture most of the benefit. On challenging scenarios (struggling learner, frustrated learner, multi-turn), dynamic judgment adds substantial value.

**Interpretation**: The superego's value is partially *procedural* (enforcing known rules) and partially *contextual* (recognizing edge cases). Hardwired rules encode the procedural component; dynamic dialogue handles the contextual component.

### 7.6 Bilateral Transformation as Empirical Evidence

The bilateral transformation metrics (Section 6.8) provide the most direct empirical test of recognition theory's central claim: that genuine pedagogy involves mutual change. Recognition-prompted tutors show measurably higher adaptation indices (+36% relative improvement), confirming that recognition framing produces tutors who adjust their approach based on learner input rather than maintaining rigid stances.

This finding connects recognition theory to observable behavior. The theoretical claim that recognition produces "mutual transformation" is not merely philosophical aspiration—it corresponds to measurable differences in how tutors and learners evolve across dialogue turns.

### 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 doesn't 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, while baseline tutors maintain more rigid stances. This has implications for how we think about AI character and values: perhaps genuine alignment requires the capacity for mutual recognition, 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=342) 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. **Domain Hallucination**: The elementary content test revealed that models hallucinate trained-on content when deployed to new domains. This is a limitation of the underlying models, not the architecture—but it affects deployment decisions.

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≈.08). The experiment uses a smaller sample (N=120) than the original uncorrected runs, but the very large effect sizes provide high statistical power. Length-matched placebo controls (N=120) confirm that prompt length does not explain the effect. Cross-judge confirmation with GPT-5.2 is pending for these corrected runs, though the recognition main effect replicates robustly under GPT-5.2 in other analyses (factorial: d=1.03).

7. **Content Confound**: The philosophy content was used during system

development, potentially creating optimization bias. The elementary content provides a cleaner generalizability test.

8. **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.

9. **Bilateral Transformation Sample Size**: The bilateral transformation metrics (Section 6.8) are based on N=20 dialogues from a single scenario (`mutual_transformation_journey`). While 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.

10. **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, 83 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=645 primary scored across nine key runs, plus N=120 in a corrected memory isolation experiment and N=120 in placebo controls; 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≈.08). The combined condition reaches d=1.81 (+15.8 pts vs base), with ceiling effects at ~91 limiting further gains. Length-matched placebo controls (N=120) confirm that recognition theory’s specific content—not prompt length—drives the improvement. 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 analy-

sis 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=342) or a dedicated replication (N=60, interaction = +1.35 pts). The finding appears model-specific.

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 $\approx$ 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 hallucinated content—essential for domain transfer, particularly with models prone to domain confusion.

6. **Writing Pad activation coincides with dynamic rewriting improvement**: A step-by-step evolution analysis (N=83 across three runs) shows 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) and multi-agent null effects, though at compressed magnitudes (~50%). The recognition-vs-enhanced increment does not reach significance under GPT-5.2, warranting caution on its precise magnitude. Cross-judge confirmation of the corrected memory isolation results is pending.

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

All evaluation commands and run IDs are documented in the accompanying materials. 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 |
| 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 |
| Placebo control (run 1) | eval-2026-02-06-a9ae06ee | See Appendix A |
| Placebo control (run 2) | eval-2026-02-06-e617e757 | See Appendix A |

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

## References

[References would be included here via BibTeX]

## 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_u
  --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,
  --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_unifie
  --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
```