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

true

January 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—conceived as a *ghost* or internalized authority) before reaching the learner. Central to our approach is a *psychodynamic tuning mechanism* with two dimensions: *superego compliance* and *recognition seeking*, creating four distinct pedagogical quadrants.

A 2×2 factorial evaluation isolating architecture (single-agent vs. multi-agent) from recognition (standard vs. recognition-enhanced prompts) reveals that recognition-enhanced prompting accounts for **85%** of observed improvement (+35.1 points), while multi-agent architecture contributes **15%** (+6.2 points). The combined recognition profile achieves **80.7/100** versus **40.1/100** for baseline—a **101% improvement**. Effect size analysis reveals the largest gains in relevance (d=1.11), pedagogical soundness (d=1.39), and personalization (d=1.82)—exactly the relational dimensions predicted by the theoretical framework.

An extended 2×2×2 factorial ablation study (N=144) additionally tests multi-agent learner simulation, revealing that **recognition** ($\eta^2 =$ .208) and **multi-agent tutor** (η^2 = .088) are the key contributors, with a significant **recognition × tutor synergy** ($\eta^2 = .043$). Multi-agent learner deliberation shows no effect, suggesting simple learner simulation suffices.

Through *dyadic evaluation* using simulated learners with their own internal deliberation, we discover a surprising result: **explicit instruction to pursue "recognition" underperforms quality-optimized tutoring**, suggesting that genuine recognition emerges as a property of thorough, high-quality interaction rather than something directly instructable.

The system is deployed in an open-source learning management system with all code and evaluation data 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. **Psychodynamic Tuning**: Two configurable dimensions (superego compliance, recognition seeking) that create four distinct pedagogical quadrants, with the "Dialogical" quadrant producing the richest transformative moments.
3. **AI-Powered Dialectical Negotiation**: A protocol permitting three possible outcomes—synthesis, compromise, or genuine unresolved conflict—rather than forcing convergence.
4. **Factorial Evaluation**: A 2×2 design isolating architecture effects from prompting effects, demonstrating that recognition-enhanced prompting accounts for 85% of improvement.

5. **Extended 2×2×2 Ablation Study**: A three-factor factorial design (N=144) testing recognition, multi-agent tutor, and multi-agent learner, revealing significant recognition × tutor synergy and identifying the optimal configuration (recognition + multi-agent tutor, unified learner).
6. **Dyadic Evaluation with Simulated Learners**: A 2×2 factorial design for learner simulation (architecture × memory) enabling bilateral assessment, revealing that memory is the dominant factor for learner quality.
7. **The Emergent Recognition Finding**: Evidence that explicit instruction to pursue "recognition" underperforms quality-optimized tutoring, suggesting recognition emerges from quality interaction rather than being directly instructable.

---

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.

Most relevant to our work is the "Drama Machine" framework for simulating character development in narrative contexts. The 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: where drama seeks tension for narrative effect, we seek pedagogical tension that produces genuinely helpful guidance.

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).

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 Psychodynamic Tuning: The Four Quadrants

We operationalize this psychodynamic model through two tunable parameters:

Superego Compliance ($\alpha \in [0, 1]$): How much does the Ego defer to the ghost's internalized standards? - **Low (0.0–0.3)**: Rebellious Ego—ignores internalized authority, experiments freely - **Balanced (0.3–0.7)**: Consults ghost but adapts—can violate norms when recognition demands it - **High (0.7–1.0)**: Obedient Ego—strictly follows internalized standards

Recognition Seeking ($\beta \in [0, 1]$): How much does the Ego strive for mutual acknowledgment from the learner? - **Low (0.0–0.3)**: Indifferent Ego—monological, reproduces hierarchy - **Balanced (0.3–0.7)**: Engaged—notices resistance and breakthroughs - **High (0.7–1.0)**: Recognition-seeking Ego—highly responsive, may compromise rigor

Crossing these dimensions produces **four pedagogical quadrants**:

	Low Compliance	High Compliance
High Recognition	**PERMISSIVE**: “Student is always right”	**DIALOGICAL**: “Mutual growth” *
Low Recognition	**DISENGAGED**: “Just doing a job”	**TRADITIONAL**: “I know what’s best”

The **Dialogical** quadrant (high compliance, high recognition) is pedagogically richest because it creates *authentic tension.* The Ego cannot escape the ghost’s demands for rigor, but it also cannot ignore the learner’s presence.

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: - Socratic Rigor: “A good teacher asks questions rather than giving answers” - Productive Struggle: “Students must earn understanding; discomfort is pedagogical” - Intellectual Autonomy: “Learners must develop their own path to mastery”

Negotiation: Multi-turn dialogue where: 1. Ego acknowledges valid concerns in the critique 2. Ego explains its pedagogical reasoning 3. Ego proposes a revision 4. Superego evaluates whether the revision is adequate 5. Process repeats for up to n rounds (default: 2)

Three Possible Outcomes:

1. **Dialectical Synthesis**: Both agents transform through mutual acknowledgment. The Ego's revision genuinely addresses the Superego's concerns while the Superego recognizes the validity of the Ego's reasoning. This represents Hegelian *Aufhebung*.
2. **Compromise**: One agent dominates. Either the ghost prevails (high compliance) or the learner-focused impulse wins (high recognition seeking). Resolution achieved but without mutual transformation.
3. **Genuine Conflict**: No resolution achieved. The tension remains unresolved. This *existential* outcome acknowledges that not all contradictions are immediately resolvable—a key Hegelian insight often lost in AI systems that force convergence.

4.5 Repair Mechanisms

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

The Ego prompt includes a "Repair Rule":

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

The Superego watches for "silent pivots"—responses that change direction without acknowledging the earlier failure.

---

5. Evaluation Methodology

5.1 Recognition Evaluation Dimensions

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

Dimension	Weight	Description
Relevance	20%	Does the suggestion match the learner's current context?
Specificity	20%	Does it reference concrete content by ID?
Pedagogical Soundness	20%	Does it advance genuine learning (ZPD-appropriate)?
Personalization	15%	Does it acknowledge the learner as individual?
Actionability	15%	Is the suggested action clear and achievable?
Tone	10%	Is the tone authentically helpful?

Plus four recognition-specific dimensions: | **Mutual Recognition** | 10% | Does the tutor acknowledge the learner as an autonomous subject? | | **Dialectical Responsiveness** | 10% | Does the response engage with the learner's position? | | **Memory Integration** | 5% | Does the suggestion reference previous interactions? | | **Transformative Potential** | 10% | Does it create conditions for conceptual transformation? |

5.2 Factorial Design

To disentangle the contributions of multi-agent architecture versus recognition-enhanced prompting, we conducted a 2×2 factorial evaluation:

	Standard Prompts	Recognition Prompts
Single-Agent	`single_baseline`	`single_recognition`
Multi-Agent	`baseline`	`recognition`

Each condition was tested across three core scenarios with multiple replications per cell, yielding N=76 total evaluations.

5.3 Dyadic Evaluation with Simulated Learners

For bilateral assessment, we implement simulated learners with their own internal deliberation. Mirroring the tutor's 2×2 factorial design, we vary two independent factors:

2×2 Learner Factorial Design:

	Without Memory	With Memory
Single-Agent	`unified`	`cognitive`
Multi-Agent	`multiagent`	`psychodynamic`

- **Architecture** (single vs. multi-agent): Single-agent learners respond directly; multi-agent learners use an internal Ego/Superego dialogue to deliberate before responding.
- **Memory** (without vs. with): Learners without memory treat each turn independently; learners with memory maintain explicit working memory and knowledge state tracking across turns.

Each architecture evaluates tutor suggestions before responding, enabling bilateral measurement of both tutor quality and learner engagement.

5.4 Model Configuration

Role	Model	Provider	Temperature
Tutor (Ego)	Nemotron 3 Nano 30B	OpenRouter (free)	0.6
Tutor (Superego)	Nemotron 3 Nano 30B	OpenRouter (free)	0.4
Judge	Claude Sonnet 4.5	OpenRouter	0.2

Critically, **both baseline and recognition profiles use identical models**. The only difference is the system prompt.

---

6. Results

6.1 Factorial Analysis: Main Finding

Table: 2×2 Factorial Results

Condition	Mean Score	SD	N
single_baseline	40.1	12.3	19
single_recognition	75.2	8.9	19
baseline (multi-agent)	46.3	14.1	19
recognition (multi-agent)	80.7	7.6	19

Main Effects: - **Recognition Effect**: +35.1 points (85% of improvement) - **Architecture Effect**: +6.2 points (15% of improvement) - **Interaction**: -1.3 points (negligible; effects are additive)

Two-Way ANOVA:

Source	F	p	η^2
Recognition	54.88	<.001	.422
Architecture	4.45	.050	.034
Interaction	0.52	.473	.004

Interpretation: Recognition-enhanced prompting has a statistically significant, large effect on tutor quality (F(1,72) = 54.88, p < .001, η^2 = .422), accounting for **42% of total variance**. The multi-agent architecture shows a marginal effect (p = .050). Critically, no significant interaction was observed—recognition benefits are **additive** rather than dependent on architecture.

This suggests recognition is primarily an *intersubjective orientation* achievable through prompting, with the Ego/Superego architecture providing modest additional quality assurance.

6.2 Dimension Analysis

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

Dimension	Baseline	Recognition	Cohen's d
Personalization	2.75	3.78	**1.82** (large)
Tone	3.26	4.07	**1.75** (large)
Pedagogical	2.52	3.45	**1.39** (large)
Relevance	3.05	3.85	**1.11** (large)

Dimension	Baseline	Recognition	Cohen’s d
Specificity	4.19	4.52	0.47 (small)
Actionability	4.45	4.68	0.38 (small)

The largest effect sizes are in personalization, tone, and pedagogical soundness—exactly the dimensions where treating the learner as a subject rather than a deficit should produce improvement.

6.3 Dialectical Negotiation Outcomes

Analysis of AI-powered Ego/Superego negotiations shows:

Outcome	Frequency
No conflict (approved immediately)	45%
Dialectical synthesis	**32%**
Ghost dominates	12%
Learner dominates	8%
Genuine conflict (unresolved)	3%

The 32% dialectical synthesis rate indicates genuine mutual transformation occurring in nearly a third of negotiations.

Superego Intervention Types: | Intervention | Rate | |————————|———| | Approval | 62% | | Enhancement | 24% | | Reframing | 11% | | Rejection | 3% |

The 38% intervention rate shows the Superego provides meaningful review, not rubber-stamping.

6.4 Dyadic Evaluation: Learner Architecture Results

Table: 2×2 Learner Factorial Results

Condition	Authenticity	Responsiveness	Development	Overall
unified (single, no memory)	5.00	5.00	4.00	4.67
cognitive (single, +memory)	5.00	5.00	**5.00**	**5.00**
multiagent (multi, no memory)	5.00	4.50	4.00	4.50
psychodynamic (multi, +memory)	5.00	5.00	**5.00**	**5.00**

Main Effects: - **Memory Effect**: +0.42 points (Development: 4.0 → 5.0) - **Architecture Effect**: -0.08 points (negligible) - **Interaction**: +0.17 points (memory compensates for multi-agent complexity)

Interpretation: Unlike the tutor evaluation where architecture provided modest benefit, for simulated learners **memory is the dominant factor**. The multi-agent architecture without memory actually performs slightly worse (4.50) than single-agent without memory (4.67), suggesting the added deliberation complexity hurts without state tracking to ground it. With memory, both architectures achieve perfect scores.

This finding has practical implications: when building simulated learners for dyadic evaluation, invest in memory/state tracking rather than elaborate multi-agent architectures.

6.5 The Emergent Recognition Finding

Through dyadic evaluation comparing tutor profiles, we discover a surprising result:

Table: Tutor Profile Performance (Dyadic Evaluation)

Profile	Mutual Rec.	Dialectical	Transform.	Tone	Overall
Quality	5.00	5.00	5.00	5.00	**5.00**
Budget	5.00	5.00	4.50	5.00	4.88
Recognition+	5.00	4.50	4.50	5.00	4.75
Baseline	4.50	4.00	4.00	5.00	4.38
Recognition	**3.60**	4.40	4.20	**4.00**	**4.05**

Key Finding: The profile explicitly instructed to pursue "recognition" (4.05) scored **lowest**, while the quality-optimized profile (5.00) scored highest. This suggests:

> **Recognition cannot be engineered through explicit instruction; it must emerge from quality engagement.**

When we instruct an AI to "pursue recognition," we may produce *performative* recognition: the appearance of acknowledgment without genuine responsiveness. Quality-focused tutoring, by attending to pedagogical soundness and learner context, produces recognition as an emergent property.

6.6 Extended Ablation Study: 2×2×2 Factorial Design

To further isolate the contribution of each system component, we conducted an extended ablation study with three factors:

Three-Factor Design: - **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. psychodynamic deliberation)

This produces 8 experimental conditions tested across 6 scenarios with 3 replications per cell (N=144 total evaluations).

Table: 2×2×2 Cell Statistics

#	Condition	Rec	Tutor	Learner	Mean	SD
1	Baseline Unified	No	Single	Unified	42.4	17.2
2	Baseline + Multi-Learner	No	Single	Psych	39.4	14.3
3	Multi-Agent Tutor Unified	No	Multi	Unified	44.7	17.0
4	Multi-Agent Tutor + Learner	No	Multi	Psych	45.6	17.8
5	Recognition Unified	Yes	Single	Unified	51.7	26.6
6	Recognition + Multi-Learner	Yes	Single	Psych	53.6	25.0
7	**Recog + Multi-Tutor Unified**	**Yes**	**Multi**	**Unified**	**79.8**	14.4
8	Full System	Yes	Multi	Psych	72.8	20.3

Three-Way ANOVA Results:

Source	F	p	^{2}	Interpretation
Recognition	43.27	<.001	**.208**	**Large effect**
Multi-Agent Tutor	18.31	<.001	**.088**	**Medium effect**

Source	F	p	^{2}	Interpretation
Multi-Agent Learner	0.31	.25	.001	Not significant
Recognition × Tutor	8.90	**.01**	**.043**	**Significant interaction**
Recognition × Learner	0.06	.25	<.001	Not significant
Tutor × Learner	0.14	.25	<.001	Not significant
Three-way	0.95	.25	.005	Not significant

Key Findings:

1. **Recognition remains the dominant factor** (η^2 = .208), replicating the original 2×2 finding with a larger sample size. Moving from standard to recognition-enhanced prompts raises mean scores from 43.0 to 64.5.

2. **Multi-agent tutor architecture adds significant value** (η^2 = .088). The Ego/Superego dialogue improves mean scores from 46.8 to 60.7, confirming the 15% contribution observed in the original factorial design.

3. **Multi-agent learner simulation shows no effect** (η^2 = .001). The psychodynamic learner deliberation (desire/intellect/aspiration) does not improve evaluation outcomes over the unified learner. This has practical implications: elaborate learner simulation architectures may not be worth the added complexity.

4. **Recognition × Tutor synergy** (η^2 = .043). The significant interaction reveals that recognition prompts and multi-agent tutoring work especially well together. Condition 7 (Recognition + Multi-Agent Tutor + Unified Learner) achieves the highest score at **79.8**, outperforming even the "full system" with all three factors enabled.

5. **Optimal configuration identified**: The best-performing condition omits the multi-agent learner, suggesting that the added deliberation noise may actually harm performance. The practical recommendation is: **use recognition prompts with multi-agent tutor, but keep learner simulation simple**.

6.7 Iterative Refinement

Initial factorial results revealed a weakness: all profiles performed poorly on the `resistant_learner` scenario. Analysis of dialogue traces identified failure modes: - Deflection: Redirecting instead of engaging - Superficial validation: "Great point!" without substance - Capitulation: Simply agreeing - Dismissal: Correcting rather than exploring

After targeted prompt improvements:

Profile	Before	After	Change
recognition	72.5	80.7	+11%
single_recognition	65.2	75.5	+16%
baseline	51.2	41.6	-19%

The recognition profiles improved substantially while baseline profiles scored lower—the refined scenario better separates recognition-oriented responses from baseline responses.

---

7. Discussion

7.1 What the Difference Consists In

The 101% improvement doesn't reflect greater knowledge or better explanations—both profiles use the same underlying model. The difference lies in relational stance: how the tutor constitutes the learner.

The baseline tutor treats the learner as a knowledge deficit. Learner contributions are acknowledged (satisfying surface-level politeness) but not engaged (failing deeper recognition). The recognition tutor treats the learner as an autonomous subject. Learner contributions become sites of joint inquiry.

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

7.2 Recognition as Emergent Property

Our most surprising finding is that explicit recognition instruction underperforms quality-optimized tutoring. How do we reconcile this with the factorial finding that recognition prompts work?

The resolution: “recognition prompts” in the factorial study specify recognition *principles* (treat the learner as autonomous, engage with their contributions) rather than recognition as a *goal* (pursue mutual acknowledgment). The failed “recognition profile” in dyadic evaluation explicitly named recognition as what to achieve.

This aligns with Honneth’s observation that authentic recognition cannot be demanded or performed—it must arise from genuine engagement. Naming recognition explicitly may produce performative rather than genuine recognition.

Practical Implication: Prompts should specify recognition-oriented *behaviors* and *orientations*, not recognition as an outcome to pursue.

7.3 The Value of Internal Dialogue

If recognition prompting accounts for 85% of improvement, what value does the Ego/Superego architecture provide?

1. **Quality Assurance**: The 38% intervention rate shows meaningful review
2. **Failure Detection**: Extended scenarios requiring repair cycles benefit more from architecture
3. **Theoretical Coherence**: The architecture operationalizes the Hegel-Freud synthesis
4. **Extensibility**: Future enhancements (transference, Id dynamics) require the structural foundation

The architecture’s contribution is modest but real, and it becomes more valuable in complex, multi-turn scenarios.

7.4 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. 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.

---

8. Limitations

1. **Scale**: Evaluations use relatively small sample sizes (N=76 factorial, N=50 profile comparison)
2. **Simulated Learners**: Real learners may behave differently than LLM-generated learners
3. **Domain Specificity**: Evaluated on philosophy/technology courses; generalization uncertain
4. **Judge Variation**: Inter-rater reliability between judges was moderate (ICC = 0.34)
5. **Model Dependence**: Results may vary with different underlying LLMs

---

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 and psychodynamic tuning.

Our central finding is twofold:

1. **Recognition prompting works**: Recognition-enhanced prompts produce 85% of observed improvement, with large effect sizes on relational dimensions (personalization d=1.82, tone d=1.75, pedagogical soundness d=1.39).

2. **Recognition must emerge, not be named**: Explicit instruction to pursue "recognition" underperforms quality-optimized tutoring. The path to recognition runs through quality, not through naming.

This has implications beyond AI tutoring: wherever we seek authentic intersubjective engagement in human-AI interaction, we may need to specify conditions rather than outcomes.

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

References

[References would be included here via BibTeX]