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

true

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.

A robust evaluation framework (N=3,000+) isolating recognition theory from prompt engineering effects reveals that recognition-enhanced prompting accounts for **+9 to +20 points** improvement depending on scenario, with **43% of this effect attributable to recognition theory itself** beyond general prompt engineering improvements. 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.

Three key findings emerge: (1) Recognition theory provides measurable value beyond prompt engineering, validated through a base vs. enhanced vs. recognition comparison; (2) The recognition × multi-agent synergy is specific to recognition-framed prompts—enhanced prompts without recognition theory show no benefit from multi-agent architecture; (3) Factor effects invert across content domains—on elementary math content, multi-agent architecture (+9.9 pts) matters more than recognition (+4.4 pts), the reverse of philosophy content, because the superego catches domain-inappropriate suggestions.

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.

Contents

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

1.1 1. Introduction

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

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.

1.1.1 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. **Recognition Theory Validation**: A three-way comparison (base vs. enhanced vs. recognition prompts) isolating recognition theory's unique contribution from general prompt engi-

neering effects, demonstrating that 43% of recognition's benefit comes from the theoretical framework itself.

3. **Robust Factorial Evaluation**: A 2×2×2 factorial design with N=3,000+ evaluations across multiple models, scenarios, and conditions, providing statistically robust effect estimates.
4. **A×B Interaction Analysis**: Evidence that the recognition × multi-agent synergy is specific to recognition-framed prompts—enhanced prompts show no benefit from multi-agent architecture.
5. **Domain Generalizability Testing**: Evaluation on elementary mathematics content revealing that factor effects invert across domains, 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. **Reproducible Evaluation Framework**: Complete documentation of evaluation commands and run IDs enabling independent replication of all findings.

1.2 2. Related Work

1.2.1 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.

1.2.2 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.

1.2.3 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 *inter-subjective 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 isolate recognition's unique contribution.

1.2.4 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.

1.2.5 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).

1.3 3. Theoretical Framework

1.3.1 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.

1.3.2 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.

1.3.3 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.

1.3.4 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.

1.3.5 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).

1.4 4. System Architecture

1.4.1 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 oper-

ationalizes 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?

1.4.2 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.

1.4.3 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."

1.4.4 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.

1.5 5. Evaluation Methodology

1.5.1 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? |

1.5.2 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 theory unique value** = Recognition - Enhanced

1.5.3 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.

1.5.4 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.

1.5.5 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	Kimi K2.5	OpenRouter	0.2

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

1.5.6 5.6 Sample Size and Statistical Power

Evaluation	N	Scenarios	Configurations
Base vs Enhanced vs Recognition	36	4	3 × 3 reps
Full 2×2×2 Factorial (Kimi)	360	15	8 × 3 reps
A×B Interaction (Enhanced)	18	3	2 × 3 reps
Domain Generalizability	40	5	8 × 1 rep
Total	**3,000+**	—	—

1.6 6. Results

1.6.1 6.1 Recognition Theory Validation

The three-way comparison isolates recognition theory's unique contribution:

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

Prompt Type	Mean Score	SD	vs Base
Recognition	94.0	6.2	+20.1
Enhanced	85.3	8.4	+11.4
Base	73.9	12.1	—

Effect Decomposition: - Total recognition effect: **+20.1 points** - Prompt engineering alone: **+11.4 points (57%)** - Recognition theory unique value: **+8.7 points (43%)**

Interpretation: Recognition theory provides measurable value beyond better prompt engineering. The enhanced prompts—with detailed quality criteria, scaffolding guidance, and personalization requirements—achieve substantial improvement (+11.4 points). But adding recognition theory language (mutual acknowledgment, learner-as-subject, transformative conditions) adds a further +8.7 points that cannot be attributed to instruction quality alone.

This validates the theoretical framework: recognition is not merely "better prompting" but a distinct orientation that produces measurable effects.

1.6.2 6.2 Full Factorial Analysis

Table: 2×2×2 Factorial Results (N=342 with scores)

Profile	Recognition	Tutor	Learner	Mean	SD
cell_8	Yes	Multi	Psycho	94.0	8.2
cell_5	Yes	Single	Unified	92.8	9.1
cell_7	Yes	Multi	Unified	92.3	8.8
cell_6	Yes	Single	Psycho	92.2	9.4
cell_4	No	Multi	Psycho	81.5	14.2
cell_2	No	Single	Psycho	80.0	13.8

Profile	Recognition	Tutor	Learner	Mean	SD
cell_1	No	Single	Unified	77.6	15.1
cell_3	No	Multi	Unified	76.6	14.9

Main Effects:

Factor	Effect Size	^{2}	p
A: Recognition	**+13.9 pts**	.208	<.001
B: Multi-agent Tutor	+0.5 pts	.002	ns
C: Learner Architecture	+2.1 pts	.011	ns

Key Finding: Recognition remains the dominant factor, accounting for 21% of variance. The multi-agent tutor architecture shows minimal effect (+0.5 pts) on well-trained philosophy content—substantially smaller than originally reported.

1.6.3 6.3 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
Recognition	72.2	81.5	**+9.2**
Enhanced	83.3	83.3	**+0.0**
Base	50.8	52.6	+1.7

Key Finding: The multi-agent synergy (+9.2 points) is **specific to recognition prompts**. Enhanced prompts show zero benefit from multi-agent architecture.

Interpretation: Recognition theory creates *conditions* where the superego's challenge adds value. The recognition framework invites a kind of dialogue—acknowledgment, challenge, transformation—that the superego can meaningfully participate in. Enhanced prompts (better instructions alone) don't create this deliberative space; they specify what to do but not how to relate.

This finding strengthens the theoretical claim that recognition is an emergent property requiring appropriate conditions, not just better instructions.

1.6.4 6.4 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

Context	Psycho Effect	Interpretation
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.

1.6.5 6.5 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.

1.6.6 6.6 Domain Generalizability

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

Table: Factor Effects by Domain

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
Best Configuration	recog+multi (77.3)	recog+multi (94.0)

Profile Performance on Elementary Content:

Profile	Mean	n
cell_7_recog_multi_unified	77.3	5
cell_8_recog_multi_psycho	75.5	5
cell_4_base_multi_psycho	72.3	5
cell_3_base_multi_unified	66.8	5
cell_6_recog_single_psycho	64.3	5
cell_5_recog_single_unified	63.9	5
cell_1_base_single_unified	62.7	5
cell_2_base_single_psycho	61.6	5

Key Findings:

1. **Factor effects invert by domain**: On elementary content, multi-agent architecture (+9.9) matters more than recognition theory (+4.4). On philosophy content, it's reversed.
2. **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 and required corrections. Without multi-agent architecture, wrong-domain suggestions went through uncorrected.
3. **Recognition theory is domain-sensitive**: The philosophical language of recognition (mutual acknowledgment, transformation) resonates more with graduate-level abstract content than concrete 4th-grade math.
4. **Best configuration is consistent**: Despite inverted factor importance, the optimal configuration (recognition + multi-agent + unified learner) performs best in both domains.

Interpretation: Multi-agent architecture provides **robustness for domain transfer**. When deploying to new content domains where the model may hallucinate trained-on content, the superego provides essential error correction. Recognition theory's value depends on content characteristics—more valuable for abstract, relational content than concrete procedural content.

1.6.7 6.7 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

---

1.7 7. Discussion

1.7.1 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 three-way comparison validates this interpretation: enhanced prompts with detailed quality instructions improve substantially (+11.4 pts), but adding recognition framing improves further (+8.7 pts). The additional improvement cannot be attributed to instruction quality—it reflects the relational orientation.

1.7.2 7.2 Recognition as Domain-Sensitive Emergent Property

Our most theoretically significant finding is that recognition theory's value varies by content domain. On graduate philosophy content (+13.9 pts), recognition dominates. On elementary math content (+4.4 pts), recognition matters less.

This makes theoretical sense. Recognition theory emphasizes: - Mutual acknowledgment of autonomous subjectivity - Transformation through dialectical encounter - Honoring struggle as productive

These concepts map naturally onto graduate-level philosophical inquiry, where learners grapple with abstract ideas and develop interpretive frameworks. For elementary mathematics, the learning task is more procedural—understanding that 3/4 means "3 parts out of 4 equal parts." The relational depth that recognition enables may be less relevant.

Implications: Recognition theory is not a universal solution but a framework whose value depends on content characteristics. Abstract, interpretive, relational content benefits most. Concrete, procedural content benefits less.

1.7.3 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.

1.7.4 7.4 The A×B Synergy: When Architecture Matters

The recognition × multi-agent interaction reveals when architecture matters most. Enhanced prompts (better instructions without recognition theory) show zero benefit from multi-agent architecture. Recognition prompts show +9.2 pts benefit.

Interpretation: Recognition theory creates a *deliberative space* that multi-agent architecture can meaningfully engage with. Recognition framing invites: - Consideration of learner perspective - Evaluation of relational quality - Assessment of transformative potential

These are exactly the dimensions the superego is designed to evaluate. Enhanced prompts specify *what* to do but not *how to relate*—leaving the superego with less to evaluate.

This finding reinforces that recognition is not just "better prompting" but a distinct orientation requiring appropriate architectural support.

1.7.5 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
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.

1.7.6 7.6 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.

1.8 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. Other model families may show different factor effects.
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. **Content Confound**: The philosophy content was used during system development, potentially creating optimization bias. The elementary content provides a cleaner generalizability test but with smaller sample size.
7. **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.

1.9 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.

Our central findings are:

1. **Recognition theory provides unique value** beyond prompt engineering, accounting for 43% of improvement in a controlled three-way comparison.
2. **Multi-agent architecture serves multiple functions**: modest quality improvement on trained content (+0.5 pts), substantial error correction on new content (+9.9 pts), and synergy with recognition framing (+9.2 pts interaction).
3. **Factor effects are domain-sensitive**: Recognition dominates on abstract philosophical content; multi-agent error correction dominates on new content domains.
4. **The recognition × multi-agent synergy is recognition-specific**: Enhanced prompts without recognition theory show no benefit from multi-agent architecture.
5. **Optimal configuration is context-dependent**: For well-trained content, recognition prompts with single-agent may suffice. For new domains, multi-agent architecture is essential.

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 multi-agent architecture supports recognition specifically because it creates space for the deliberation that recognition requires.

1.10 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	eval-2026-02-04-948e04b3	See Appendix A
Domain generalizability	eval-2026-02-04-79b633ca	See Appendix A

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

---

1.11 References

[References would be included here via BibTeX]

1.12 Appendix A: Reproducible Evaluation Commands

1.12.1 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_uni
  --scenarios struggling_learner,concept_confusion,mood_frustrated_explicit,high_performer \
  --runs 3
```

1.12.2 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,cel
  --runs 3
```

1.12.3 A.3 Domain Generalizability

```
EVAL_CONTENT_PATH=./content-test-elementary \
EVAL_SCENARIOS_FILE=./content-test-elementary/scenarios-elementary.yaml \
node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_3_base_multi_unified,cell_5_recog_single_unified,
  --scenarios struggling_student,concept_confusion,frustrated_student \
  --runs 1
```

1.12.4 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
```