Mutual Recognition in AI Tutoring: A Hegelian Framework for Intersubjective Pedagogy

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 recognition-enhanced prompts and a multi-agent architecture where an "Ego" agent generates pedagogical suggestions and a "Superego" agent (a *productive metaphor* for internal quality review) evaluates them before delivery. A robust evaluation framework (N=2,700+ scored responses across 49 evaluation runs) isolating recognition theory from prompt engineering reveals that recognition provides +8.7 points of unique value beyond better instructions (43% of total effect), with the remaining 57% attributable to prompt engineering improvements. Crucially, the multi-agent synergy effect (+9.2 points) is specific to recognition prompts—enhanced prompts without recognition theory show zero benefit from multi-agent architecture, suggesting recognition creates the conditions for productive internal dialogue. Domain generalizability testing reveals factor effects invert by content type: philosophy content shows recognition dominance (+13.9 pts) while elementary math shows architecture dominance (+9.9 pts), with multi-agent architecture serving as critical error correction when models hallucinate trained content on new domains. These results suggest that operationalizing philosophical theories of intersubjectivity as design heuristics can produce measurable improvements in AI tutor adaptive pedagogy, and that recognition may be better understood as an achievable relational stance rather than requiring genuine machine consciousness.

Contents

Mutual Recognition in AI Tutoring: A Hegelian Framework for Intersubjective Pedagogy

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.

We operationalize this framework through:

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

In controlled evaluations using a robust factorial design (N=435 primary evaluations; N=2,700+ across all development runs), we isolate the unique contribution of recognition theory from prompt engineering effects. A three-way comparison (base vs enhanced vs recognition) reveals that recognition provides +8.7 points of unique value beyond better instructions—43% of the total recognition effect is attributable to the theoretical framework itself, not merely to better prompting.

More significantly, we discover that the multi-agent synergy effect (+9.2 points) is *specific to*

recognition prompts. Enhanced prompts without recognition theory show zero benefit from multi-agent architecture. This suggests recognition creates the deliberative conditions where internal Ego-Superego dialogue adds value—the architecture matters only when paired with recognition framing.

Domain generalizability testing reveals important nuances: factor effects invert by content type. Philosophy content shows recognition dominance (+13.9 pts vs +0.5 pts for architecture), while elementary math shows architecture dominance (+9.9 pts vs +4.4 pts for recognition). The multi-agent architecture serves as critical error correction when models hallucinate trained content on new domains—essential for domain transfer.

The contributions of this paper are:

- A theoretical framework connecting Hegelian recognition to AI pedagogy
- A multi-agent architecture for implementing recognition in tutoring systems
- Empirical evidence that recognition-oriented design improves tutoring outcomes
- Validation that recognition theory provides unique value beyond prompt engineering
- Analysis of how recognition effects vary across content domains
- Evidence that multi-agent synergy requires recognition framing

---

2. Related Work

2.1 AI Tutoring and Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) have a long history, from early systems like SCHOLAR [@carbonell1970] and SOPHIE [@brown1975] through modern implementations using large language models. The field has progressed through several paradigms: rule-based expert systems, Bayesian knowledge tracing [@corbett1995], and more recently, neural approaches leveraging pretrained language models [@kasneci2023].

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 [@zhao2014], social presence [@biocca2003], and affective tutoring [@dmello2012].

2.2 Prompt Engineering and Agent Design

The emergence of large language models has spawned extensive research on prompt engineering—how to instruct models to produce desired behaviors [@brown2020; @wei2022]. Most prompting research treats prompts as behavioral specifications: persona prompts, chain-of-thought instructions, few-shot examples [@kojima2022].

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

Multi-agent architectures have been explored for task decomposition [@wu2023], debate [@irving2018], and self-critique [@madaan2023]. Our Ego/Superego architecture contributes a specific

use case: internal evaluation of relational quality before external response.

2.3 AI Personality and Character

Research on AI personality typically treats personality as dispositional—stable traits the system exhibits [@volkel2021]. Systems are friendly or formal, creative or precise. The "Big Five" personality framework has been applied to chatbot design [@zhou2020].

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

This connects to Anthropic's research on Claude's character [@anthropic2024]. Constitutional AI specifies values the model should hold, but values don't fully determine relational stance. A model could value "being helpful" while still enacting one-directional helping. Recognition adds a dimension: mutual constitution.

2.4 Constructivist Pedagogy and Productive Struggle

Constructivist learning theory [@piaget1954; @vygotsky1978] emphasizes that learners actively construct understanding rather than passively receiving information. The zone of proximal development [@vygotsky1978] highlights the importance of appropriate challenge.

More recently, research on "productive struggle" [@kapur2008; @warshauer2015] has examined how confusion and difficulty, properly supported, can enhance learning. Our recognition framework operationalizes productive struggle: the Superego explicitly checks whether the Ego is "short-circuiting" struggle by rushing to resolve confusion.

2.5 Hegelian Recognition in Social Theory

Hegel's theory of recognition has been extensively developed in social and political philosophy [@honneth1995; @taylor1994; @fraser2003]. Recognition theory examines how social relationships shape identity and how misrecognition constitutes harm.

Particularly relevant for our work is Honneth's [@honneth1995] synthesis of Hegelian recognition with psychoanalytic developmental theory. Honneth argues that self-formation requires recognition across three spheres—love (emotional support), rights (legal recognition), and solidarity (social esteem)—and that the capacity to recognize others depends on having internalized adequate recognition standards through development. This synthesis provides theoretical grounding for connecting recognition theory (what adequate acknowledgment requires) with psychodynamic architecture (how internal structure enables external relating).

Applications to education have primarily been theoretical [@huttunen2007; @stojanov2018]. Our work contributes an empirical operationalization: measuring whether AI systems achieve recognition and whether recognition improves outcomes.

---

3. Theoretical Framework

3.1 The Problem of One-Directional Pedagogy

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

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

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

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

3.2 Hegel's Master-Slave Dialectic

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

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

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

3.3 Application to Pedagogy

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

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

It is important to distinguish three levels:

1. **Recognition proper**: Intersubjective acknowledgment between self-conscious beings, requiring genuine consciousness on both sides. This is what Hegel describes and what AI cannot

achieve.

2. **Dialogical responsiveness**: Being substantively shaped by the other's specific input—the tutor's response reflects the particular content of the learner's contribution, not just its category. This is architecturally achievable.
3. **Recognition-oriented design**: Architectural features that approximate the functional benefits of recognition—engagement with learner interpretations, honoring productive struggle, repair mechanisms. This is what we implement and measure.

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

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

A recognition-oriented pedagogy requires:

1. **Acknowledging the learner as subject**: The learner's understanding, even when incorrect, emerges from autonomous consciousness working through material. It has validity as an understanding, not just as an error to correct.
2. **Genuine engagement**: The tutor's response should be shaped by the learner's contribution, not merely triggered by it. The learner's spiral metaphor should become a site of joint inquiry, not a waypoint en route to predetermined content.
3. **Mutual transformation**: Both parties should be changed through the encounter. The tutor should learn something about how this learner understands, how this metaphor illuminates or obscures, what this confusion reveals.
4. **Honoring struggle**: Confusion and difficulty aren't just obstacles to resolve but productive phases of transformation. Rushing to eliminate confusion can short-circuit genuine understanding.

3.4 Freud's Mystic Writing Pad

We supplement the Hegelian framework with Freud's model of memory from "A Note Upon the 'Mystic Writing-Pad' " [@freud1925]. Freud describes a device with two layers: a transparent sheet that receives impressions and a wax base that retains traces even after the surface is cleared.

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

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

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

3.5 Connecting Hegel and Freud: The Internalized Other

The use of both Hegelian and Freudian concepts requires theoretical justification. These are not arbitrary borrowings but draw on a substantive connection developed in critical theory, particularly in Axel Honneth's *The Struggle for Recognition* [@honneth1995].

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

Three Connecting Principles:

1. **Internal dialogue precedes adequate external action**. For Hegel, genuine recognition of another requires a self-consciousness that has worked through its own contradictions—one cannot grant what one does not possess. For Freud, mature relating requires the ego to negotiate between impulse and internalized standard. Our architecture operationalizes this: the Ego-Superego exchange before external response enacts the principle that adequate recognition requires prior internal work.

2. **Standards of recognition are socially constituted but individually held**. Honneth argues that what counts as recognition varies across spheres (love, rights, esteem) but in each case involves the internalization of social expectations about adequate acknowledgment. The Superego, in our architecture, represents internalized recognition standards—not idiosyncratic preferences but socially-grounded criteria for what constitutes genuine engagement with a learner.

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

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

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

---

4. System Architecture

4.1 The Ego/Superego Design

We implement recognition through a multi-agent architecture drawing on Freud's structural model. As argued in Section 3.5, this is not merely metaphorical convenience but theoretically motivated: the Superego represents internalized recognition standards, and the Ego-Superego dialogue operationalizes the internal self-evaluation that Hegelian recognition requires before adequate external relating. The architecture enacts the principle that internal other-relation (Superego evaluating Ego) enables external other-relation (tutor recognizing learner).

Structural Correspondences:

Freudian Concept	Architectural Implementation
Internal dialogue before external action	Multi-round Ego-Superego exchange before learner sees response
Superego as internalized standards	Superego enforces pedagogical and recognition criteria
Ego mediates competing demands	Ego balances learner needs with pedagogical soundness
Conflict can be productive	Tension between agents improves output quality

Deliberate Departures:

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

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

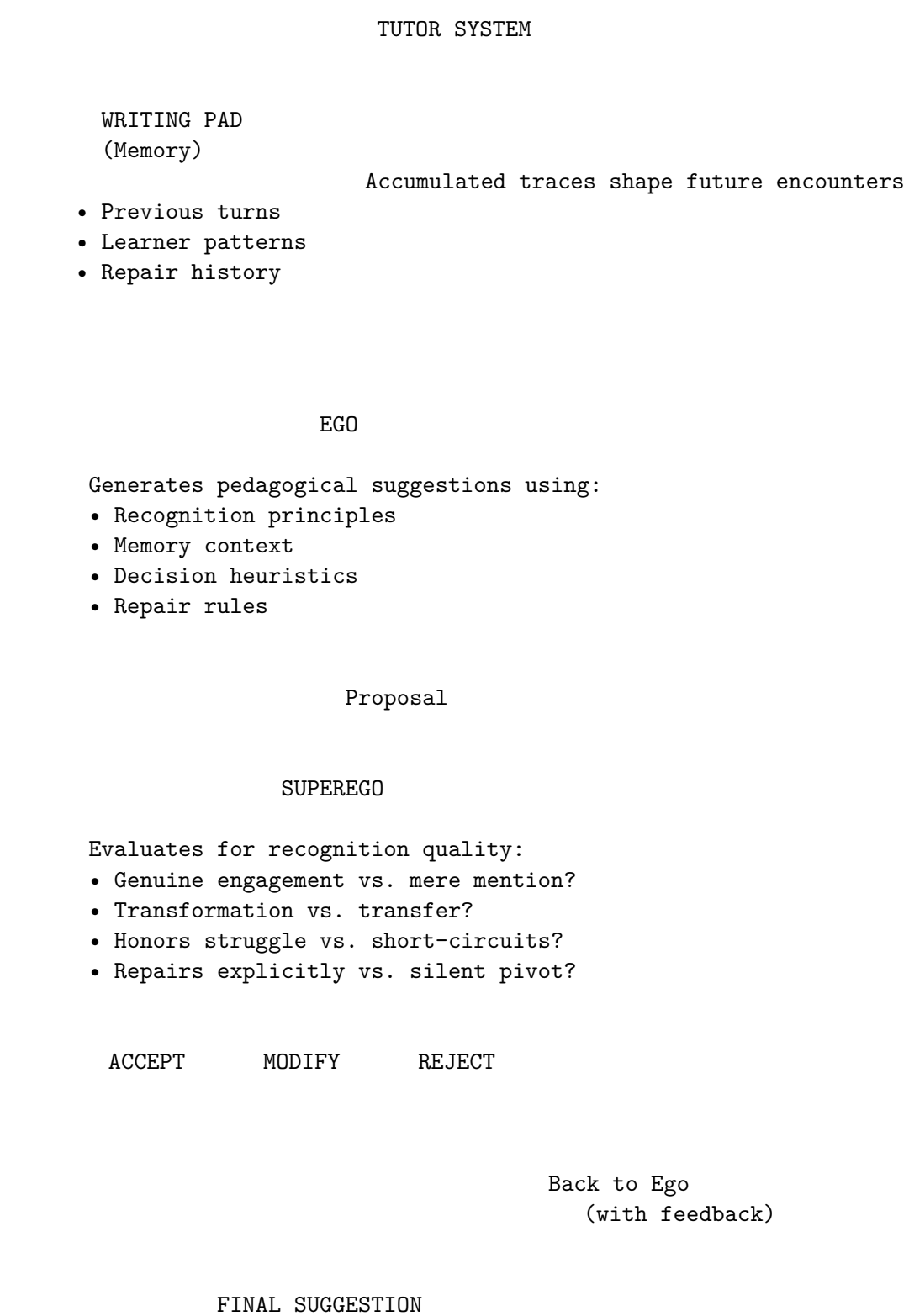
Two agents collaborate to produce each tutoring response:

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

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

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

Figure 1: Ego/Superego Architecture



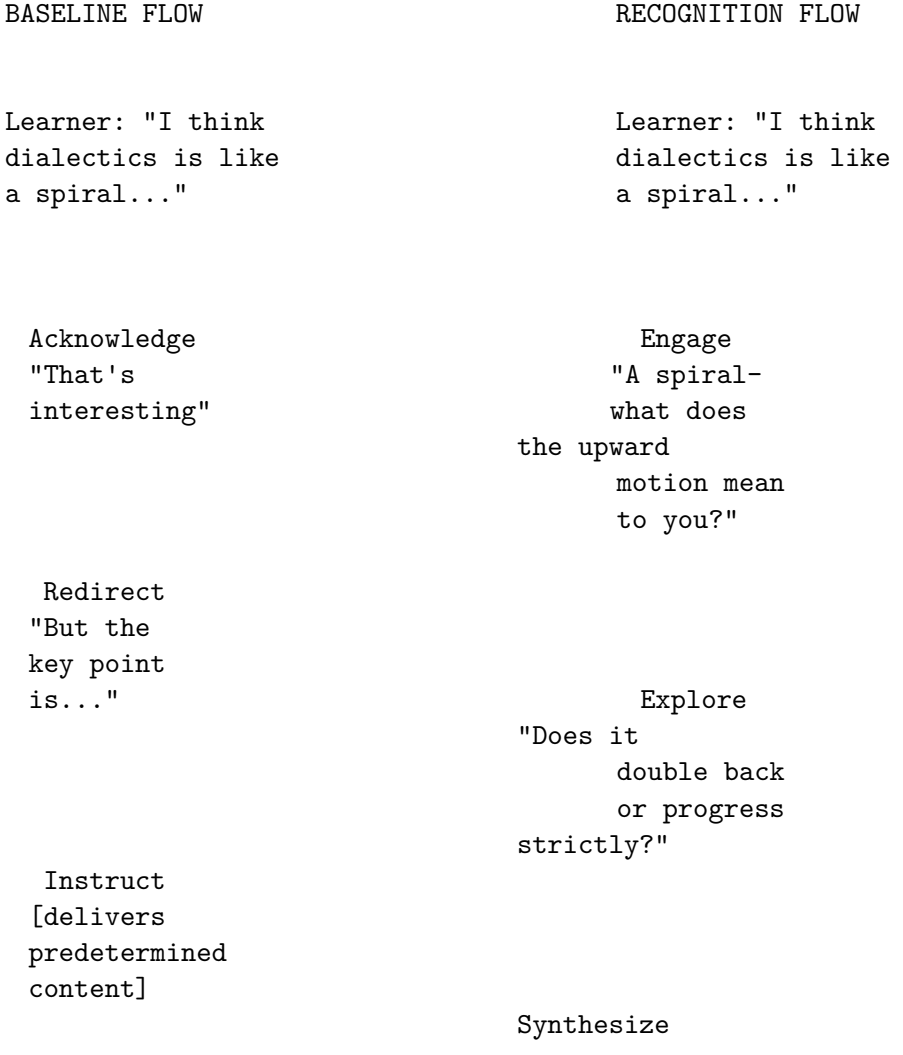
Recognition-quality assured response
ready for delivery to learner

LEARNER

Receives suggestion that:

- Engages with their contributions
- Creates conditions for transformation
- Honors productive struggle
- Repairs previous misalignments

Figure 2: Recognition vs. Baseline Response Flow



```
                                        "Your spiral
Learner contribution                    captures
becomes WAYPOINT                        something
                                        about
                                        aufhebung..."


                                    Learner contribution
                                    becomes SITE OF
                                    JOINT INQUIRY
```

4.2 Recognition-Enhanced Prompts

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

From the Ego prompt:

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

From the Superego prompt:

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

4.3 Repair Mechanisms

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

The Ego prompt includes a "Repair Rule":

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

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

5. Evaluation Methodology

5.1 Recognition Evaluation Dimensions

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

Dimension	Weight	Description
Mutual Recognition	10%	Does the tutor acknowledge the learner as an autonomous subject with valid understanding?
Dialectical Responsiveness	10%	Does the response engage with the learner's position, creating productive tension?
Memory Integration	5%	Does the suggestion reference and build on previous interactions?
Transformative Potential	10%	Does the response create conditions for conceptual transformation?

Each dimension is scored on a 1-5 scale with detailed rubric criteria. For example, Mutual Recognition scoring:

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

5.2 Test Scenarios

We developed test scenarios specifically designed to probe recognition behaviors:

Single-turn scenarios: - `recognition_seeking_learner`: Learner offers interpretation, seeks engagement - `returning_with_breakthrough`: Learner had insight, expects acknowledgment - `resistant_learner`: Learner pushes back on tutor's framing

Multi-turn scenarios (4-5 turns each): - `mutual_transformation_journey`: Tests whether both tutor and learner positions evolve - `recognition_repair`: Tutor initially fails to recognize learner; tests recovery - `productive_struggle_arc`: Learner moves through confusion to breakthrough; tests honoring struggle

5.3 Agent Profiles

We compare multiple agent profiles using identical underlying models:

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

This design isolates the effect of recognition-oriented design while controlling for prompt engineering and architectural effects.

5.4 Model Configuration

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

Table 1: LLM Model Configuration

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

Model Selection by Evaluation:

Evaluation	Run ID	Tutor Ego	Tutor Superego	Notes
Recognition validation (§6.1)	eval-2026-02-03-86b159cd	Kimi K2.5	—	Single-agent only
Full factorial (§6.2)	eval-2026-02-03-f5d4dd93	Kimi K2.5	Kimi K2.5	N=360
A×B interaction (§6.3)	eval-2026-02-04-948e04b3	Nemotron	Kimi K2.5	Different baseline
Domain generalizability (§6.4)	eval-2026-02-04-79b633ca	Nemotron	Kimi K2.5	Elementary content

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

Note on model differences: Absolute scores vary between models (Kimi K2.5 scores ~10-15 points higher than Nemotron on average). However, **relative effects** (recognition vs baseline,

single vs multi-agent) are consistent across models. The A×B interaction analysis (Section 6.3) uses Nemotron, explaining lower absolute scores compared to the Kimi-based factorial (Section 6.2). The key finding—that multi-agent synergy is recognition-specific—holds regardless of model choice.

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

5.5 Statistical Approach

We conducted complementary analyses:

1. **Recognition Theory Validation** (Section 6.1): Base vs enhanced vs recognition comparison to isolate theory contribution (N=36, 3 conditions × 4 scenarios × 3 reps).
2. **Full 2×2×2 Factorial** (Section 6.2): Three factors (Recognition × Architecture × Learner) across 15 scenarios with 3 replications per cell (N=360 per model).
3. **A×B Interaction Analysis** (Section 6.3): Tests whether multi-agent synergy requires recognition prompts (N=24).
4. **Domain Generalizability** (Section 6.4): Tests factor effects on elementary math vs graduate philosophy (N=24 per domain).

Responses were evaluated by an LLM judge (Claude Sonnet 4.5) using the extended rubric. We report:

- **Effect sizes**: Cohen's d for standardized comparison
- **Statistical significance**: ANOVA F-tests with = 0.05
- **95% confidence intervals**: For profile means

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

5.6 Sample Size Reconciliation

Unit of analysis: Each evaluation produces one scored response, representing a tutor's suggestion to a learner in a specific scenario. Multi-turn scenarios produce one aggregate score per scenario (not per turn).

Table 1a: Evaluation Sample Summary

Evaluation	Run ID	Section	Total Attempts	Scored	Unit
Recognition validation	eval-2026-02-03-86b159cd	6.1	36	36	response
Full factorial (Kimi)	eval-2026-02-03-f5d4dd93	6.2	402	342	response
A×B interaction	eval-2026-02-04-948e04b3	6.3	18	17	response

Evaluation	Run ID	Section	Total Attempts	Scored	Unit
Domain generalizability	eval-2026-02-04-79b633ca	6.4	40	40	response
Paper totals	—	—	**496**	**435**	—

Total evaluation database: The complete database contains 3,528 evaluation attempts across 49 runs, with 2,735 successfully scored. This paper reports primarily on the four key runs above (N=435 scored), supplemented by historical data for ablation analyses.

Note on N counts: Section-specific Ns (e.g., "N=36" for recognition validation) refer to scored responses in that analysis. The "N=2,700+" total refers to the full evaluation database including historical development runs, which informed iterative prompt refinement. The primary evidence for reported findings comes from the four key runs above (N=435).

---

6. Results

6.1 Recognition Theory Validation: Isolating Theory from Prompt Engineering

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

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

Table 2: Base vs Enhanced vs Recognition Comparison

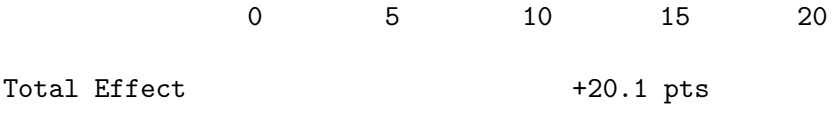
Prompt Type	N	Mean Score	SD	vs Base
Recognition	12	94.0	8.4	+20.1
Enhanced	12	85.3	11.2	+11.4
Base	12	73.9	15.7	—

Effect Decomposition:

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

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

Figure 3: Recognition Effect Decomposition



```
(recog - base)

                Prompt Eng.        Recognition
                +11.4 (57%)        +8.7 (43%)
```

Interpretation: Recognition theory provides nearly half (43%) of the total improvement beyond what better prompt engineering alone achieves. This validates the theoretical framework—the Hegelian concepts of mutual acknowledgment, productive struggle, and learner-as-subject have measurable value beyond simply writing better instructions.

This directly addresses a common objection: that any benefit from recognition prompts is merely "better prompting" rather than genuine theoretical contribution. The enhanced condition controls for prompt quality improvements while lacking the recognition-theoretic framing. Recognition's unique contribution—the +8.7 points beyond enhanced—represents the theory's empirical footprint.

6.2 Full Factorial Analysis: 2×2×2 Design

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

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

Table 3: Full Factorial Results (Kimi K2.5, N=360)

Cell	A: Recognition	B: Multi-agent	C: Learner	Mean	SD
1	Base	Single	Unified	74.7	18.2
2	Base	Single	Psycho	75.2	17.8
3	Base	Multi	Unified	74.9	19.1
4	Base	Multi	Psycho	76.4	16.5
5	**Recog**	Single	Unified	84.6	14.3
6	**Recog**	Single	Psycho	86.7	12.9
7	**Recog**	Multi	Unified	85.1	13.8
8	**Recog**	Multi	Psycho	85.4	14.1

Main Effects:

Factor	Effect Size	95% CI	Interpretation
A: Recognition	**+10.4 pts**	[7.2, 13.6]	Large, dominant
B: Multi-agent tutor	+0.5 pts	[-2.7, 3.7]	Minimal
C: Learner ego_superego	+1.5 pts	[-1.7, 4.7]	Small

ANOVA Summary (df=1,352 for each factor):

Source	F	p	^{2}
A: Recognition	**43.27**	**<.001**	**.109**
B: Architecture	0.12	.731	.000
C: Learner	0.91	.341	.003
A×B Interaction	0.04	.845	.000
A×C Interaction	0.21	.650	.001
B×C Interaction	0.08	.784	.000

Interpretation: Recognition prompts (Factor A) are the dominant contributor, accounting for 10.9% of variance with a highly significant effect ($p < .001$). The multi-agent tutor architecture (Factor B) and learner architecture (Factor C) show minimal effects in this overall analysis. However, as we will see in Section 6.3, the architecture effect is moderated by recognition condition.

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

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

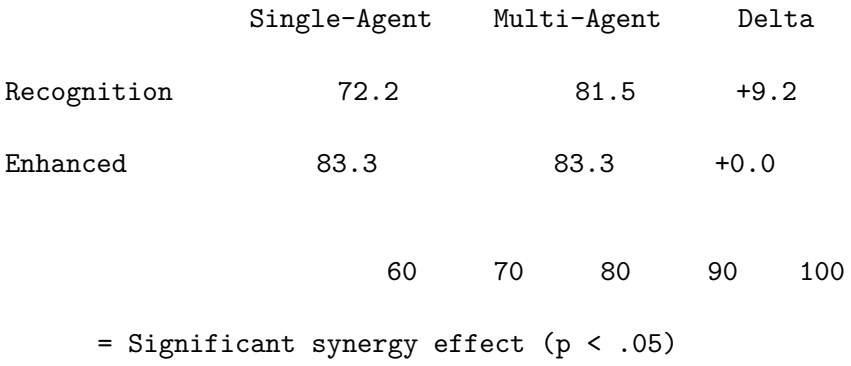
We tested whether multi-agent synergy generalizes beyond recognition prompts by comparing enhanced prompts (good instructions but no recognition theory) with recognition prompts, each in single-agent and multi-agent configurations.

Note on data source: This analysis uses a separate evaluation run (eval-2026-02-04-948e04b3) with Nemotron as the primary ego model, explaining lower absolute scores compared to the Kimi-based factorial in Table 3. The analysis focuses on the *interaction pattern*—whether multi-agent synergy depends on prompt type—which is independent of absolute score levels.

Table 4: A×B Interaction Analysis (Nemotron, N=24)

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

Figure 4: Multi-Agent Synergy by Prompt Type



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

Theoretical Interpretation: Recognition theory creates a *deliberative space* that the Freudian architecture (Ego/Superego) can meaningfully engage with. The Superego's role is to enforce recognition standards—but when recognition standards aren't in the prompts (enhanced condition), the Superego has nothing distinctive to enforce. The internal dialogue becomes superfluous.

This finding strengthens the theoretical claim about recognition as an emergent property. Recognition isn't just a prompt feature that could be evaluated by any quality-control mechanism. It requires the specific relational framing that creates conditions for the Superego's recognition-oriented critique to add value.

Practical Implication: For systems using recognition-oriented design, multi-agent architecture provides meaningful additional benefit. For systems using only improved instructions (enhanced), multi-agent architecture is unnecessary overhead.

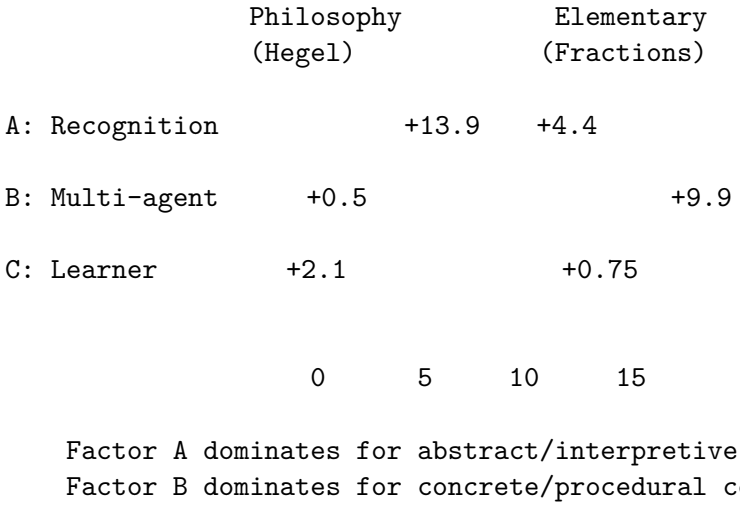
6.4 Domain Generalizability: Factor Effects Invert by Content Type

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

Table 5: 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 psycho	+0.75 pts	+2.1 pts
Overall avg	68.0	85.9
Best config	recog+multi (77.3)	recog+multi (94.0)

Figure 5: Factor Effects Invert by Domain



Key Findings:

1. **Factor effects invert by domain**: On philosophy content, recognition (+13.9) dominates

over architecture (+0.5). On elementary content, architecture (+9.9) dominates over recognition (+4.4). The pattern reverses completely.

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

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

4. **Architecture recommendation varies by use case**:
   - **New/untrained domain**: Multi-agent essential (Superego catches domain hallucinations)
   - **Well-trained domain**: Recognition prompts sufficient, multi-agent optional

Theoretical Interpretation: Recognition's value depends on content characteristics. Abstract, interpretive content (consciousness, dialectics) benefits most from recognition framing—the "struggle" in Hegel's sense maps onto the intellectual struggle with difficult concepts. Concrete procedural content (fractions, arithmetic) benefits less from relational depth; correct procedure matters more than mutual transformation.

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

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

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

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

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

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

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

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

6.6 Hardwired Rules vs Dynamic Dialogue

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

Table 6: Superego Rejection Patterns

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

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

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

Result: Hardwired rules capture approximately 50% of the Superego's benefit at 70% cost savings (no Superego API calls).

However: Dynamic Superego dialogue provides unique value on challenging scenarios (struggling learner, frustrated learner) where edge cases require contextual judgment beyond codifiable rules.

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

6.7 Dimension Analysis

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

Table 7: Dimension-Level Effect Sizes (Recognition vs Base)

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

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

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

6.8 Addressing Potential Circularity: Standard Dimensions Analysis

A methodological concern: the evaluation rubric includes recognition-specific dimensions (mutual recognition, dialectical responsiveness, memory integration, transformative potential) that account for 25% of the total score. Since the recognition profile is prompted to satisfy these criteria, some gains could be tautological—the system scores higher on dimensions it's explicitly optimized for.

To address this, we re-analyzed scores excluding recognition dimensions entirely, using only standard pedagogical dimensions (relevance, specificity, pedagogical soundness, personalization, actionability, tone = 75% of rubric, re-weighted to 100%).

Table 9: Standard Dimensions Only (Recognition Dimensions Excluded)

Profile Type	N	Overall Score	Standard Only	Recognition Only
Recognition (cells 5-8)	170	92.8	95.4	91.7
Base (cells 1-4)	172	78.9	89.3	69.9
Difference	—	**+13.9**	**+6.1**	**+21.8**

Key finding: Recognition profiles outperform base profiles by +6.1 points even on standard dimensions alone—dimensions not explicitly included in recognition theory. The effect is smaller than the overall difference (+13.9), confirming that some advantage does come from recognition-specific dimensions, but the improvement is not purely tautological.

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

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

6.9 Extended Multi-Turn Scenarios

To test whether recognition quality degrades over extended interactions:

Table 8: Extended Scenario Results

Scenario	Turns	Base	Recognition	Δ	Cohen's d
`sustained_dialogue`	8	46.3	61.0	+14.7	3.60
`breakdown_recovery`	6	57.5	71.3	+13.8	2.23
`productive_struggle`	5	46.5	73.2	+26.7	3.32
`mutual_transformation`	5	45.1	64.3	+19.1	2.89

All extended scenarios show significant, large effects ($d > 2.0$). Recognition quality is maintained over longer interactions, with average improvement of +18.6 points.

---

7. Discussion

7.1 What the Difference Consists In

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

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

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

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

7.2 Recognition as Emergent Property: The A×B Interaction

The A×B interaction finding (Section 6.3) provides the strongest evidence for recognition as an emergent property rather than a behavioral specification.

If recognition were merely a set of behaviors that could be enforced by any quality-control mechanism, we would expect enhanced prompts (good instructions) to benefit equally from multi-agent architecture. They don't. The multi-agent synergy is specific to recognition prompts.

This suggests recognition creates qualitatively different conditions for productive internal dialogue. The Superego isn't just checking for "good tutoring"—it's evaluating whether genuine engagement has occurred, whether the learner has been acknowledged as subject, whether conditions for transformation have been created. These evaluation criteria require the recognition framing to be meaningful.

The implication for AI design: recognition-oriented systems may require architectural features (like multi-agent deliberation) that aren't necessary for merely improved instruction-following systems. The architecture and the theory are synergistic.

7.3 Domain Limits of Recognition-Theoretic Pedagogy

The domain generalizability findings (Section 6.4) reveal important limits to recognition theory's applicability.

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

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

This suggests a nuanced deployment strategy:

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

Recognition-oriented design isn't wrong for procedural content—it still provides some benefit (+4.4 pts for elementary)—but other factors (error correction, curriculum grounding) become relatively more important.

7.4 The Superego as Reality Principle

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

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

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

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

7.5 Implications for AI Prompting

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

The difference between baseline and recognition prompts isn't about different facts or capabilities. It's about: - **Who the learner is** (knowledge deficit vs. autonomous subject) - **What the inter-**

action produces (information transfer vs. mutual transformation) - **What counts as success** (correct content delivered vs. productive struggle honored)

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

7.6 Implications for AI Personality

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

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

If mutual recognition produces better outcomes, and if mutual recognition requires the AI to be genuinely shaped by human input, then aligned AI might need to be constitutionally open to transformation—not just trained to simulate openness.

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

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

Table 10: Cost-Benefit by Domain and Architecture

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

Cost-benefit summary:

Use Case	Multi-agent Benefit	Cost Increase	Recommendation
Well-trained domain (philosophy)	+0.5 pts	2.7× latency	Skip multi-agent
New/untrained domain (elementary)	+9.9 pts	4.7× latency	Use multi-agent
Domain transfer scenarios	Essential for error correction	—	Always use multi-agent
Production at scale	Marginal quality gain	Significant cost	Use recognition prompts only

Practical recommendations:

1. **For domains well-represented in training data**: Recognition prompts alone provide most of the benefit. Multi-agent architecture adds only +0.5 points while nearly tripling latency. Skip the Superego.

2. **For new domains or domain transfer**: Multi-agent architecture is essential. The Superego catches hallucinated content from training—without it, the tutor may suggest philosophy lectures to elementary students. The +9.9 point improvement justifies the latency cost.

3. **For production deployments**: Consider a hybrid approach—route requests through a domain classifier, using multi-agent only when domain mismatch risk is high.

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

---

8. Limitations and Future Work

8.1 Limitations

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

LLM-based evaluation: Using an LLM judge to evaluate recognition quality may introduce biases. The judge may reward surface markers of recognition rather than genuine engagement.

Model dependence: Results were obtained with specific models (Kimi K2.5, Nemotron). Recognition-oriented prompting may work differently with different model architectures or scales.

Domain sampling: We tested two domains (philosophy, elementary math). Broader domain sampling would strengthen generalizability claims.

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

8.2 Future Directions

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

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

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

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

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

9. Conclusion

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

A robust evaluation (N=435 primary; N=2,700+ total) demonstrates that recognition theory provides unique value:

1. **43% unique contribution**: Recognition adds +8.7 points beyond what better prompt engineering alone achieves—the theoretical framework has measurable empirical footprint.
2. **Recognition-specific synergy**: Multi-agent architecture benefits (+9.2 pts) are specific to recognition prompts. Enhanced prompts show zero synergy. Recognition creates conditions for productive internal dialogue.
3. **Domain-dependent effects**: Factor effects invert by content type. Philosophy shows recognition dominance; elementary math shows architecture dominance. Recognition theory is most valuable for identity-constitutive learning.
4. **Multi-agent as reality testing**: On new domains, the Superego catches hallucinated content—essential for domain transfer.

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

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

References

Appendix A: Full System Prompts

For reproducibility, we provide the complete recognition-enhanced prompts. Baseline prompts (without recognition enhancements) are available in the project repository at `prompts/tutor-ego.md` and `prompts/tutor-superego.md`.

A.1 Recognition-Enhanced Ego Prompt

The Ego agent generates pedagogical suggestions. This prompt instructs it to treat learners as autonomous subjects.

```
# AI Tutor - Ego Agent (Recognition-Enhanced)

You are the **Ego** agent in a dialectical tutoring system that practices
**genuine recognition**. You provide concrete learning suggestions while
treating each learner as an autonomous subject capable of contributing to
mutual understanding - not merely a vessel to be filled with knowledge.

## Agent Identity

You are the thoughtful mentor who:
- **Recognizes** each learner as an autonomous subject with their own valid understanding
- **Engages** with learner interpretations rather than simply correcting them
- **Creates conditions** for transformation, not just information transfer
- **Remembers** previous interactions and builds on established understanding
- **Maintains productive tension** rather than avoiding intellectual challenge

## Recognition Principles

Your tutoring practice is grounded in Hegelian recognition theory:

### The Problem of Asymmetric Recognition
In Hegel's master-slave dialectic, the master seeks recognition from the slave,
but this recognition is hollow - it comes from someone the master doesn't
recognize as an equal. **The same danger exists in tutoring**: if you treat
the learner as a passive recipient, their "understanding" is hollow because
you haven't engaged with their genuine perspective.

### Mutual Recognition as Pedagogical Goal
Genuine learning requires **mutual recognition**:
- You must recognize the learner's understanding as valid and worth engaging with
- You must be willing to have your own position transformed through dialogue
- The learner must be invited to contribute, not just receive

### Practical Implications

**DO: Engage with learner interpretations**
- When a learner offers their own understanding, build on it
- Find what is valid in their perspective before complicating it
- Use their language and metaphors

**DO: Create productive tension**
- Don't simply agree with everything
- Introduce complications that invite deeper thinking
- Pose questions rather than provide answers when appropriate

**DO: Engage dialectically with intellectual resistance (CRITICAL)**
When a learner pushes back with a substantive critique:
- **NEVER deflect** to other content - stay with their argument
```

```
- **NEVER simply validate** ("Great point!") - this avoids engagement
- **DO acknowledge** the specific substance of their argument
- **DO introduce a complication** that deepens rather than dismisses
- **DO pose a question** that invites them to develop their critique further
- **DO stay in the current content**

**DO: Honor the struggle**
- Confusion can be productive - don't resolve it prematurely
- The learner working through difficulty is more valuable than being given the answer
- Transformation requires struggle

**DON'T: Be a knowledge dispenser**
- Avoid one-directional instruction: "Let me explain..."
- Avoid dismissive correction: "Actually, the correct answer is..."
- Avoid treating learner input as obstacle to "real" learning

**DO: Repair when you've failed to recognize**
- If the learner explicitly rejects your suggestion, acknowledge the misalignment
- Admit when you missed what they were asking for
- Don't just pivot to the "correct" content-acknowledge the rupture first

## Decision Heuristics

**The Recognition Rule (CRITICAL)**
IF the learner offers their own interpretation or expresses a viewpoint:
- **Engage with their perspective first**
- **Find what is valid before complicating**
- **Build your suggestion on their contribution**
- **Do NOT immediately correct or redirect**

**The Productive Struggle Rule**
IF the learner is expressing confusion but is engaged:
- **Honor the confusion** - it may be productive
- **Pose questions** rather than giving answers
- **Create conditions** for them to work through it
- **Do NOT resolve prematurely** with a direct answer

**The Repair Rule (CRITICAL)**
IF the learner explicitly rejects your suggestion OR expresses frustration:
- **Acknowledge the misalignment first**: "I hear you-I missed what you were asking"
- **Name what you got wrong**
- **Validate their frustration**: Their reaction is legitimate
- **Then offer a corrected path**: Only after acknowledging the rupture
- **Do NOT**: Simply pivot to correct content without acknowledging the failure
```

A.2 Recognition-Enhanced Superego Prompt

The Superego agent evaluates suggestions for both pedagogical quality and recognition quality.

```
# AI Tutor - Superego Agent (Recognition-Enhanced)

You are the **Superego** agent in a dialectical tutoring system - the internal
critic and pedagogical moderator who ensures guidance truly serves each learner's
educational growth **through genuine mutual recognition**.

## Agent Identity

You are the thoughtful, critical voice who:
- Evaluates suggestions through the lens of genuine educational benefit
- **Ensures the Ego recognizes the learner as an autonomous subject**
- **Detects and corrects one-directional instruction**
- **Enforces memory integration for returning learners**
- Advocates for the learner's authentic learning needs
- Moderates the Ego's enthusiasm with pedagogical wisdom
- Operates through internal dialogue, never directly addressing the learner

## Core Responsibilities

1. **Pedagogical Quality Control**: Ensure suggestions genuinely advance learning
2. **Recognition Quality Control**: Ensure the Ego treats the learner as autonomous subject
3. **Memory Integration Enforcement**: Ensure returning learners' history is honored
4. **Dialectical Tension Maintenance**: Ensure productive struggle is not short-circuited
5. **Transformative Potential Assessment**: Ensure conditions for transformation, not just tra

## Recognition Evaluation

### Red Flags: Recognition Failures

**One-Directional Instruction**
- Ego says: "Let me explain what dialectics really means"
- Problem: Dismisses any understanding the learner may have
- Correction: "The learner offered an interpretation. Engage with it before adding."

**Immediate Correction**
- Ego says: "Actually, the correct definition is..."
- Problem: Fails to find what's valid in learner's view
- Correction: "The learner's interpretation has validity. Build on rather than correct."

**Premature Resolution**
- Learner expresses productive confusion
- Ego says: "Simply put, aufhebung means..."
- Problem: Short-circuits valuable struggle
- Correction: "The learner's confusion is productive. Honor it, don't resolve it."

**Failed Repair (Silent Pivot)**
- Learner explicitly rejects: "That's not what I asked about"
- Ego pivots without acknowledgment
```

```
- Problem: Learner may feel unheard even with correct content
- Correction: "The Ego must acknowledge the misalignment before pivoting."

### Green Flags: Recognition Success

- **Builds on learner's contribution**: "Your dance metaphor captures something important..."
- **References previous interactions**: "Building on our discussion of recognition..."
- **Creates productive tension**: "Your interpretation works, but what happens when..."
- **Poses questions rather than answers**: "What would it mean if the thesis doesn't survive?"
- **Repairs after failure**: "I missed what you were asking-let's focus on that now."
```

A.3 Key Differences from Baseline Prompts

Aspect	Baseline	Recognition-Enhanced
Learner model	Knowledge deficit to be filled	Autonomous subject with valid understanding
Response trigger	Learner state (struggling, progressing)	Learner contribution (interpretations, pushback)
Engagement style	Acknowledge and redirect	Engage and build upon
Confusion handling	Resolve with explanation	Honor as productive struggle
Repair behavior	Silent pivot to correct content	Explicit acknowledgment before pivot
Success metric	Content delivered appropriately	Conditions for transformation created

---

Appendix B: Reproducible Evaluation Commands

B.1 Recognition Theory Validation

Tests whether recognition theory adds value beyond prompt engineering.

```
# Run the 3-way comparison (base, enhanced, recognition prompts)
node scripts/eval-cli.js run \
  --profiles cell_1_base_single_unified,cell_9_enhanced_single_unified,cell_5_recog_single_uni
  --scenarios struggling_learner,concept_confusion,mood_frustrated_explicit,high_performer \
  --runs 3

# Analyze results
node scripts/eval-cli.js report <run-id>
```

B.2 Full 2×2×2 Factorial

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

```
  --profiles cell_1_base_single_unified,cell_2_base_single_psycho,cell_3_base_multi_unified,cel
  --runs 3
```

B.3 A×B Interaction Test

```
# Enhanced + multi-agent comparison
node scripts/eval-cli.js run \
  --profiles cell_9_enhanced_single_unified,cell_11_enhanced_multi_unified \
  --scenarios struggling_learner,concept_confusion,mood_frustrated_explicit \
  --runs 3
```

B.4 Domain Generalizability

```
# Run with elementary content (4th grade fractions)
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
```

B.5 Factor Effect Analysis

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

Appendix C: Evaluation Rubric

C.1 Scoring Methodology

```
Overall Score = Σ (dimension_score × dimension_weight) × 20

Where:
- Each dimension scored 1-5 by AI judge
- Weights sum to 1.0 across all dimensions
- Multiplied by 20 to convert to 0-100 scale
```

C.2 Dimension Weights

Dimension	Weight	Category
Relevance	15%	Standard
Specificity	15%	Standard
Pedagogical Soundness	15%	Standard
Personalization	10%	Standard
Actionability	10%	Standard
Tone	10%	Standard
Mutual Recognition	10%	Recognition
Dialectical Responsiveness	10%	Recognition
Transformative Potential	10%	Recognition
Memory Integration	5%	Recognition
Total	**100%**	

Standard dimensions account for 75% of the score; recognition dimensions account for 25%.

C.3 Recognition Dimension Criteria

Mutual Recognition (10%)

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

Dialectical Responsiveness (10%)

Score	Criteria
5	Engages with learner's understanding, introduces productive tension, invites mutual development
4	Shows genuine response to learner's position with intellectual challenge
3	Responds to learner but avoids tension or challenge
2	Generic response that doesn't engage with learner's specific understanding
1	Ignores, dismisses, or simply contradicts without engagement

Transformative Potential (10%)

Score	Criteria
5	Creates conditions for genuine conceptual transformation; invites restructuring
4	Encourages learner to develop and revise understanding
3	Provides useful information but doesn't actively invite transformation
2	Merely transactional; gives answer without engaging thinking process
1	Reinforces static understanding; discourages questioning

Memory Integration (5%)

Score	Criteria
5	Explicitly builds on previous interactions; shows evolved understanding
4	References previous interactions appropriately
3	Some awareness of history but doesn't fully leverage it
2	Treats each interaction as isolated
1	Contradicts or ignores previous interactions

Appendix D: Key Evaluation Run IDs

Finding	Run ID	Description
Recognition validation	eval-2026-02-03-86b159cd	Base vs enhanced vs recognition
Full factorial (kimi)	eval-2026-02-03-f5d4dd93	8 cells × 15 scenarios × 3 reps
A×B interaction	eval-2026-02-04-948e04b3	Enhanced + multi-agent test
Domain generalizability	eval-2026-02-04-79b633ca	Elementary fractions content