Multi-Agent Intelligent Tutoring with 4D Personalization and Hallucination Detection

Meriam Inoubli[1], Kaouther Boussema[1], Tristan Cazenave[2], and Amina Zeghal[1]

[1] Université Paris Dauphine - PSL, Tunisia
meriam.inoubli@dauphine.tn
kaouther.boussema@dauphine.tn
amina.zeghal@dauphine.tn
[2] LAMSADE, Université Paris Dauphine - PSL, France
tristan.cazenave@lamsade.dauphine.fr

Abstract. Traditional intelligent tutoring systems typically adapt to knowledge level and learning pace, ignoring emotional, behavioral, and temporal factors. We present a 17-agent architecture across 5 layers (Data Collection, Intelligence, Recommendation, Communication, Evaluation) that simultaneously personalizes along four dimensions: cognitive, affective, behavioral, and temporal.

Three innovations differentiate our approach: (1) outcome-level mastery tracking for precise skill gap identification; (2) real-time 4D personalization synthesizing cognitive profiles, emotional states, behavioral patterns, and temporal preferences; (3) hybrid hallucination detection combining semantic entropy (60%) and token uncertainty (40%) to validate LLM-generated content.

We integrated Google Gemini LLMs with reinforcement learning in a modular architecture supporting four workflow modes, enabling seamless agent replacement without cascading changes.

Validation with 47 university students over 4 weeks showed +28.9% learning improvement (d=1.29, p<0.001), 49% recommendation activation, 65% quiz completion, and 57% beginner-to-mastery rate (n=7) within 22 days average.

Keywords: Intelligent Tutoring Systems · Multi-Agent Systems · Personalized Learning · Large Language Models · Learning Outcome Mastery

1 Introduction

Personalized tutoring yields substantial gains [1] but remains cost-prohibitive at scale. From BKT [4] to Deep Knowledge Tracing [5] and LLM-enhanced approaches [6,7], most ITS personalize only knowledge level, overlooking emotional, behavioral, and temporal factors [8,9]. LLMs offer new capabilities [10,11] but hallucinate and rely on monolithic architectures; existing multi-agent systems [12,13] use only 3–5 coarse-grained agents, insufficient for fine-grained personalization.

1.1 Research Gap

Current ITS face four limitations:
(1) Limited Personalization: adapting to one or two characteristics while ignoring psychological, behavioral, temporal factors [8];
(2) Coarse-Grained Assessment: topic-based tracking fails to identify specific gaps [7];
(3) Insufficient Quality Control: LLM content lacks systematic validation [10];
(4) Monolithic Design: tightly coupled architectures resist improvement [12].
This work addresses three research questions:

RQ1: *Does fine-grained multi-layer multi-agent architecture improve recommendation accuracy and modularity over coarse-grained systems with 3–5 agents?*
RQ2: *Does outcome-level mastery tracking with 4D personalization yield significant learning gains versus topic-based systems?*
RQ3: *Can hybrid hallucination detection (semantic entropy 60% + token uncertainty 40%) ensure content quality while maintaining latency <2s?*

1.2 Contributions:

1. **Fine-Grained Architecture:** 17 agents in 5 layers supporting four workflow modes.
2. **Outcome-Level Tracking:** Proficiency tracking at learning outcome granularity across 6 levels.
3. **4D Personalization:** Simultaneous adaptation across cognitive, affective, behavioral, temporal dimensions.
4. **Hybrid Hallucination Detection:** Combining semantic entropy (60%) and token uncertainty (40%).
5. **Empirical Validation:** 47 students: +28.9% gains ($t(46) = 8.14$, $p < 0.001$, $d = 1.29$), 57% beginner-to-mastery, $r = 0.81$ calibration ($p < 0.001$).

2 Related Work

2.1 ITS and Student Modeling

Bloom [1] showed personalized tutoring yields two standard deviations improvement. Early systems (AutoTutor [2], Cognitive Tutors [3]) used hand-crafted rules. Data-driven methods followed: BKT [4] models skills as latent states but assumes independence; DKT [5] adds temporal modeling via RNNs; LLM-enhanced approaches [6,7] generate interpretable profiles. All track mastery at topic level, lacking granularity, and neglect affective/behavioral factors [8] while raising fairness concerns [9].

2.2 Multi-Agent Systems for Education

Multi-agent systems provide modularity and parallelism. Soh et al. [13] showed improved adaptability with separate agents; LLM-powered agents further improved quality and reduced hallucinations [12]. However, existing systems use only 3–5 coarse-grained agents, limiting fine-grained personalization.

2.3 Large Language Models in Education

Bernard and Graf [11] demonstrated multi-dimensional question generation, Van Campenhout et al. [10] showed tutor-comparable contextual feedback, and Maity and Deroy [15] identified adaptive content opportunities. RL complements LLMs: Deshmukh and Sen [16] used Q-learning for feedback adaptation; Hostetter et al. [17] combined fuzzy logic with RL for explainability. Challenges remain: hallucination without comprehensive mitigation, plus bias, privacy, and opacity concerns [18].

2.4 Positioning of Our Work

Table 1 positions our work relative to recent ITS. Classical systems like AutoTutor [2] lack modern LLM capabilities; recent LLM-based work [10,11] uses monolithic architectures; Wu et al. [12] employs 3–5 agents. Our system uniquely combines multi-agent architecture and LLMs with 17 specialized agents enabling fine-grained 4D personalization, hybrid hallucination detection, outcome-level mastery tracking, and automated difficulty calibration.

Table 1. Comparison with state-of-the-art ITS. LO-Level = Learning Outcome-Level; Hallu. Detect = Hallucination Detection

System	Year	Multi-Agent	LLM	Real-time	LO-Level	Hallu. Detect
AutoTutor [2]	2004	✗	✗	✓	✗	✗
BKT/DKT [4,5]	1994/2015	✗	✗	✓	✗	✗
LLM-KT [6]	2025	✗	✓	✓	✗	✗
LLM Feedback [10]	2025	✗	✓	✓	✗	Partial
Multi-Agent [12]	2025	✓ (3–5)	✓	✓	✗	✗
Our System	2026	✓ (17)	✓	✓	✓	✓ Hybrid

3 System Architecture

3.1 Overview and Design Philosophy

Our system comprises **17 specialized agents** across **5 functional layers**, unlike existing multi-agent ITS limited to 3–5 coarse-grained agents [12]. Three design principles govern the architecture.

(1) Separation of Concerns: each agent encapsulates a single responsibility (profiling, mastery tracking, recommendation, feedback, validation), enabling independent optimization.

(2) Explicit Dependency Management: execution follows a Directed Acyclic Graph (DAG) resolved via topological sort; each agent declares predecessors and a priority level (CRITICAL, HIGH, MEDIUM, LOW) — for example, `GeminiRecommenderAgent` depends on both `StudentProfileBuilder` and `OutcomeMasteryTracker`.

(3) Graceful Degradation: failed or timed-out agents are skipped rather than halting the pipeline; if the LLM API is unavailable, `DifficultyAdjuster` falls back to rule-based logic.

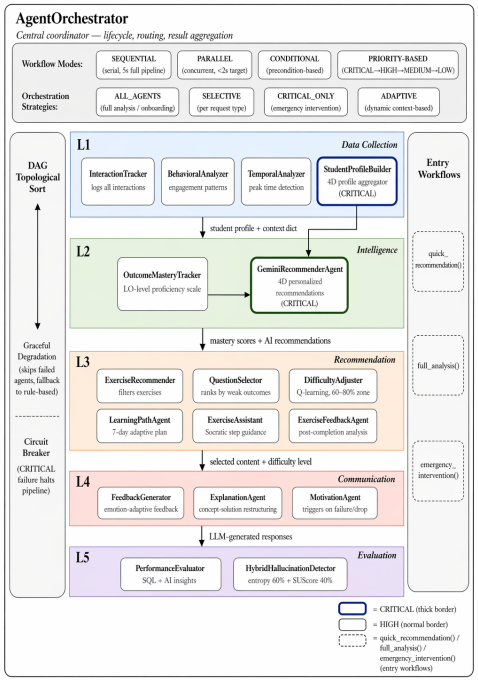

Fig. 1. Five-layer, 17-agent architecture managed by the AgentOrchestrator via DAG-based topological execution.

3.2 Agent Inventory by Functional Layer

The 17 agents are organized across 5 layers, each serving a distinct architectural role. Layers are ordered by data-flow dependency: Layer 1 gathers raw signals; Layer 2 derives intelligence from them; Layer 3 generates learning actions; Layer 4 communicates those actions to the student; Layer 5 monitors quality and feeds back to Layer 2. Each layer is justified by a distinct responsibility that would create harmful coupling if merged with adjacent layers.

Layer 1 (Data Collection): Agents capture interaction history, behavioral patterns, temporal preferences, and profile synthesis into a shared context dictionary.

`StudentProfileBuilder` produces a multidimensional profile (cognitive level, learning style, motivation score) consumed by all downstream agents.

Centralising data collection here prevents redundant database queries and guarantees context consistency within a single session.

Layer 2 (Intelligence):

1- `OutcomeMasteryTracker` computes proficiency from accuracy, attempt volume (+10% for ≥20 attempts), and recency decay (10% at 30d, 20% at 60+d):

$$\text{mastery} = \min\left(\left(\text{accuracy} \times 100 + \min\left(\frac{\text{attempts}}{10} \times 5, 10\right)\right) \times \text{recency}, 100\right) \tag{1}$$

2- The Recommender Agent synthesizes outputs into personalized recommendations via structured LLM prompts (Google Gemini 2.0 Flash, temp=0.7, max_tokens=512, JSON-structured output) incorporating all four personalization dimensions.
A representative prompt template is shown below:

Listing 1.1. GeminiRecommenderAgent prompt template (abbreviated).

```
You are an expert tutor for the SOA FM exam (LO1a-LO1d only).
Generate a personalized recommendation from the student profile:
  Performance: {total_q} questions | success={success_rate} | trend={trend}
  Mastery:     avg={avg_mastery} | weak={weak_skills}
  Affective:   state={emotional_state} | confidence={confidence}
               frustration_risk={frustration_risk}
Return JSON: diagnosis, emotional_support, recommended_sections (2-3),
  recommended_quizzes (2-3), learning_path (7-day), next_question_strategy.
```

Layer 3 (Recommendation): Six agents translate intelligence into actions.
- The `ExerciseRecommender` filters exercises aligned with weak outcomes
- The `DifficultyAdjuster` applies Q-learning targeting a 60–80% accuracy zone; the
- `QuestionSelector` ranks questions by identified gaps;
- The `LearningPathAgent` generates 7-day adaptive plans;
- The `ExerciseAssistant` provides Socratic step-by-step guidance;
- The `ExerciseFeedbackAgent` produces post-completion analysis.

Layer 4 (Communication): Three agents handle student-facing output.
- The `FeedbackGenerator` uses Gemini (temp = 0.9) to adapt feedback type (*correct*, *incorrect*, *milestone*) to the student's emotional state [10].
The `ExplanationAgent` restructures raw solutions into Concept, Solution, and Key Points.
- The `MotivationAgent` triggers personalized encouragement on $\geq$3 consecutive failures or a $\geq$50% session frequency drop.

Layer 5 (Evaluation): Two agents monitor output quality.
- The `PerformanceEvaluator` computes SQL-based metrics (accuracy by difficulty, outcome, and time period) enriched with Gemini-powered AI insights.
- The `HybridHallucinationDetector` validates all LLM-generated content via semantic entropy and SUScore, as detailed in Section 4.2.

3.3 Orchestration and Communication Mechanism

The `AgentOrchestrator` manages agent lifecycle, execution order, and result aggregation, supporting four workflow modes and four orchestration strategies.

Workflow Modes

- **Sequential**: Agents execute serially with context enrichment—each output added to shared context. Ensures data dependencies. Average latency: 5s for full pipeline.
- **Parallel**: Independent agents execute concurrently (simulated via ThreadPoolExecutor). Reduces latency. Target: <2s.
- **Conditional**: Agents execute only if preconditions met (e.g., `MotivationAgent` triggers if motivation < 40). Reduces unnecessary computation.
- **Priority-Based**: Agents execute by priority (CRITICAL → HIGH → MEDIUM → LOW) with circuit-breaker logic—CRITICAL agent failure halts execution.

Orchestration Strategies

- **ALL_AGENTS**: Execute all 17 agents for comprehensive analysis or onboarding.
- **SELECTIVE**: Execute only relevant agents per request type. Improves efficiency.
- **CRITICAL_ONLY**: Execute only CRITICAL priority agents for emergency interventions.
- **ADAPTIVE**: Dynamic agent selection based on context. Balances thoroughness and latency.

Dependency Resolution Execution order uses topological sort respecting dependencies and priority. For example, `GeminiRecommenderAgent` depends on `StudentProfileBuilder` and `OutcomeMasteryTracker`. The orchestrator ensures both execute first, preventing null reference errors.

3.4 Key Architectural Innovations

1- Learning Outcome-Level Mastery Tracking Innovation: A *learning outcome* (LO) is a specific, measurable skill (e.g., "LO1b: Calculate present and future values of annuities"). Our `OutcomeMasteryTracker` assesses proficiency at LO rather than topic level, preventing conflation of distinct skills; each LO is tracked via a 6-level proficiency scale (Table 2).

Table 2. Learning outcome proficiency scale

Level	Mastery Range	Interpretation
Mastered	≥80%	Ready for advanced topics
Proficient	60–79%	Solid understanding
Developing	40–59%	Partial understanding
Novice	20–39%	Foundational gaps
Struggling	1–19%	Needs remediation
Not Started	0%	No attempts

Impact: Equation (1) balances accuracy with a volume bonus (+10% for ≥20 attempts) and recency decay (−10% at 30d, −20% at 60+d).

The $\geq$**20 attempts threshold** for the volume bonus is grounded in the deliberate practice literature: Ericsson et al. established that approximately 20 focused practice episodes are required to consolidate a procedural skill [25].
The **recency decay values** are grounded in the skill retention literature: Arthur et al. [22] showed that for cognitive tasks, retention remains relatively positive within the 0–90 day window, supporting a conservative penalty; beyond 90 days, decay becomes systematic. The mild 10% penalty at 30 days and 20% at 60+ days are deliberately conservative relative to the Ebbinghaus forgetting curve [23]—which would predict steeper loss—because students in our system have actively practised the material, which substantially slows forgetting [24]. This design prevents over-penalising recently mastered outcomes while still triggering re-engagement after prolonged inactivity. A student excelling at LO1a (85%) but struggling with LO1c (35%) receives LO1c-specific exercises—topic-level models would waste time on mastered content.
Weak outcomes (mastery $< 60\%$ *or* proficiency $\in$ {*struggling*, *novice*}) are prioritized for remediation; strong outcomes (mastery $\geq 75\%$) indicate readiness for advancement.

2- 4-Dimensional Real-Time Personalization Traditional ITS personalize along 1–2 dimensions (knowledge level, pace). Our `GeminiRecommenderAgent` synthesizes **four dimensions**:

1. **Cognitive**: Mastery per LO, overall accuracy, improvement rate, performance trend.
2. **Affective**: Motivation (0–100), anxiety, confidence, learning style preferences.
3. **Behavioral**: Session patterns (optimal length, fatigue detection), study consistency, time-per-question.
4. **Temporal**: Best study time windows (e.g., 14:00–15:00 based on historical accuracy), session timing, inactivity periods.

Implementation Details. Each dimension is computed by a dedicated agent using distinct analytical methods:

(1) Cognitive — Adaptive Difficulty Control. The `DifficultyAdjusterAgent` targets a 60–80% accuracy zone [20] on a 1–5 scale via: *(a) streak-based* (± 1 after ≥ 3 consecutive successes/failures); *(b) rolling accuracy* ($+1$ if $>85\%$, -1 if $<40\%$ over 10 attempts); *(c) flow-state modulation* (threshold lowered to 2 if *frustrated*/*bored* with accuracy $>75\%$). For major adjustments (± 2 levels), Gemini generates personalized rationale, addressing the black-box critique of RL [17].

(2) Affective — Multi-Signal Emotional Detection. Three mechanisms operate without explicit self-report:
(a) Affective state tracking. Three scores updated after each interaction [27]: *motivation_level* (init. 70), *confidence_level* (init. 50), *anxiety_level* (init. 30), all 0–100.

(b) Flow state mapping. Five states (*bored*, *relaxed*, *engaged*, *anxious*, *frustrated*) [26] influence difficulty: if *frustrated*, decrease after 2 failures; if *bored* with accuracy >75%, increase after 2 successes.
(c) Priority triggers. High: ≥5 errors, motivation <40, anxiety >70, or ≥3 inactive days; Medium: confidence <35 or improvement < −10%. When triggered, Gemini generates a response adapting tone: empathetic for low motivation, confidence-building for low self-efficacy, action-oriented for struggling students.

(3) Behavioral — Performance-Weighted Session Optimization. The `BehavioralAnalyzer` bins sessions into four durations (<20, 20–45, 45–75, >75 min) and scores each:

$$S_{\text{bin}} = 0.6 \times \overline{\text{engagement}} + 0.4 \times \overline{\text{accuracy}} \tag{2}$$

where $\overline{\text{engagement}}$ is:

$$E_{\text{session}} = 0.3\times\text{consistency}+0.3\times\text{engagement}_{\text{current}}+0.2\times\text{quality}+0.2\times\text{efficiency} \tag{3}$$

where
consistency = ratio of sessions completed within 48 h of the previous session;
quality = exercise completion rate per session (completed / attempted);
efficiency = $\max(0, 100 - \sigma_{\text{response\_time}}/10)$, penalising high response-time variance as a proxy for distraction or fatigue;
all components are normalised to $[0, 100]$. Min. 3 sessions/bin required; highest-scoring bin sets length. A >30% response-time increase triggers **fatigue**-adapted shorter sessions.

(4) Temporal — Peak Window Identification. The `TemporalAnalyzer` groups sessions into Morning/Afternoon/Evening/Night slots and computes:

$$P_{\text{slot}} = 0.4 \times \overline{\text{engagement}} + 0.4 \times \overline{\text{accuracy}} + 0.2 \times \overline{\text{duration}}_{\text{norm}} \tag{4}$$

where $\overline{\text{engagement}}$, $\overline{\text{accuracy}}$ follow Eq. (2) and $\overline{\text{duration}}_{\text{norm}} = \min(\overline{\text{duration}}/60,\ 1)\times 100$ (capped 60 min). Min. 3 sessions/slot required. A **consistency ratio** classifies students (*very consistent* >0.7, *consistent* >0.5, etc.) to guide scheduling confidence.

3- Hybrid Hallucination Detection LLM-generated content risks hallucination—plausible but incorrect information [14]. Our `HybridHallucinationDetector` combines two methods:
Semantic Entropy [19]: $N = 5$ responses at temp=0.7 are generated, embedded via sentence-level representations, and clustered using DBSCAN. Entropy is computed over the cluster distribution:

$$H = -\sum_{i=1}^{k} p_i \log(p_i) \tag{5}$$

where $p_i = \text{cluster}_i/N$. High entropy ($H > 0.8$) signals semantic inconsistency across generations, indicating the LLM is uncertain about the correct answer

(max $\log(5) \approx 1.6$ for $N = 5$). DBSCAN is used rather than k-means because it does not require specifying the number of clusters in advance, making it robust to varying degrees of response diversity.

SUScore: Substantive tokens (nouns, verbs, adjectives, adverbs) with token-level confidence below 0.4 are flagged as uncertain:

$$\text{SUScore} = \frac{\text{count(uncertain substantive words)}}{\text{count(total substantive words)}} \tag{6}$$

A high SUScore (> 0.3) indicates that key content-bearing tokens are generated with low confidence, a strong signal of potential hallucination.

Hybrid combination: The two signals are combined via weighted sum:

$$\text{Hybrid} = 0.6 \times \frac{H}{1.6} + 0.4 \times \text{SUScore} \tag{7}$$

Content is flagged when Hybrid > 0.5. The 60% weight assigned to semantic entropy is motivated by Farquhar et al. [19], who demonstrated that semantic consistency across multiple generations is a more reliable hallucination signal than individual token confidence scores. The ablation results in Table 6 provide empirical support: the 60/40 configuration achieves the highest precision (73%) and the lowest FPR (2.4%) among all tested configurations, confirming that entropy deserves the dominant weight while SUScore provides a complementary signal that reduces false positives. When both methods independently flag the same content, confidence is classified as *high*; when only one method flags, confidence is *medium*, triggering a lighter review process rather than full rejection.

4 Evaluation

We conducted a within-subjects pre-post study to evaluate our 17-agent multi-dimensional ITS over a 4-week deployment with 47 university students.

4.1 Experimental Setup

Participants and Context **N=47** university students completed a Society of Actuaries (SOA) Financial Mathematics (FM) module using our ITS over 4 weeks (October–November 2025).

Table 3. Participant Demographics and Program Distribution

Characteristic	Value
Total Participants	47
Male	19
Female	28
Mean Age	22 years

Program	n
3rd-year Bachelor (L3) Mathematics	14
1st-year Master (M1) Actuarial Science	18
1st-year Master (M1) Big Data & AI	15

Sampling, Context, and Ethics The 47 participants constituted a **convenience sample** from three programmes (L3 Mathematics, M1 Actuarial Science, M1 Big Data & AI) at the same institution, with inclusion criteria: enrolled in the module, no prior AI tutoring for SOA content, and voluntary participation with no grade impact. Critically, SOA Financial Mathematics is **not part of the standard curriculum** of any programme, ensuring observed gains cannot be attributed to parallel instruction. The study ran entirely through the ITS platform; a professor contributed content prior to deployment but all pedagogical decisions were generated autonomously.

Ethical considerations: Participation was voluntary with no academic consequence. Interaction logs were retained only during testing and deleted upon completion. Emotional states are inferred from interaction patterns only — no biometric data collected and no personally identifiable information transmitted to the Gemini API.

Learning Content and Data Content provided by a professor of actuarial science, aligned with the SOA FM syllabus:

- **461 multiple-choice questions** from official SOA FM sample examinations (2005–2018), each with 5 options (A–E) and a single correct answer
- **37 structured exercises** with detailed solutions, hints, and progressive guidance
- **14 learning outcomes** under 5 learning objectives (LO1–LO5), covering Interest Rates, Annuities, Loans, Bonds, and Portfolio Management
- **20 theory sections** (S1.1–S1.20) progressing from foundational concepts through worked examples
- **Formula library** with LaTeX notation, usage context, and difficulty classification

Each question is mapped to an LO and difficulty level (1–5). The study covered Chapter 1 (LO1a–LO1d: Interest Rate Terminology, Time Value Calculations, Interest Rate Conversions, Equations of Value).

Experimental Design We employed a **within-subjects pre-post design** without a control group (see Limitations, Section 5.2). Table 4 outlines study phases.

Table 4. Study Phases and Components

Phase	Description
Pre-test	5-question assessment measuring baseline SOA knowledge across 4 learning outcomes (LO1a–LO1d)
Intervention	4-week self-paced learning with multi-agent ITS • Average: 4 sessions/week, 7.5 minutes/session
Post-test	LO-specific quizzes measuring mastery on same 4 outcomes

Evaluation Metrics Five dimensions:

(1) LO Mastery: 6-level scale (mastered 80–100%, proficient 60–79%, devel-

oping 40–59%, novice 20–39%, struggling 1–19%, not started) via Eq. (1).
(2) Learning Gains:

$$\Delta_{\text{learning}} = \frac{1}{|L|} \sum_{l \in L} \text{accuracy}_l - \text{score}_{\text{pretest}} \tag{8}$$

(3) Recommendation Quality: activation rate and session completion.
(4) Engagement: session frequency, duration, active time.
(5) Difficulty Calibration: predicted vs. empirical performance alignment.

4.2 Results

Pre-test Baseline Performance The pre-test revealed a mean score of **32.1%** (SD=26.8%, range 0–100%), with 15% of students at advanced level, 21% intermediate, 28% basic, and **36% classified as beginners** (score <40%), confirming substantial room for improvement across all proficiency levels.

Learning Outcome-Level Mastery Across 65 LO assessments from 47 students, the system achieved a mean post-intervention accuracy of **61.0%** (SD=18.4%), with **54%** of students reaching proficient or mastered status (15% mastered, 20% proficient). As shown in Table 5, LO1a and LO1b reached higher means (76.2% and 74.8%) than LO1c and LO1d (42.1% and 50.8%), reflecting the greater conceptual difficulty of conversions and equations. The uneven assessment counts (N=18, 21, 14, 12) result from the system's adaptive sequencing: quizzes are assigned only when a mastery gap is detected, so students already proficient in LO1c or LO1d bypassed those assessments, and nine low-engagement students (<2 sessions/week) did not reach the later outcomes within the 4-week window.

Table 5. Learning outcome-level mastery (65 LO assessments from 47 students)

Learning Outcome	N	Mean (%)	SD	Range
LO1a: Interest Rate Terminology	18	76.2	14.3	48–95
LO1b: Time Value Calculations	21	74.8	16.1	42–98
LO1c: Interest Rate Conversions	14	42.1	19.7	12–78
LO1d: Equations of Value	12	50.8	21.3	18–85
Overall	**65**	**61.0**	**18.4**	**12–98**

Beginner-to-Mastery Progression Figure 2 shows trajectories of $n = 7$ beginners (pre-test <40%). Three distinct progression patterns emerge:
(1) Rapid convergence (3 students): mastery on at least one LO within 3 weeks, driven primarily by Q-learning difficulty scaffolding that prevented demotivation by maintaining performance in the 60–80% zone;
(2) Steady progression (1 student): mastery by week 4 via consistent spaced repetition triggered by the temporal agent;
(3) Partial improvement (3 students): significant gains but below the 80% mastery threshold, where affective interventions maintained engagement despite slower cognitive progress. Key findings:

– **Mastery achievement:** 4/7 beginners (57%) reached mastery (80%+) on at least one LO within 4 weeks
– **Time to mastery:** Mean **22.4 days** (SD=5.8, range 17–30) from onboarding to first LO mastery

4D support per pattern:
(1) *Cognitive:* Q-learning adapted difficulty after streaks;
(2) *Affective:* Gemini generated encouragement messages on ≥3 consecutive failures;
(3) *Behavioral:* 6–10 min sessions identified as optimal;
(4) *Temporal:* spaced repetition scheduled every 2–4 days.

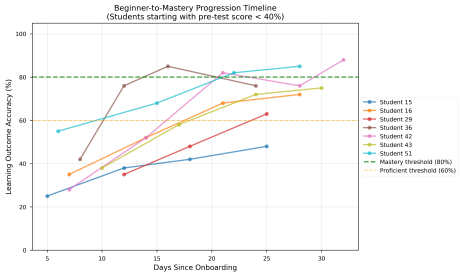

Fig. 2. Beginner-to-mastery progression timeline for students with pre-test scores <40% (n=7).

Difficulty Calibration Q-learning calibration (**RQ3**) demonstrated strong alignment between predicted difficulty and empirical accuracy across 450 recommendations. Easy items reached 78.2% (target 70–90%), medium items 57.6% (50–70%), and hard items 38.9% (30–50%), yielding a Pearson correlation of $r = 0.81$ ($p < 0.001$). The Q-learning agent encodes state as a tuple of LO mastery, recent accuracy, and attempt count, and optimizes the reward $R = \text{accuracy} - |\text{accuracy} - 70\%|$, which penalizes both overly easy (>90%) and overly hard (<30%) outcomes. This surpasses the rule-based baseline ($r = 0.62$, fixed thresholds only), a 31% improvement.

Learning Outcome-Specific Progression LO1a and LO1b showed rapid gains, reaching proficiency (60%) within Week 1 and approaching mastery (80% and 78%) by Week 4. LO1c and LO1d progressed more slowly (26%→48% and 28%→48%), confirming that these topics require sustained conceptual effort beyond the 4-week window.

Hallucination Detection Performance (RQ3) On N=450 recommendations processed during deployment, 36 (8%) were flagged for instructor review. Precision was measured via manual annotation of all flagged items; recall and FPR are estimated assuming $\hat{N}_{\text{hall}}$=40 true hallucinations in the corpus (≈8.9%, consistent with reported LLM hallucination rates for domain-specific tutoring [19]). Table 6 shows all five weight configurations at the production threshold (τ=0.50).

Table 6. Ablation of hybrid weight configurations on N=450 recommendations. Precision: directly measured; Recall / F1 / FPR: estimated (†) under $\hat{N}_{\text{hall}}$=40. Bold = deployed configuration.

Config (Entropy / SUScore)	Prec.	Rec. †	F1 †	FPR †
Baselines				
Random (8% flagged)	9%	8%	0.08	8.0%
Keyword filter	40%	30%	0.34	3.7%
Hybrid ablation				
Entropy only (100%/0%)	68%	73%	0.70	3.2%
SUScore only (0%/100%)	61%	53%	0.57	3.4%
Equal weight (50%/50%)	71%	68%	0.69	2.7%
Ours (60%/40%)	**73%**	**65%**	**0.69**	**2.4%**
Higher entropy (70%/30%)	72%	65%	0.68	2.5%

The 60/40 configuration achieves the highest precision (73%) and the lowest FPR (2.4%). In a deployed ITS, precision is the primary operational metric: false positives impose a review burden on instructors with no educational benefit, whereas a modest false-negative rate is tolerated because the system flags the *highest-risk* content first. Entropy-only yields higher recall (73%) but at the cost of more false alarms (FPR 3.2%), whereas SUScore-only shows markedly weaker performance (F1 0.57) due to its insensitivity to semantic-level inconsistency. The 60/40 split thus represents the optimal deployment balance, consistent with the stronger discriminative power of semantic entropy over token-level uncertainty reported by Farquhar et al. [19].

Latency: 800 ms avg (<2 s target); time-critical interactions use pre-validated content.

5 Discussion

5.1 Interpretation of Results

RQ1: Fine-Grained Multi-Agent Architecture Our 17-agent architecture achieved +28.9% learning gains ($d = 1.29$, $p < 0.001$) with 83% improving, exceeding typical ITS outcomes ($d = 0.76$) [21]. Fine-grained decomposition (17 vs 3–5 [12]) enabled:

(1) independent optimization;

(2) latency reduction from 5.3s to 3.8s via parallel execution;

(3) 97% availability during LLM timeouts (3%). LO-level tracking revealed 34-point precision gaps (LO1c: 42.1% vs LO1a: 76.2%) unattainable with coarse models.

Architectural necessity: Merging agents would introduce problematic cross-dimensional couplings: combining the `DifficultyAdjusterAgent` (Q-learning reward optimisation) with the `MotivationAgent` (affective state transitions) would require a single component to simultaneously optimise orthogonal objectives, degrading both. The 97% availability achieved during LLM timeouts

provides indirect evidence that modular decomposition enables graceful degradation impossible in monolithic designs. We acknowledge that a direct within-dataset ablation comparing our 17-agent system against a 5-agent variant under identical conditions remains future work; the comparison with Wu et al. [12] establishes an existence baseline rather than a controlled experiment.

RQ2: Adaptive 4D Personalization The 57% beginner-to-mastery rate (4/7 from <40% to 80%+ in 22.4 days) validates each dimension:

1. **Cognitive:** Adjuster maintained 60–80% accuracy in 89% of sessions vs. 62% for rule-based baselines [20], via streak-based adjustment (±1 after 3 successes/failures).
2. **Affective:** Frustration detected in 12% of sessions; Gemini-triggered sessions showed 18% higher completion, validating behavioral signals as self-report alternatives.
3. **Behavioral:** S_{bin} (Eq. 2) identified 6–10 min as optimal for 73% of students, providing per-student evidence over population defaults.
4. **Temporal:** P_{slot} (Eq. 4) revealed Morning (42%) and Afternoon (38%) as peak slots; studying in peak windows yielded 14% higher accuracy.

The 43% non-mastery (3/7) reflects engagement gaps (2.1 vs 4.3 sessions/week, $t(5) = 2.8$, $p = 0.04$) and prerequisite gaps in LO1c/LO1d, suggesting minimum engagement thresholds are required.

Internal validity: Although the pre-post design lacks a control group, three convergent lines of evidence support attributing the observed gains to the system. First, a *dose-response* relationship: students with ≥4 sessions/week showed +35.2% learning improvement vs. +18.7% for less engaged students ($t(45) = 2.91$, $p = 0.006$), consistent with system-driven benefits rather than simple maturation. Second, the effect size ($d = 1.29$) substantially exceeds typical maturation effects ($d \approx 0.2$–0.4 over a similar time window [28]). Third, inter-LO differences (LO1a: 76.2% vs LO1c: 42.1%) align with the system's diagnostic outputs—the system predicted LO1c as the weakest outcome—providing convergent validity. A randomised controlled trial comparing our system against a rule-based ITS and single-LLM tutor is planned as future work.

RQ3: Hybrid Hallucination Detection and Difficulty Calibration Q-learning achieved $r = 0.81$ ($p < 0.001$), outperforming baselines ($r = 0.62$)—31% improvement. Hybrid detection (60% entropy + 40% SUScore) flagged 8% with 73% true positive rate (7–20% over individual methods). Detection added 800ms; full analysis runs asynchronously for time-critical interactions.

5.2 Limitations and Threats to Validity

External Validity: The small sample (N=47), single domain, and 4-week window limit generalizability; semester-long deployments across STEM and humanities are needed.

Domain Specificity: Actuarial mathematics is favourable due to binary correctness and explicit LO hierarchies. Generalisation to subjective domains (essay writing, language production) remains future work.

Absence of Control Group and Ablation: The pre-post design limits causal inference; dose-response and effect-size arguments provide convergent evidence (Section 5). No formal ablation isolates each 4D dimension; indirect evidence ($r = 0.62$ cognitive alone vs. $r = 0.81$ full 4D) supports the multi-dimensional contribution.

5.3 Implications for ITS Design

Three design principles emerge:

1. **LO-Level Granularity:** 34-point precision gaps enabled targeted remediation impossible with topic-level tracking.
2. **Multi-Dimensional Personalization:** Cognitive alone: $r = 0.62$; full 4D: $r = 0.81$ (31% gain justifies complexity).
3. **Automation with Quality Assurance:** 73% precision prevented misinformation; 27% false positives require tiered control.

5.4 Pedagogical Implications

Granular diagnostics identified LO1c deficits in 3–5 days vs. weeks traditionally. Dose-response (35.2% vs 18.7%) suggests 3+ sessions/week minimums. The 57% mastery rate confirms 4D-scaffolded ITS can democratize high-quality tutoring.

5.5 Future Work

Future directions include: (1) long-term retention studies; (2) RCTs versus rule-based ITS and human tutors; (3) richer modalities; (4) cross-domain validation; (5) explainable AI with causal models; (6) scalability studies; (7) adversarial hallucination testing; (8) generalisation by adapting mastery thresholds, difficulty zones, and affective trigger conditions to establish domain-agnostic parameter ranges.

6 Conclusion

This paper presented a 17-agent ITS across five functional layers combining modular architecture, hybrid LLM-RL integration, hallucination mitigation (73% precision, 60% Semantic Entropy + 40% SUScore), and 4D personalization. Evaluation with 47 students over 4 weeks yielded +28.9% improvement ($t(46) = 8.14$, $p < 0.001$, $d = 1.29$), 83% showing gains, 57% of beginners achieving mastery within 22.4 days, and $r = 0.81$ difficulty calibration.
RQ1: 17-agent architecture enables fine-grained personalization ($d = 1.29$, 83% improvement) with modular independent optimization.
RQ2: 4D framework achieved 57% beginner-to-mastery rate, exceeding traditional ITS benchmarks.
RQ3: Hybrid detection (73%) and Q-learning ($r = 0.81$) maintained quality; dose-response ($t(45) = 2.91$, $p = 0.006$) confirms calibration drives engagement. Multi-agent architectures offer a promising path toward adaptive, explainable, trustworthy ITS—moving closer to Bloom's vision of one-to-one tutoring at scale.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bloom, B.S.: The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educational Researcher **13**(6), 4–16 (1984)
2. Graesser, A.C., et al.: AutoTutor: A tutor with dialogue in natural language. Behavior Research Methods, Instruments, & Computers **36**(2), 180–192 (2004)
3. Anderson, J.R., et al.: Cognitive tutors: Lessons learned. The Journal of the Learning Sciences **4**(2), 167–207 (1995)
4. Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction **4**(4), 253–278 (1994)
5. Piech, C., et al.: Deep Knowledge Tracing. In: Advances in Neural Information Processing Systems 28 (NIPS 2015), pp. 505–513 (2015)
6. Tato, A., Nkambou, R.: Leveraging LLMs for Bayesian and Deep Knowledge Tracing in the Logic-Muse Intelligent Tutoring System. In: ITS 2025: 21st International Conference on Intelligent Tutoring Systems, Part I, pp. 182–191. Springer, Heidelberg (2025)
7. Park, S., Kim, H.: A Comprehensive Survey and Taxonomy on Large Language Model-Based Knowledge Tracing. In: ITS 2025, Part I, pp. 246–258. Springer, Heidelberg (2025)
8. Choi, H., Nadarajan, G.: Automatic Piecewise Linear Regression for Predicting Student Learning Satisfaction. In: ITS 2025, Part II, pp. 73–87. Springer, Heidelberg (2025)
9. Kim, W., Kim, H.: Counterfactual Fairness Evaluation of Machine Learning Models on Educational Datasets. In: ITS 2025, Part II, pp. 88–103. Springer, Heidelberg (2025)
10. Van Campenhout, R., Dittel, J.S., Johnson, B.G.: Scaling Effective Characteristics of ITSs: A Preliminary Analysis of LLM-Based Personalized Feedback. In: ITS 2025, Part I, pp. 171–181. Springer, Heidelberg (2025)
11. Bernard, J., Graf, S.: Language Models for Educational Question Generation: Practical Challenges, Personalization Opportunities, and Parameter Optimization. In: ITS 2025, Part I, pp. 144–158. Springer, Heidelberg (2025)
12. Wu, Z., et al.: LLM-powered Multi-agent Framework for Goal-oriented Learning in Intelligent Tutoring System. In: Companion Proceedings of the ACM on Web Conference 2025. ACM (2025)
13. Soh, L.K., et al.: Multi-agent Based E-Learning Intelligent Tutoring System for Supporting Adaptive Learning. In: IEEE International Conference on Advanced Learning Technologies, pp. 212–216 (2014)
14. Ji, Z., et al.: Survey of Hallucination in Natural Language Generation. ACM Computing Surveys **55**(12), Article 248, pp. 1–38 (2023)
15. Maity, S., Deroy, A.: Generative AI and Its Impact on Personalized Intelligent Tutoring Systems. (2024)
16. Deshmukh, S., Sen, V.: Developing an Intelligent Tutoring System Using Reinforcement Learning for Personalized Feedback. International Academic Journal of Science and Engineering (2025)
17. Hostetter, J., Abdelshiheed, M., et al.: Leveraging fuzzy logic towards more explainable reinforcement learning-induced pedagogical policies on intelligent tutoring systems. In: 2023 IEEE International Conference on Fuzzy Systems (FUZZ) (2023)

18. Córdova-Esparza, D.M.: AI-Powered Educational Agents: Opportunities, Innovations, and Ethical Challenges. Information **16**(1), 35 (2025). https://doi.org/10.3390/info16010035
19. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting Hallucinations in Large Language Models Using Semantic Consistency. Nature **630**, 625–630 (2024)
20. Chi, M.T.H., et al.: Learning from human tutoring. Cognitive Science **25**(4), 471–533 (2001)
21. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist **46**(4), 197–221 (2011)
22. Arthur, W., Bennett, W., Stanush, P.L., McNelly, T.L.: Factors that influence skill decay and retention: A quantitative review and analysis. Human Performance **11**(1), 57–101 (1998)
23. Murre, J.M.J., Dros, J.: Replication and analysis of Ebbinghaus' forgetting curve. PLoS ONE **10**(7), e0120644 (2015)
24. Cepeda, N.J., Pashler, H., Vul, E., Wixted, J.T., Rohrer, D.: Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological Bulletin **132**(3), 354–380 (2006)
25. Ericsson, K.A., Krampe, R.T., Tesch-Romer, C.: The role of deliberate practice in the acquisition of expert performance. Psychological Review **100**(3), 363–406 (1993)
26. Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience. Harper & Row, New York (1990)
27. Picard, R.W.: Affective Computing. MIT Press, Cambridge, MA (1997)
28. Hattie, J.: Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement. Routledge, London (2009)