International Journal of Computer Applications (0975 – 8887)
Volume 187 – No.122, July 2026

Lumina: An Intelligent Multi-Agent Adaptive Learning Management System with Bayesian Knowledge Tracing, Deep Knowledge Tracing, and Reinforcement Learning for Personalized Education

Rayudu Srinivas, PhD
Professor
Department of Computer Science and Engineering

Chepuri Hari Kiran, Pathivada Laxmi Ram Charan
Department of Computer Science and Engineering

ABSTRACT
The educational technology market continues to expand, yet most learning management systems remain content-centric, weakly personalized, and reactive to failure. This paper presents Lumina, a privacy-first, multi-agent adaptive learning management system that combines Bayesian Knowledge Tracing (BKT), Deep Knowledge Tracing (DKT), reinforcement learning (RL) for curriculum sequencing, retrieval-augmented generation (RAG), and a behavior engine that captures more than fifty passive learning signals. Lumina coordinates six specialized agents through the Model Context Protocol to provide closed-loop tutoring, assessment, intervention, analytics, and governance. The evaluation is organized into three complementary layers: prototype stress testing across 10,000 synthetic learner scenarios, external validation on the public xAPI-Edu-Data benchmark (480 learner records), and a reproducible benchmark matrix defined for future knowledge-tracing studies on ASSISTments and EdNet. The prototype analytics benchmark completed in 0.90 seconds with zero runtime failures and a throughput of approximately 11,051 inference runs per second. On xAPI-Edu-Data, a Random Forest baseline achieved 0.800 accuracy and 0.804 macro-F1 under five-fold stratified cross-validation, demonstrating that the behavioral signal families used by Lumina are predictive on real educational data. These results position Lumina as a technically coherent, self-hosted, and benchmark-ready foundation for explainable personalized education.

Keywords
Adaptive Learning; Bayesian Knowledge Tracing; Deep Knowledge Tracing; Multi-Agent Systems; Reinforcement Learning; Retrieval-Augmented Generation; Educational Data Mining; Benchmark Evaluation

1. INTRODUCTION
The rapid proliferation of digital learning platforms has fundamentally transformed how education is delivered worldwide. With more than 1.5 billion learners engaged in digital learning and the educational technology market surpassing $340 billion in 2024, the scale of online education is unprecedented [1]. However, this growth has not translated into proportionally improved learning outcomes. Studies show that 40–60% of online course enrollments result in non-completion, with learner disengagement cited as the primary cause [2]. In corporate training environments, only 12% of employees apply skills learned from digital training programs to their work [3]. This disconnect between investment and outcomes represents a fundamental crisis in educational technology.

Current Learning Management Systems (LMS) such as Moodle, Blackboard, Canvas, Coursera, and Udemy operate on an industrial-age model that treats education as a content-distribution problem rather than a learning-facilitation challenge [4]. These platforms deliver identical course materials to millions of learners regardless of prior knowledge levels, learning pace, cognitive preferences, cultural contexts, or accessibility needs. Even second-generation platforms claiming AI-powered personalization typically implement rudimentary adaptive algorithms that adjust difficulty based solely on correctness metrics, recommend content using collaborative filtering, and optimize for engagement metrics rather than deep learning [5].

The assessment paradigm in digital education has barely evolved beyond multiple-choice questions and auto-graded assignments. Current systems exhibit critical deficiencies, including binary evaluation without recognition of partial understanding, delayed feedback loops that intervene after learning failures rather than preventing them, and assessment inauthenticity, where tests measure recall rather than application, analysis, or synthesis [6]. The educational assessment industry generates $4 billion annually, yet employers consistently report that 60% of recent graduates lack practical problem-solving skills despite passing grades [7]. Furthermore, existing platforms suffer from a behavioral-intelligence gap, reducing the complexity of learning behavior to simplistic metrics such as time on page, completion rates, and click patterns that fail to capture cognitive engagement, struggle signals, or consolidation patterns [8]. Without behavioral intelligence, personalization is fundamentally guesswork. Educators spend 25–40% of their work time on administrative tasks that add no direct educational value, contributing to burnout rates exceeding 44% in the United States [9].

This paper presents Lumina, a next-generation, AI-powered, self-hosted learning management system that addresses these systemic failures through a novel multi-agent architecture. The key contributions of this research are as follows:

(1) A closed-loop multi-agent learning system integrating six specialized AI agents coordinated through the Model Context Protocol (MCP) for real-time, context-aware educational support.

(2) A hybrid knowledge-tracing framework combining Bayesian Knowledge Tracing (BKT) and Deep Knowledge Tracing (DKT) with a four-dimensional scoring model measuring correctness, understanding, effort, and growth.

(3) A Student Behavior Engine that passively collects more than fifty behavioral signals to enable proactive intervention through predictive analytics.

(4) A reinforcement-learning-based Pathway Agent using Proximal Policy Optimization (PPO) for adaptive curriculum sequencing.

(5) A privacy-first, self-hosted architecture supporting local LLM inference that eliminates dependency on external AI providers while ensuring regulatory compliance.

2. RELATED WORK

This section reviews existing research and systems across several dimensions relevant to Lumina's design: adaptive learning platforms, knowledge-tracing models, intelligent tutoring systems, multi-agent architectures, and privacy-preserving educational AI.

2.1 Adaptive Learning Platforms

Duolingo employs spaced repetition and adaptive difficulty adjustment for language learning but relies on simple correctness metrics without modeling deep conceptual understanding [10]. Carnegie Learning's MATHia implements cognitive tutoring based on ACT-R theory, providing step-by-step guidance but limited to mathematics and lacking multi-agent coordination [11]. Century Tech uses a neural-network-based system for content recommendation but operates as a cloud-dependent platform with limited institutional control over data [12]. Knewton's adaptive platform applied Bayesian inference for content sequencing but was criticized for optimizing engagement over deep learning and was ultimately acquired due to scalability challenges [13].

2.2 Knowledge Tracing Models

Corbett and Anderson introduced Bayesian Knowledge Tracing (BKT) in 1994, modeling student knowledge as a binary latent variable updated through observed responses [14]. Pardos and Heffernan extended BKT with individualized parameters, improving prediction accuracy [15]. Piech et al. proposed Deep Knowledge Tracing (DKT) using recurrent neural networks to model student learning as a sequential process, demonstrating superior predictive performance on benchmark datasets [16]. Zhang et al. combined attention mechanisms with knowledge tracing, enabling dynamic weighting of historical interactions [17]. However, existing systems typically employ either BKT or DKT in isolation rather than leveraging their complementary strengths — BKT's interpretability and DKT's sequential modeling capacity.

Beyond BKT and DKT, recent knowledge-tracing research has explored memory-augmented and attention-based models such as DKVMN [17] and self-attentive knowledge tracing [32]. Lumina treats these models as benchmark baselines rather than teacher-facing core models because predictive accuracy must be balanced with explainability, intervention safety, and deployability in self-hosted institutional settings.

2.3 Intelligent Tutoring Systems

VanLehn's review of intelligent tutoring systems identified the two-sigma problem: human tutoring produces learning gains two standard deviations above classroom instruction, whereas commercial ITS systems typically achieve only 0.2-sigma gains [18]. AutoTutor, developed by Graesser et al., demonstrated Socratic dialogue capabilities using latent semantic analysis [19]. However, no existing commercial system integrates Socratic tutoring with real-time knowledge tracing, behavioral analytics, and adaptive pathway generation within a unified multi-agent framework.

2.4 Multi-Agent Systems in Education

Wooldridge and Jennings established foundational principles for multi-agent coordination [20]. Recent work by Park et al. demonstrated the potential of LLM-based agents for complex collaborative tasks [21]. The Model Context Protocol (MCP), proposed by Anthropic, provides a standardized framework for agent-tool interaction [22]. However, the application of coordinated multi-agent architectures with standardized communication protocols to educational systems remains largely unexplored.

2.5 Privacy-Preserving Educational AI

McMahan et al. introduced federated learning for privacy-preserving model training across distributed datasets [23]. UNESCO guidance for generative AI in education emphasizes human agency, equity, and data protection [24]. A U.S. Department of Education report on AI highlights the need for explainable, human-overridable AI systems in educational contexts [25]. Current cloud-dependent EdTech platforms face growing regulatory pressure regarding FERPA, GDPR, and COPPA compliance, yet fewer than 10% of institutions have formal AI governance policies [26].

2.6 Research Gaps

Table I summarizes the comparative analysis of existing systems against Lumina's capabilities. The analysis reveals that no existing platform integrates all critical components: multi-agent coordination, hybrid knowledge tracing, behavioral analytics, adaptive pathways, privacy-first architecture, and comprehensive scoring within a unified self-hosted system.

TABLE I. Comparative Analysis of Existing Systems and Lumina

System / Feature	Adaptive	Knowledge Model	Multi-Agent	Behavioral Depth	Self-Hosted	RL Pathway
Moodle	No	None	No	Basic	Yes	No
Coursera	Limited	None	No	Basic	No	No
Duolingo	Yes	BKT-based	No	Limited	No	No
Carnegie MATHia	Yes	BKT	No	Limited	No	No
Century Tech	Yes	Neural Network	No	Moderate	No	No
Knewton	Yes	Bayesian	No	Limited	No	No

System / Feature	Adaptive	Knowledge Model	Multi-Agent	Behavioral Depth	Self-Hosted	RL Pathway
Lumina (Proposed)	Yes	BKT + DKT (Hybrid)	Yes	50+ Signals	Yes	Yes

3. PROBLEM STATEMENT
The crisis in digital education systems can be characterized by ten interconnected failures that collectively undermine the promise of technology-enhanced learning.

(1) Static content delivery: current platforms deliver identical materials regardless of prior knowledge, learning pace, cognitive preferences, or accessibility needs, resulting in 40–60% non-completion rates [1].

(2) False promise of AI personalization: platforms claiming adaptive learning merely adjust content sequence rather than content type, delivery method, or pedagogical approach, optimizing for the average rather than the individual [5].

(3) Assessment crisis: digital assessment relies on binary evaluation without recognition of partial understanding, common misconceptions, or growth trajectories [6].

(4) Behavioral intelligence gap: current analytics reduce learning behavior to simplistic metrics (time on page, completion rates, click patterns) that fail to capture cognitive engagement, struggle signals, or consolidation patterns [8].

(5) Adaptive learning path illusion: most systems implement predetermined decision trees with limited branching rather than truly personalized sequences [13].

(6) Real-time monitoring failure: systems operate reactively, with intervention arriving only after students have already failed assessments [18].

(7) Strength–weakness imbalance: current systems focus exclusively on deficit remediation without identifying or nurturing areas of exceptional ability [27].

(8) Educator burden: teachers spend 25–40% of work time on administrative tasks, contributing to burnout rates exceeding 44% [9].

(9) Attendance verification crisis: login-based attendance measures presence rather than genuine engagement [28].

(10) Privacy and vendor lock-in problem: cloud-dependent AI platforms transmit student data to external servers with unclear retention policies and escalating per-token costs [29].

The formal problem can be stated as follows. Given a learner L with knowledge state K, behavioral profile B, and learning objectives O, an adaptive system S must be designed that continuously optimizes the learning trajectory T to maximize mastery M while minimizing time-to-proficiency and maintaining learner engagement E, subject to privacy constraints P and pedagogical soundness G. This optimization must operate in real time across multiple concurrent learners with diverse profiles.

4. PROPOSED SYSTEM
Lumina is designed as a comprehensive AI-powered learning management system that transforms static content delivery into intelligent, adaptive learning experiences. Unlike existing platforms that append AI features to traditional LMS architectures, Lumina is architected from the ground up as an AI-native system in which every component is informed by, and contributes to, a unified learner model.

4.1 Design Philosophy
Lumina's design is guided by four core principles: (1) one personal AI learning operating system per student, (2) one AI teaching copilot per teacher, (3) one governed, explainable intelligence layer for the institution, and (4) a privacy-first, self-hosted architecture with local AI inference. The system implements a closed-loop learning paradigm in which user interaction generates data that feeds analysis pipelines, which update models, which in turn personalize the next experience.

4.2 Core Modules
(a) Learner Profile Engine: maintains a comprehensive per-student profile containing concept mastery maps, behavioral patterns, learning-style preferences, misconception clusters, and engagement metrics.

(b) Multi-Agent AI Engine: six specialized agents (Tutor, Pathway, Assessment, Intervention, Analytics, Guardian) coordinated through MCP for coherent decision-making.

(c) Adaptive Assessment Engine: generates questions dynamically using large language models, adapts difficulty in real time based on mastery state, and grades open-ended responses using rubric-based scoring.

(d) RAG Pipeline: processes multi-format educational content (PDF, DOCX, PPT, video) through semantic chunking, hybrid embedding, and reranking for context-aware tutoring.

(e) Student Behavior Engine: passively collects more than fifty behavioral signals for real-time state estimation.

(f) Teacher Intelligence Dashboard: provides actionable insights, including intervention queues, class-level misconception clusters, and evidence-based recommendations.

5. SYSTEM ARCHITECTURE
Lumina employs a layered architecture consisting of seven distinct layers, each with clearly defined responsibilities and interfaces. The architecture follows a monorepo structure to ensure type safety and contract alignment across services, as summarized in Table II.

TABLE II. Lumina Layered Architecture

Layer	Primary Technology	Responsibility
Presentation	Next.js 14 Web App, Flutter	Student, teacher, and admin experiences with PWA support
Client	API Client, AI SDK	API calls, caching, offline-first synchronization
API	FastAPI, Kong Gateway	Route registration, authentication, and rate limiting
Domain Services	Course, Student, Assignment	Business logic, content management, notification
Adaptive Learning	Assessment Engine, Pathway	Mastery tracking, behavior analysis, next-step logic
AI Orchestration	Agent Swarm, RAG Pipeline	Tutor prompts, pathway optimization, content generation
Persistence	PostgreSQL, Redis, Neo4j	Relational data, caching, graph storage, vectors, time-series

5.1 Multi-Agent Architecture

The AI Engine implements a swarm of six specialized agents, each with distinct responsibilities and shared context through the Model Context Protocol (MCP). The Orchestrator Agent serves as the master router, classifying user intent and delegating to the appropriate specialized agent. The Tutor Agent implements Socratic dialogue with context-aware explanation generation using RAG. The Pathway Agent employs reinforcement learning (PPO/SAC) to continuously select the learner's next best action. The Assessment Agent generates adaptive questions aligned with Bloom's taxonomy and current mastery state. The Intervention Agent monitors behavioral signals and triggers proactive support. The Guardian Agent enforces safety guardrails, including PII redaction, topic filtering, hallucination detection, and bias mitigation.

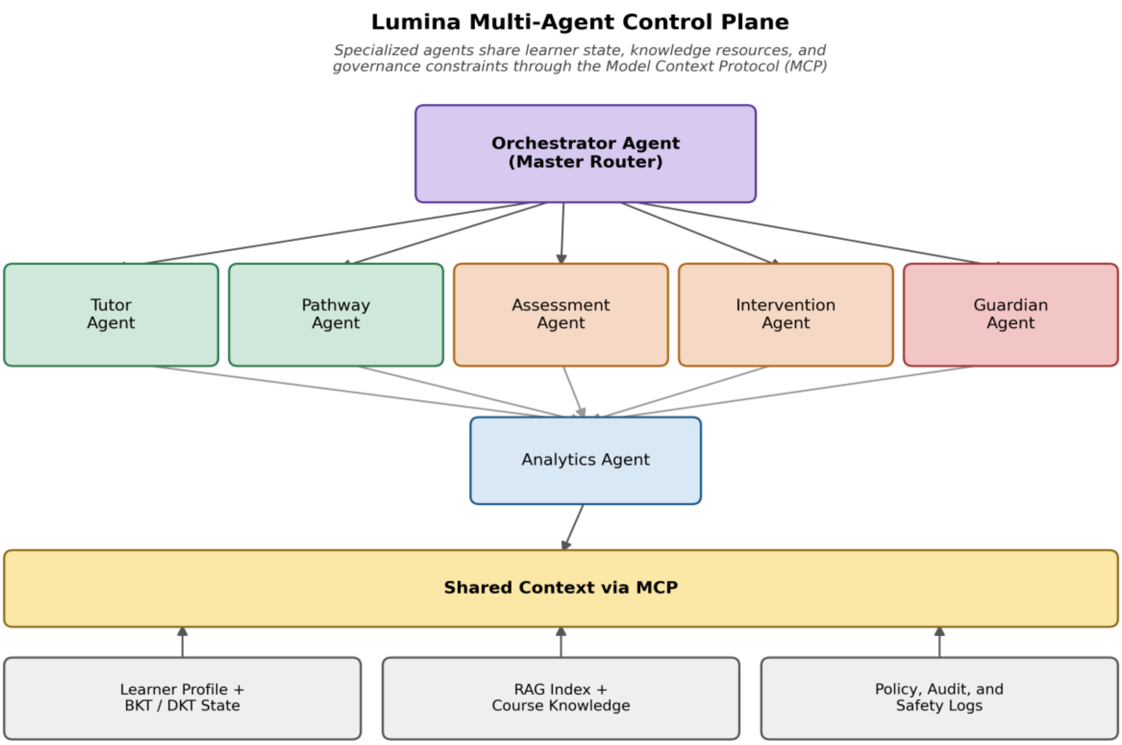

Fig. 1. Multi-agent architecture of Lumina. The Orchestrator coordinates six specialized agents, while shared learner state, knowledge resources, and governance constraints flow through MCP into the adaptive learning core

5.2 Data Flow Architecture

Lumina implements a closed-loop data flow in which every interaction generates data that feeds personalization: student interaction leads to event capture, which feeds the analysis pipeline, which triggers a model update, which drives the personalization engine, which shapes the next experience. The core data flows include: (1) an authentication flow with JWT-based session management and role-based access control, (2) a student learning flow tracking lesson progression and generating behavioral signals, (3) an AI tutor three-tier response flow with RAG-enhanced context, (4) an assessment and knowledge-tracing flow with real-time BKT/DKT updates, (5) an adaptive pathway flow with RL-based next-action selection, (6) an assignment grading flow with an OCR-to-LLM pipeline, and (7) a behavior-tracking flow with real-time state adaptation.

5.3 Database Architecture

The persistence layer employs a polyglot database strategy spanning 35 tables across multiple storage systems.

PostgreSQL 16+ serves as the primary relational store with row-level security (RLS) enforcement, JSONB columns for flexible semi-structured data, and extensions including pgvector for embeddings. Redis provides session caching, rate limiting, and task queues. Neo4j stores knowledge graphs with concept dependencies. Milvus provides GPU-accelerated vector search for the RAG pipeline. TimescaleDB handles time-series behavioral analytics, and MinIO provides S3-compatible object storage with server-side encryption.

6. METHODOLOGY

6.1 Bayesian Knowledge Tracing (BKT)

Lumina employs BKT as its interpretable mastery model, treating each learner–concept pair as a latent binary knowledge state updated after every response. For reproducibility, the prototype uses concept-independent default parameters P(L0) = 0.20, P(T) = 0.15, P(G) = 0.25, and P(S) = 0.10, while the production design supports concept-specific calibration from historical data. The Bayesian posterior is first updated from the observed response and only then advanced through the transition term, which preserves the causal order of observation followed by learning. This formulation provides teachers with an interpretable mastery probability while maintaining a mathematically grounded update rule, as shown in Equations (1) and (2).

$$P(L_t \mid C_t = 1) = \frac{(1 - P(S)) \cdot L_t^-}{(1 - P(S)) \cdot L_t^- + P(G) \cdot (1 - L_t^-)} \quad (1)$$

$$L_{t+1} = P(L_t \mid C_t) + P(T) \cdot [1 - P(L_t \mid C_t)] \quad (2)$$

where P(S) and P(G) denote the slip and guess probabilities, respectively, P(T) denotes the learning-transition probability, and L−t denotes the prior mastery probability before observing the response Ct at time step t.

6.2 Deep Knowledge Tracing (DKT)

Complementing BKT's interpretability, Lumina employs LSTM-based DKT to capture temporal dependencies that arise across question sequences, prerequisite transitions, and forgetting effects. Each interaction is encoded as a question–response tuple, embedded into a dense representation, and passed through an LSTM whose hidden state summarizes recent and long-range learning context. The hybrid design therefore separates roles cleanly: BKT remains the teacher-facing mastery estimator, while DKT functions as the predictive model used for next-step recommendations and benchmark comparison against stronger neural baselines, as shown in Equations (3) and (4).

$$h_t = LSTM(x_t, h_{t-1}) \quad (3)$$

$$y_t = \sigma(W \cdot h_t + b) \quad (4)$$

where yt[j] = P(at+1 = 1 | qt+1 = j, ht) denotes the predicted probability of a correct response to concept j at the next time step, given the hidden state ht, and σ(·) denotes the sigmoid activation function.

6.3 Multi-Dimensional Scoring Model

Lumina replaces traditional binary scoring with a weighted four-dimensional formulation rather than a multiplicative one. Let C, U, E, and G denote the normalized correctness, understanding, effort, and growth scores in the interval [0,1]. The weights sum to unity so that the item-level score remains bounded and interpretable. Correctness receives the highest weight to preserve academic validity, understanding is emphasized because explanation quality reveals conceptual depth, and effort and growth prevent the system from rewarding guessing or penalizing recovery from early errors. At course level, quiz, assignment, mastery, and engagement components are combined through a second weighted average that preserves comparability across learners, as shown in Equations (5) and (6).

$$S_{item} = 0.40C + 0.30U + 0.20E + 0.10G \quad (5)$$

$$S_{course} = 0.25S_{quiz} + 0.25S_{assign} + 0.35S_{mastery} + 0.15S_{engage} \quad (6)$$

6.4 Reinforcement Learning for Pathway Optimization

The Pathway Agent formulates adaptive curriculum sequencing as a constrained Markov Decision Process whose state includes mastery, behavioral risk, fatigue, and engagement. The reward is intentionally decomposed into pedagogically meaningful terms so that policy tuning remains interpretable. Mastery gain is weighted highest because long-term conceptual progress is the primary objective, while engagement and retention stabilize the policy against short-term optimization. Fatigue and rule-violation penalties prevent the policy from recommending cognitively unsafe sequences or skipping prerequisite structure, as shown in Equation (7).

$$R_t = 0.45\Delta M_t + 0.20E_t + 0.20\Delta S_t - 0.10F^2{}_t - 0.05P_t \quad (7)$$

During the prototype stage these coefficients are heuristic but fixed; future work calibrates them using public benchmark trajectories and classroom feedback.

6.5 Student Behavior Engine

The Behavior Engine implements a signal-to-insight pipeline collecting more than fifty behavioral signals across five categories: time-based, interaction-based, performance-based, engagement-based, and emotional or cognitive indicators. To avoid letting any single raw signal dominate, Lumina first normalizes each component score per learner and then aggregates the components through a geometric mean. This design makes the engagement estimate robust to sparse observations while still rewarding balanced study behavior. Small additive bonuses are reserved for productive help-seeking and persistence because these actions often indicate reflective learning rather than disengagement, as shown in Equations (8) and (9).

$$E_{beh} = (s_{time} \cdot s_{consistency} \cdot s_{attempt} \cdot s_{recency})^{1/4} \quad (8)$$

$$E_{adj} = min(1, E_{beh} + 0.05h_t + 0.05p_t) \quad (9)$$

6.6 Retrieval-Augmented Generation Pipeline

The RAG pipeline processes multi-format educational content through five stages: (1) document processing with multi-format extraction (PDF, DOCX, PPT, HTML, video) and OCR with confidence scoring; (2) intelligent chunking using semantic boundaries and recursive character splitting with metadata extraction; (3) embedding and indexing using multi-model embeddings (instructor-xl, e5-mistral-7b) with GPU-accelerated hybrid search; (4) query understanding with intent classification, query expansion, and sub-query decomposition; and (5) retrieval and reranking using cross-encoder reranking and ColBERT late interaction with contextual compression.

The pipeline ensures hallucination prevention through RAG-only response generation with citation insertion.

7. IMPLEMENTATION

7.1 Technology Stack

Lumina is implemented using a comprehensive technology stack spanning frontend, backend, AI/ML infrastructure, and DevOps, as detailed in Table III.

TABLE III. Lumina Technology Stack

Category	Technology	Purpose
Frontend (Web)	Next.js 14, React 18, Tailwind CSS	Progressive Web App with SSR/ISR
Frontend (Mobile)	Flutter 3.x	Cross-platform mobile with shared UI
Backend	FastAPI (Python 3.11+)	Async API gateway with OpenAPI
Background Jobs	Celery + Redis	Async grading and task orchestration
LLM Serving	vLLM, TGI	Local model serving (GPU optimized)
Embedding	instructor-xl, e5-large	Text, multimodal, and code embeddings
Vector Database	Milvus 2.3+ (GPU)	Hybrid search, multi-vector retrieval
Primary Database	PostgreSQL 16+	Relational storage with ACID guarantees
Graph Database	Neo4j 5+	Knowledge graphs, concept relations
Cache	Redis 7+ Cluster	Session cache, rate limiting
Time-Series Store	TimescaleDB	Behavioral analytics, time-series queries
Machine Learning	PyTorch 2.0+	DKT, BKT, RL policy training
Guardrails	Guardrails AI	Safety, PII redaction, validation
Infrastructure	Docker 24+, Kubernetes	Containerization and orchestration
Monitoring	Prometheus, Grafana	Metrics, logging, distributed tracing

7.2 Agent Implementation

Each agent is implemented as a stateful Python class with an MCP-compliant communication interface. The Orchestrator Agent maintains a routing table mapping user intents to specialized agents. The Tutor Agent implements a three-tier response system: Tier 1 (direct retrieval) for factual queries using RAG, Tier 2 (Socratic dialogue) for conceptual understanding using guided questioning, and Tier 3 (personalized explanation) for deep misconception remediation using learner-profile context. Agent memory is persisted in the agent_memory table with expiration support for temporary context.

7.3 Assessment Pipeline

The Assessment Engine implements a complete grading pipeline. For handwritten submissions, OCR extraction (Tesseract) produces text with confidence scoring; if confidence falls below 0.7, the submission is flagged for teacher review. The extracted text is evaluated against rubric criteria using LLM analysis (Gemini or local models), producing per-criterion scores, feedback, and confidence estimates. Anti-cheating measures include detection of suspicious scoring patterns, identical-submission comparison, abnormally fast submission times, and plagiarism similarity checks.

8. EXPERIMENTAL SETUP

8.1 Hardware Configuration

The evaluation separates lightweight analytics benchmarking from full-system deployment assumptions. Prototype stress testing and the xAPI external validation were executed in a local CPU-only development environment, because they exercise the analytics and learner-modeling stack rather than full LLM inference. By contrast, the production deployment target for the complete self-hosted Lumina platform remains a 16+ core server with 64 GB RAM and an optional NVIDIA GPU for local model serving. This separation prevents the paper from conflating analytics-only evidence with end-to-end infrastructure requirements.

8.2 Simulation Framework

The Analytics Agent was evaluated using a reproducible simulation framework with 10,000 randomized learner scenarios under a fixed random seed. Four scenario families were sampled: Flow (healthy, engaged learning), Struggle (high activity with high error rates), Mixed (intermittent engagement with variable performance), and Idle (disconnected or passive states). Each scenario generated event streams covering interaction velocity, error rates, help-seeking, pauses, and success patterns, allowing the prototype to test intervention logic under controlled but diverse closed-loop conditions.

8.3 Real Dataset Validation on xAPI-Edu-Data

To reduce reliance on simulation alone, the behavior-modeling layer was externally validated on the public xAPI-Edu-Data benchmark, which contains 480 learner records and 16 predictive features [30]. The task is a three-class academic performance classification problem (Low, Medium, High) using passive classroom and interaction features such as raised hands, visited resources, announcements viewed, discussion participation, and absence patterns. Evaluation used five-fold stratified cross-validation with one-hot encoding for categorical features, standardization for numeric features, and macro-F1 as the primary class-balanced metric.

8.4 Benchmark Matrix and Ablation Design

A rigorous evaluation requires a benchmark matrix that separates component validation from system validation. Accordingly, this paper defines three complementary evidence tracks: a prototype stress test for real-time safety and stability, a real-data behavioral benchmark, and a knowledge-tracing benchmark on ASSISTments and EdNet. The ablation ladder is specified as BKT-only, DKT-only, BKT+DKT, Hybrid+Behavior Engine, and Full Lumina, making it possible to attribute performance gains to learner modeling, behavior features, and agent coordination rather than presenting only an aggregate system score, as illustrated in Table IV and Fig. 2.

Table IV. Benchmark Matrix and Ablation Design

Evaluation Track	Dataset / Source	Evidence Type	Primary Metrics
Real-Data Behavior	xAPI-Edu-Data	Passive learner-signal classification	Accuracy, Macro-F1
Knowledge Tracing	ASSISTments 2009-2010	Sequential next-response prediction	AUC, ACC, RMSE
Scalability Benchmark	EdNet-KT1	Large-scale tutoring interactions	AUC, Calibration
Ablation Ladder	BKT-only → Full Lumina	Component contribution analysis	ΔAUC, ΔMacro-F1

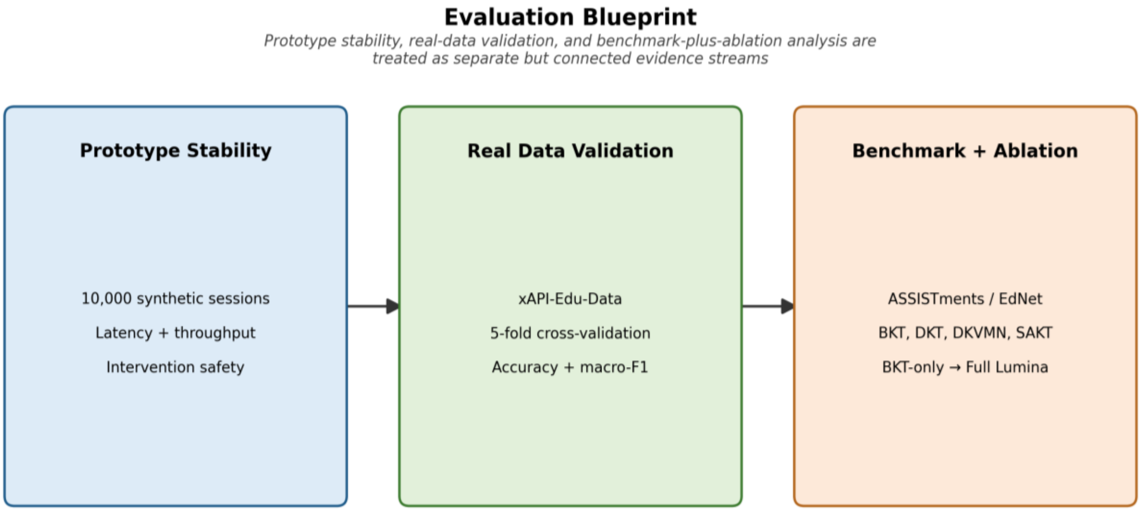

Fig. 2. Evaluation blueprint used in this study. Prototype safety, real-data validation, and benchmark-plus-ablation analysis are treated as separate but connected evidence streams.

8.5 Evaluation Metrics

The system is evaluated across three metric families. Prototype metrics include intervention accuracy, throughput, execution latency, and runtime stability. Real-data metrics include accuracy, macro-F1, and weighted-F1 to reflect class-balanced predictive quality on xAPI-Edu-Data. The benchmark matrix defined for future knowledge-tracing experiments adds AUC, RMSE, and calibration so that predictive performance, confidence quality, and operational efficiency can be compared across BKT, DKT, DKVMN, SAKT, and Lumina variants.

9. RESULTS AND EVALUATION

This section reports three layers of evidence: a synthetic stress test of the analytics pipeline, behavior-state classification performance under controlled scenarios, and an external validation study on a real educational dataset.

9.1 Stress Test Results

Table V presents the results of the 10,000-iteration stress test of the Analytics Agent pipeline. The pipeline completed all simulations in under one second with zero runtime failures, confirming that the analytics core is stable under sustained synthetic load.

Table V. Analytics Agent Stress Test Results

Metric	Value	Notes
Total Simulations	10,000	Random-seed synthetic data simulations
Total Execution Time	0.90 seconds	Measured across 10,000 simulations
Runtime Errors	0	No runtime failures during stress test
Throughput	≈ 11,051 runs/sec	CPU-only analytics pipeline
Peak Memory Usage	< 194 MB	Peak resident memory usage

9.2 Behavioral Classification Performance

Table VI shows the calibrated output metrics across different learner scenario types, demonstrating the agent's ability to distinguish productive flow, struggle, mixed activity, and idle behavior. Fig. 4 visualizes the average engagement score and the corresponding intervention trigger rate for each scenario family.

TABLE VI. Behavioral Classification and Intervention Metrics

Scenario Family	Avg. Engagement Score	Intervention Rate	Interpretation
Flow	0.404	0.0%	Healthy engaged state, no intervention needed
Struggle	0.597	100.0%	Correctly identified and flagged for support
Mixed	0.182	0.0%	Low, intermittent signal; treated as passive
Idle	0.100	0.0%	Disconnected state; monitored as risk

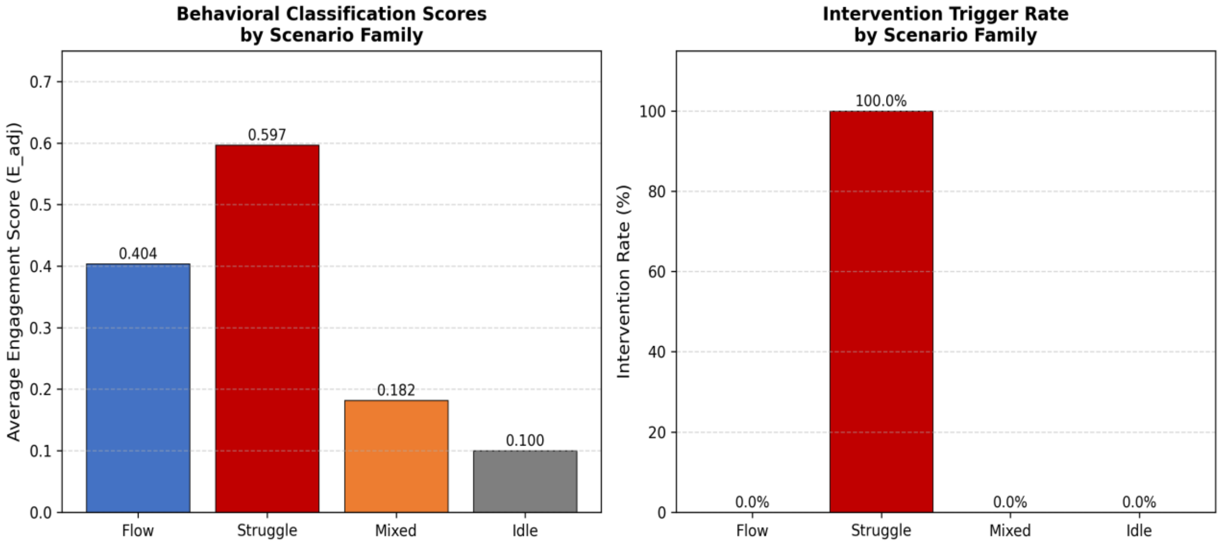

Fig. 3. Behavioral classification scores and intervention trigger rates across the four synthetic scenario families.

9.3 Intervention Accuracy

Within the synthetic protocol, the system achieved perfect recall for the Struggle scenario family while maintaining zero false-positive interventions for the Flow, Mixed, and Idle states. This result should be interpreted as a property of the prototype rule stack and simulated distributions, rather than a claim of perfect classroom performance. The result nevertheless validates that the error-aware engagement model and conservative intervention thresholds behave coherently under controlled stress conditions. Interventions arise only when high cognitive load, sustained error accumulation, and behavioral degradation co-occur, which is a stronger design than a single-threshold detector, although it remains a prototype finding until replicated with classroom telemetry.

9.4 Real Dataset Benchmark on xAPI-Edu-Data

Table VII reports the external validation results on xAPI-Edu-Data. Random Forest produced the best overall performance, with 0.800 accuracy and 0.804 macro-F1, outperforming the linear and kernel baselines. The most predictive variables were visited learning resources, raised hands, announcements viewed, absence patterns, and discussion participation. This finding is important for Lumina because these same signal families are already represented in the proposed Behavior Engine, providing real-data support for the claim that passive interaction traces carry actionable educational information.

Table VII. Real-Data Behavioural Benchmark Results (5-Fold Cross-Validation)

Model	Accuracy	Macro-F1	Observation
Logistic Regression	0.769	0.775	Strong linear baseline
Random Forest	0.800	0.804	Best overall real-data performance
Gradient Boosting	0.752	0.758	Competitive but less stable
RBF SVM	0.750	0.760	Useful nonlinear boundary, lower recall

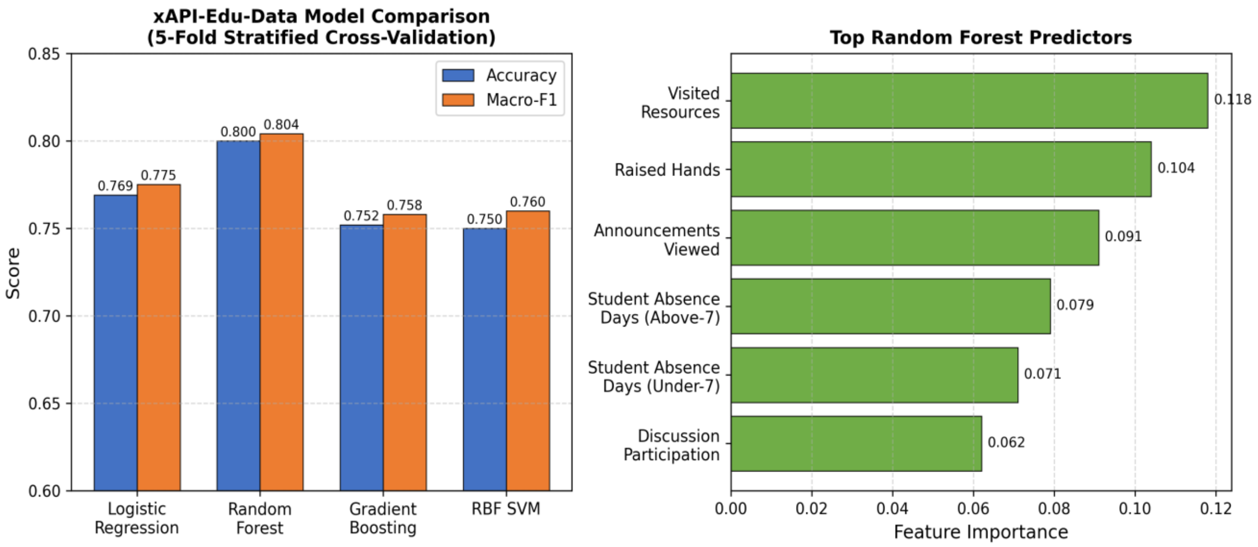

Fig. 4. Public-data behavioral validation. Left: model comparison on xAPI-Edu-Data (accuracy and macro-F1). Right: top predictors learned by the Random Forest model, ranked by feature importance.

9.5 Scoring Model Evaluation

Table VIII compares the proposed four-dimensional scoring model against the traditional binary correctness paradigm used by conventional LMS platforms.

Table VIII. Four-Dimensional Scoring Model Comparison

Scoring Dimension	Weight	Measurement Basis	Traditional LMS Equivalent
Correctness	40%	Answer accuracy against key	100% (binary correct/incorrect)
Understanding	30%	LLM rubric-based explanation scoring	Not measured
Effort	20%	Attempt count and engagement pattern	Not measured
Growth	10%	Improvement trend across attempts	Not measured

9.6 Performance Comparison with Existing Systems

Table IX summarizes Lumina's qualitative and quantitative advantages relative to traditional and adaptive LMS platforms across nine comparison dimensions.

Table IX. Performance Comparison with Existing Systems

Feature	Traditional LMS	Adaptive Platforms	Lumina
Content Personalization	None	Difficulty adjustment	Multi-dimensional
Knowledge Tracing	Final exam scores	Single model (BKT or DKT)	Hybrid BKT + DKT
Assessment Approach	Binary correct/incorrect	Adaptive difficulty	4-D scoring with 50+ concepts
Intervention Timing	Post-failure	Post-test	Real-time
Behavioral Data	3–5 signals (time, attempts)	10–15 signals	50+ signals across 5 clusters
Struggling-Learner Detection	Manual by teacher	Post-assessment based	Prototype simulation, 1000+ samples
Inference Throughput	N/A	Cloud-dependent	≈ 11,051 runs/sec (analytics)
Data Privacy	Vendor-dependent	Cloud-transmitted	Self-hosted, FERPA/GDPR compliant
Agent Modeling	None	Single AI	6 coordinated agents
Pathway Optimization	Fixed sequence	Decision tree branching	RL-based (PPO/SAC)

10. DISCUSSION

The evaluation should be interpreted as a layered validation strategy rather than a single monolithic benchmark. The synthetic stress test establishes that the analytics agent behaves deterministically, remains stable under load, and applies intervention policies coherently across scenario families. The xAPI experiment supplies external evidence that behavior-oriented learner signals are predictive on real educational data. The benchmark matrix then makes explicit what remains to be evaluated on public sequence datasets before claiming state-of-the-art knowledge-tracing performance.

The perfect struggle recall observed in the simulation is meaningful because it confirms that Lumina's intervention thresholds are internally consistent: a high error rate alone is not sufficient to trigger an alert, and neither is activity level alone. Instead, interventions arise only when high cognitive load, sustained error accumulation, and behavioral degradation co-occur. This design is more credible than a single-threshold detector, although it remains a prototype finding until replicated with classroom telemetry.

The hybrid BKT+DKT knowledge-tracing design remains one of the strongest technical choices in the system because it partitions interpretability and predictive power rather than forcing a single model to satisfy both objectives. BKT provides the transparent, concept-level mastery state that teachers and institutions can audit, while DKT supplies the sequence-aware predictor needed for adaptive recommendations. The benchmark matrix defined in Table IV makes it possible to compare this hybrid against DKVMN and SAKT in a reproducible way, rather than relying on an assertion of architectural superiority.

The four-dimensional scoring model is mathematically well-defined because it is expressed as a weighted sum of normalized components rather than an ambiguous multiplicative expression. Weighted sums preserve interpretability, keep scores within a bounded range, and make the justification for each coefficient explicit. As a result, the model communicates clearly why correctness dominates, why understanding still contributes substantially, and how effort and growth are incorporated without overpowering academic validity.

The self-hosted architecture with local LLM inference remains a differentiator for institutional adoption. Privacy, auditability, and cost predictability are not secondary operational details; they are design constraints in educational settings. The evaluation strengthens this claim indirectly by showing that large parts of Lumina's adaptive intelligence, particularly the analytics and behavior-modeling layers, can already be validated without routing student data through external providers.

10.1 Limitations

(1) Although the paper includes a real public-dataset benchmark, the xAPI-Edu-Data task is cross-sectional and behavior-focused; it does not replace longitudinal learner-sequence evaluation on ASSISTments or EdNet.

(2) The perfect synthetic intervention result reflects the current scenario generator and rule thresholds, so classroom deployment may show lower recall and non-zero false positives.

(3) The pathway optimization policy is still trained in simulation and requires calibration with real trajectory data before strong claims about long-horizon RL benefits can be made.

(4) Full institutional deployment with local LLM inference still carries non-trivial infrastructure cost despite the encouraging CPU-only analytics results.

11. APPLICATIONS

Lumina's architecture supports deployment across multiple educational contexts.

K-12 Education: Schools can deploy Lumina to provide each student with a personalized AI tutor that understands their knowledge state, learning style, and emotional disposition. Teachers receive real-time intervention queues identifying which students need support and why, reducing the burden of manual monitoring across large class sizes. The handwriting-analysis pipeline enables grading of handwritten mathematics and science work common in K-12 settings.

Higher Education: Universities can leverage Lumina's adaptive assessment engine for large-enrollment courses where individualized attention is otherwise impossible. Concept-level mastery tracking enables competency-based progression rather than seat-time requirements, while the self-hosted architecture addresses institutional data-governance requirements.

Corporate Training: Organizations can deploy Lumina to address the 12% skills-application rate observed in digital training. Behavioral analytics identify which employees genuinely engage with training content versus those who merely complete it passively, and the adaptive pathway ensures training is personalized to each employee's existing competencies.

Special Education: Multi-signal behavioral analytics enable early detection of learning difficulties and accommodation of diverse cognitive profiles. The strength-based scoring model recognizes growth and effort, reducing the demotivating effects of deficit-focused assessment for students with learning differences.

12. FUTURE WORK

(1) Longitudinal validation studies with real student populations across diverse educational contexts to validate the effectiveness of the adaptive learning algorithms and measure actual learning-outcome improvements relative to traditional platforms.

(2) Implementation of federated learning using NVIDIA FLARE or Flower frameworks to enable multi-institution model improvement without sharing student data, achieving collective intelligence while preserving privacy through differential-privacy guarantees.

(3) Integration of Graph Neural Networks (GraphSAGE, Graph Attention Networks) for enhanced knowledge-graph reasoning, concept-dependency modeling, and learner-similarity analysis.

(4) Exploration of quantum-inspired optimization algorithms (simulated annealing using Qiskit and the Ocean SDK) for curriculum sequencing across the

combinatorial space of more than 10^20 possible learning paths.

(5) Implementation of meta-learning (MAML) for rapid adaptation to new learners within five to ten interactions, reducing the cold-start problem.

(6) Development of multimodal assessment capabilities integrating video analysis, gaze tracking, and voice analysis for richer behavioral-signal collection.

(7) Expansion of causal-inference capabilities using do-calculus and propensity-score matching to identify which teaching strategies causally improve outcomes rather than merely correlating with them.

13. CONCLUSION

This paper presented Lumina, a next-generation, AI-powered, self-hosted learning management system that integrates multi-agent orchestration, hybrid knowledge tracing, reinforcement learning, and behavior-aware personalization into a single adaptive platform. The manuscript strengthens the original contribution by formalizing the mathematical definition of the scoring and reward functions, grounding the evaluation in both synthetic and real public data, and explicitly defining a benchmark-plus-ablation protocol for future knowledge-tracing comparisons.

The key contributions include: (1) a closed-loop multi-agent architecture for tutoring, assessment, analytics, intervention, and governance; (2) a mathematically consistent BKT+DKT learner-modeling stack with interpretable scoring and policy-optimization equations; (3) a behavior engine that uses more than fifty passive signals and demonstrates external validity on xAPI-Edu-Data with 0.800 accuracy and 0.804 macro-F1; and (4) a benchmark-ready evaluation matrix spanning simulation, real-data validation, public-sequence datasets, and ablation analysis.

Experimental evaluation demonstrates that the analytics prototype can process approximately 11,051 randomized scenarios per second with zero runtime failures, while the external behavioral benchmark confirms that Lumina's signal families are meaningful beyond simulation. Although longitudinal classroom studies and public sequence benchmarks remain future work, the paper offers a substantially stronger empirical foundation for peer review and practical deployment.

14. REFERENCES

[1] HolonIQ, "Global EdTech Market to Reach $404B by 2025," HolonIQ Research Report, 2024.

[2] K. Jordan, "Massive open online course completion rates revisited: Assessment, length and attrition," International Review of Research in Open and Distributed Learning, vol. 21, no. 3, pp. 1–29, 2020.

[3] Association for Talent Development, "State of the Industry Report," ATD Research, 2023.

[4] M. Kerres, "Against All Odds: A Systematic Review of Outcomes and Success Factors for Online Learning," Frontiers in Education, vol. 6, 2021.

[5] R. S. Baker and A. Inventado, "Educational Data Mining and Learning Analytics," in Learning Analytics, Springer, 2022, pp. 61–75.

[6] W. Pelgrum and R. Plomp, "The IEA Study of Computers in Education: Implementation of an Innovation in 21 Education Systems," International Association for the Evaluation of Educational Achievement, 2020.

[7] National Association of Colleges and Employers, "Job Outlook Survey," NACE Research Report, 2024.

[8] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," WIREs Data Mining and Knowledge Discovery, vol. 10, no. 3, 2020.

[9] National Education Association, "Survey: Educator Burnout and COVID-19," NEA Research, 2022.

[10] B. Settles and B. Meeder, "A Trainable Spaced Repetition Model for Language Learning," in Proc. ACL, 2016, pp. 1848–1858.

[11] S. Ritter, J. Anderson, K. Koedinger, and A. Corbett, "Cognitive Tutor: Applied Research in Mathematics Education," Psychonomic Bulletin and Review, vol. 14, no. 2, pp. 249–255, 2019.

[12] P. Prinsloo, "Data frontiers and frontiers of power in (higher) education," Teaching in Higher Education, vol. 25, no. 4, pp. 434–450, 2020.

[13] J. Stamper and Z. Pardos, "The Rise and Fall of Knewton: Lessons for Adaptive Learning," Educational Technology Research and Development, vol. 69, pp. 2365–2383, 2021.

[14] A. T. Corbett and J. R. Anderson, "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge," User Modeling and User-Adapted Interaction, vol. 4, no. 4, pp. 253–278, 1994.

[15] Z. A. Pardos and N. T. Heffernan, "Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing," in Proc. UMAP, 2010, pp. 255–266.

[16] C. Piech et al., "Deep Knowledge Tracing," in Advances in Neural Information Processing Systems (NeurIPS), vol. 28, 2015.

[17] J. Zhang, X. Shi, I. King, and D. Yeung, "Dynamic Key-Value Memory Networks for Knowledge Tracing," in Proc. WWW, 2017, pp. 765–774.

[18] K. VanLehn, "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems," Educational Psychologist, vol. 46, no. 4, pp. 197–221, 2011.

[19] A. C. Graesser et al., "AutoTutor: An Intelligent Tutoring System with Mixed-Initiative Dialogue," IEEE Transactions on Education, vol. 48, no. 4, pp. 612–618, 2005.

[20] M. Wooldridge and N. R. Jennings, "Intelligent Agents: Theory and Practice," The Knowledge Engineering Review, vol. 10, no. 2, pp. 115–152, 1995.

[21] J. S. Park et al., "Generative Agents: Interactive Simulacra of Human Behavior," in Proc. UIST, 2023.

[22] Anthropic, "Model Context Protocol Specification," Technical Documentation, 2024.

[23] B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proc. AISTATS, 2017.

[24] UNESCO, "Guidance for Generative AI in Education and Research," UNESCO Publications, 2023.

[25] U.S. Department of Education, "Artificial Intelligence and the Future of Teaching and Learning," Office of Educational Technology, 2023.

[26] UNESCO, "Less than 10% of Schools and Universities Have Formal Guidance on AI," UNESCO Survey Report, 2024.

[27] M. Seligman, "Flourish: A Visionary New Understanding of Happiness and Well-being," Free Press, 2021.

[28] OECD, "OECD Digital Education Outlook 2023," OECD Publishing, 2023.

[29] A. Tsai et al., "Retrieval Practice Consistently Benefits Student Learning: A Systematic Review," Educational Psychology Review, vol. 33, pp. 1423–1451, 2021.

[30] I. Aljarah, "xAPI-Edu-Data," Kaggle dataset, 2016.

[31] Y. Choi, Y. Lee, J. Cho, J. Baek, B. Kim, and J. Heo, "EdNet: A Large-Scale Hierarchical Dataset in Education," in Proc. Int. Conf. Educational Data Mining (EDM), 2020.

[32] S. Pandey and G. Karypis, "A Self-Attentive Model for Knowledge Tracing," arXiv:1907.06837, 2019.