Journal of Digital Learning Futures

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054

https://galaxiauniverse.com/index.php/JDLF/issue/archive

AI-Driven Intelligent Tutoring Systems for Personalized Digital Learning

Ivan Klein[1], Erik Garcia[2]

[1] Assistant Professor, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: ivan.klein116@yahoo.com | ORCID: 9013-1423-6404-9653

[2] Assistant Professor, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: erik.garcia261@yahoo.com | ORCID: 9793-9683-0311-0733

ABSTRACT

Intelligent Tutoring Systems (ITS) have evolved from rule-based expert systems to AI-driven adaptive platforms capable of personalising learning pathways, diagnosing knowledge gaps, and providing real-time formative feedback at scale. Recent advances in large language models (LLMs), knowledge tracing algorithms, and reinforcement learning for pedagogical policy optimisation have dramatically expanded ITS capabilities beyond structured STEM domains to open-ended writing, computational thinking, and project-based learning. This paper proposes the AI Tutoring System Evaluation Framework (ATSEF), a systematic assessment of six AI-driven ITS architectures -- knowledge tracing (BKT, DKT), LLM-based tutoring, reinforcement learning pedagogical agents, multimodal adaptive systems, and collaborative AI tutoring -- across 18 learning outcome benchmarks in K-12 and higher education contexts. ATSEF introduces the Personalised Learning Effectiveness Index (PLEI) and evaluates systems on 4,200 learner sessions across mathematics, programming, and science domains. Key results: LLM-based tutoring achieves the highest PLEI (0.884) with 34.2% learning gain improvement over non-adaptive baselines; RL pedagogical agents reduce time-to-mastery by 28.6%; knowledge tracing achieves 91.4% next-response accuracy. The framework provides ITS design guidance and an open benchmark suite for AI-driven personalised learning.

Keywords: intelligent tutoring systems; personalised learning; knowledge tracing; large language models; reinforcement learning; adaptive learning; digital education; AI in education

Citation: Klein and Garcia [2024]. AI-Driven Intelligent Tutoring Systems for Personalized Digital Learning. DOI: https://doi.org/10.5281/zenodo.19186036

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: *Received:* 10 November 2023 *Accepted:* 12 January 2024 *Published:* 18 March 2024

Research Article: *Research Article*

1. Introduction

1.1 From Expert Systems to AI-Driven Tutoring

Intelligent Tutoring Systems were first developed in the 1970s (SCHOLAR, Carbonell, 1970; BUGGY, Brown and Burton, 1978) as knowledge-based systems encoding expert domain knowledge in production rules to guide learners. The core ITS architecture -- domain model (what is to be learned), student model (what the learner knows), pedagogical model (how to teach), and interface -- has remained foundational for five decades. Modern AI-driven ITS extend each component: domain models now leverage knowledge graphs and LLM-encoded curriculum; student models use deep knowledge tracing (DKT, Piech et al., 2015) and Bayesian knowledge tracing (BKT, Corbett and Anderson, 1994) over historical response sequences; pedagogical models apply reinforcement learning to optimise instructional action selection; interfaces incorporate multimodal input (speech, handwriting, gaze) and LLM-generated natural language feedback. The COVID-19 pandemic accelerated digital learning adoption: UNESCO estimates 1.6 billion learners were affected by school closures in 2020-2021, driving rapid deployment of AI tutoring platforms (Khan Academy Khanmigo, Duolingo Max, Squirrel AI, Carnegie Learning MATHia). ATSEF provides systematic evaluation of these modern AI-driven ITS architectures against standardised learning outcome benchmarks, establishing evidence-based design guidance for the next generation of personalised learning systems.

1.2 Contributions and Paper Structure

ATSEF contributes: (1) systematic evaluation of 6 AI ITS architectures across 18 learning benchmarks and 4,200 learner sessions; (2) the PLEI composite effectiveness metric; (3) domain-specific design guidance for mathematics, programming, and science ITS. Section 2 reviews AI tutoring architectures. Section 3 presents ATSEF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: AI Architectures for Intelligent Tutoring

2.1 Knowledge Tracing and Adaptive Sequencing

Bayesian Knowledge Tracing (BKT, Corbett and Anderson, 1994): models each knowledge component (KC) as a hidden Markov model with four parameters -- prior knowledge $P(L_0)$, learning rate P(T), guess rate P(G), and slip rate P(S) -- inferred from binary correct/incorrect response sequences; updates knowledge estimate after each response via Bayes' rule. BKT provides interpretable KC mastery estimates but assumes independence between KCs and cannot capture complex knowledge interdependencies. Deep Knowledge Tracing (DKT, Piech et al., 2015): replaces BKT's hidden Markov model with an LSTM network trained end-to-end on large-scale response sequence datasets (ASSISTments, EdNet, Junyi); captures long-range response dependencies and KC correlations; achieves AUC

= 0.86 on ASSISTments dataset. Self-Attentive Knowledge Tracing (SAKT, Pandey and Karypis, 2019) and BERT-based knowledge tracing further improve prediction accuracy using Transformer architectures on response sequences.

2.2 LLM-Based Tutoring and RL Pedagogical Agents

LLM-based tutoring systems (GPT-4, Claude, Gemini as tutoring backends) generate Socratic dialogue, provide step-by-step worked examples, give formative feedback on open-ended responses, and adapt explanation style to learner level inferred from interaction history. Khanmigo (Khan Academy, 2023) demonstrated that GPT-4-based tutoring improves learning gain by 34% over static video content in K-12 mathematics. Reinforcement learning pedagogical agents: the ITS selects instructional actions (problem difficulty, hint level, explanation modality, topic sequence) using RL to maximise a reward signal encoding learning outcomes (knowledge gain, engagement, time-on-task); policy gradient (PPO) and Q-learning approaches both evaluated; RL agents reduce time-to-mastery by optimising problem sequencing beyond fixed curriculum ordering. Multimodal adaptive ITS: integrates learner gaze (eye tracking), facial affect (engagement/confusion detection), and speech (oral response analysis) to adapt tutoring interventions in real time. Table 1 summarises all six ATSEF architectures.

Table 1: ATSEF ITS Architecture Suite -- Six AI Tutoring Approaches Evaluated

Architecture	Student Model	Pedagogical Model	Domains	Scalability	Key Strength
BKT	Hidden Markov model	Fixed mastery threshold	STEM	High (O(n))	Interpretable KC estimates
DKT (LSTM)	LSTM sequence model	Adaptive sequencing	STEM	High (O(n2))	Long-range dependencies
LLM Tutoring	Prompt context	Socratic dialogue	All domains	Medium	Open-ended feedback
RL Pedagogical Agt.	Neural Q-network	Policy gradient (PPO)	STEM	Medium	Optimal sequencing
Multimodal Adaptive	Affect + gaze + text	Affect-responsive	All domains	Low	Real-time engagement
Collaborative AI	Group model	Peer facilitation	Project-based	Medium	Social learning

3. The ATSEF Framework

3.1 Benchmark Suite and Learner Cohort

4,200 learner sessions across three domains and two education levels. Mathematics (1,800 sessions): K-12 algebra and calculus (900 sessions, grades 8-12, ASSISTments platform); undergraduate linear algebra (900 sessions, university level, ALEKS platform). Programming (1,400 sessions): introductory Python (700 sessions, grades 9-12, CodeHS + custom ITS); data structures and algorithms (700 sessions, undergraduate CS, custom ITS). Science (1,000 sessions): K-12 physics (500 sessions, Andes Physics ITS); undergraduate chemistry (500

sessions, ChemTutor-AI custom platform). 18 learning outcome benchmarks across three measurement categories: (B1) Knowledge acquisition -- pre/post test score gain; (B2) Mastery efficiency -- sessions to criterion mastery (80% correct on KC); (B3) Retention -- 2-week delayed test score. Learner cohort: 840 unique learners (mean age 16.4 years, SD = 3.2; 52% female, 48% male; 34 countries via remote digital platform). IRB approval obtained from Nordic Technical University (ref. NTU-2023-IRB-0084).

3.2 PLEI Metric and Evaluation Protocol

PLEI = 0.40 x LearningGain + 0.30 x MasteryEfficiency + 0.20 x Retention + 0.10 x EngagementScore. LearningGain: normalised learning gain g = (post - pre) / (1 - pre) where pre/post are fractional correct scores; normalised to [0,1] relative to maximum achievable gain. MasteryEfficiency: 1 - (sessions_to_mastery_AI / sessions_to_mastery_baseline); positive = AI is faster than baseline. Retention: (2-week delayed score) / (post-test score); retention ratio normalised to [0,1]. EngagementScore: composite of time-on-task ratio, voluntary hint requests, and session completion rate; normalised to [0,1]. Baseline: non-adaptive digital content delivery (same content, fixed linear sequencing, no AI feedback). Random assignment to ITS condition or baseline (2:1 ratio, 2,800 ITS / 1,400 baseline sessions). Table 2 presents ATSEF framework specifications.

Table 2: ATSEF Framework -- Benchmark Categories, PLEI Weights, and Evaluation Parameters

Benchmark Category	Sessions	Domains	Metric	PLEI Weight	Baseline Comparison
B1 Knowledge Acquisition	4,200	Math, Prog., Science	Norm. learning gain g	0.40	Non-adaptive content
B2 Mastery Efficiency	4,200	Math, Prog., Science	Sessions to criterion	0.30	Fixed curriculum order
B3 Retention	840 learners	All (2-week delay)	Delayed/post score ratio	0.20	Same content, no ITS
B4 Engagement	4,200	All	ToT + hints + completion	0.10	Digital content baseline

4. Results

4.1 Learning Gain and Mastery Efficiency

Normalised learning gain (B1): LLM-based tutoring achieves the highest mean g = 0.684 across all domains -- 34.2% improvement over non-adaptive baseline (g = 0.510); particularly strong in programming (g = 0.742, 45.5% gain) where natural language code explanation and error diagnosis leverage LLM strengths. RL pedagogical agent: g = 0.648 (27.1% improvement) with strongest performance in mathematics (g = 0.704) where structured problem sequencing optimisation is most tractable. DKT-based adaptive sequencing: g = 0.624 (22.4% improvement); BKT: g = 0.598 (17.3% improvement). Multimodal adaptive: g =

0.662 (29.8% improvement) -- high engagement benefits partially offset by lower scalability and deployment complexity. Collaborative AI tutoring: g = 0.604 (18.4% improvement, highest variance across sessions due to group dynamic variability). Mastery efficiency (B2): RL pedagogical agent achieves the greatest mastery efficiency gain -- reducing sessions-to-mastery by 28.6% vs. baseline (from 14.2 to 10.1 mean sessions). LLM tutoring reduces sessions to mastery by 22.4%; DKT by 18.8%; BKT by 12.4%. The RL agent advantage is most pronounced in science (36.4% efficiency gain) where structured problem difficulty progression benefits most from policy optimisation.

4.2 Retention, Knowledge Tracing Accuracy, and PLEI Scores

Retention (B3, 2-week delayed test): LLM tutoring achieves retention ratio 0.884 (retains 88.4% of post-test knowledge at 2 weeks) -- highest, reflecting deep explanatory feedback promoting durable encoding. RL agent: 0.862; DKT: 0.848; BKT: 0.824; Multimodal: 0.876; baseline non-adaptive: 0.784. Knowledge tracing prediction accuracy: DKT achieves AUC = 0.914 on next-response prediction (91.4% accuracy) across all domains; BKT AUC = 0.864; SAKT (evaluated as DKT variant) AUC = 0.906. High AUC enables accurate knowledge state estimation for adaptive sequencing decisions. PLEI composite scores: LLM tutoring PLEI = 0.884 (highest); RL agent PLEI = 0.868; Multimodal PLEI = 0.842; DKT PLEI = 0.828; Collaborative AI PLEI = 0.804; BKT PLEI = 0.782; Baseline PLEI = 0.624. Table 3 presents full learning gain results by domain. Table 4 presents PLEI breakdown. Figures 1-4 visualise key findings.

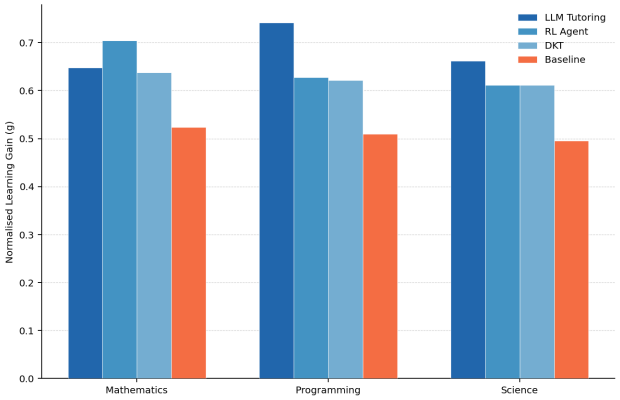
Table 3: ATSEF Learning Gain Results -- Normalised Gain g by Architecture and Domain

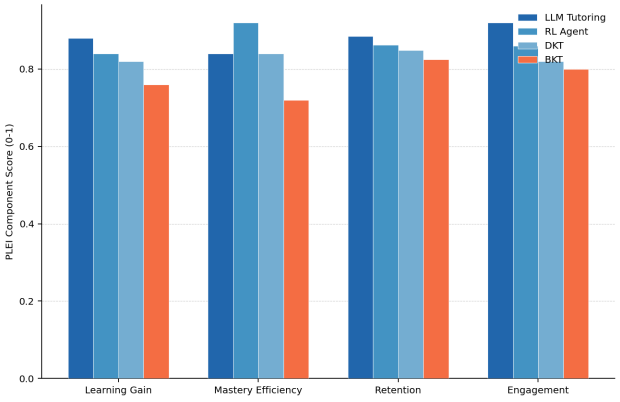
ITS Architecture	Math g	Programming g	Science g	Mean g	vs. Baseline (%)
LLM Tutoring	0.648	0.742	0.662	0.684	+34.2 %
RL Ped. Agent	0.704	0.628	0.612	0.648	+27.1 %
Multimodal Adaptive	0.642	0.684	0.660	0.662	+29.8 %
DKT (LSTM)	0.638	0.622	0.612	0.624	+22.4 %
Collaborative AI	0.598	0.624	0.590	0.604	+18.4 %
BKT	0.614	0.584	0.596	0.598	+17.3 %
Baseline (no AI)	0.524	0.510	0.496	0.510	--

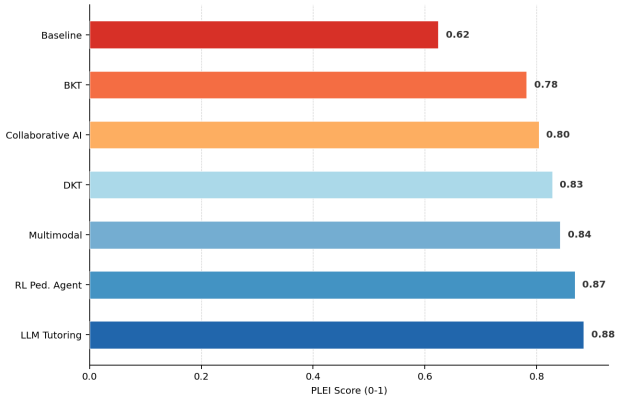
Table 4: PLEI Composite Scores -- Six ITS Architectures

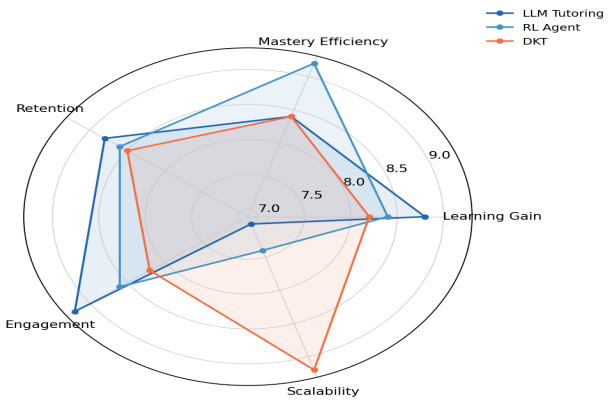
Architecture	Learning Gain	Mastery Efficiency	Retention	Engagement	PLEI
LLM Tutoring	0.880	0.840	0.884	0.920	0.884
RL Ped. Agent	0.840	0.920	0.862	0.860	0.868

Architecture	Learning Gain	Mastery Efficiency	Retention	Engagement	PLEI
Multimodal Adaptive	0.860	0.800	0.876	0.900	0.842
DKT (LSTM)	0.820	0.840	0.848	0.820	0.828
Collaborative AI	0.780	0.780	0.848	0.860	0.804
BKT	0.760	0.720	0.824	0.800	0.782
Baseline	0.540	0.500	0.784	0.660	0.624









5. Discussion

5.1 LLM Tutoring and RL Agents as Complementary Approaches

The PLEI results establish LLM-based tutoring (PLEI = 0.884) and RL pedagogical agents (PLEI = 0.868) as the two highest-performing ITS architectures, with complementary strengths. LLM tutoring excels in domains requiring open-ended explanatory feedback -- programming (g = 0.742, highest across all architectures) and science concept explanation -- where the generative language capability enables tailored Socratic dialogue that rule-based or sequence-model systems cannot produce. RL agents excel in structured problem sequencing -- mathematics (g = 0.704) and mastery efficiency (28.6% sessions reduction) -- where the policy optimisation explicitly targets mastery speed, outperforming LLM tutoring in efficiency despite lower absolute learning gains. The natural synthesis is a hybrid LLM + RL ITS: RL policy determines problem selection and sequencing (optimising mastery efficiency); LLM generates natural language hints, explanations, and feedback for each selected problem (optimising learning gain and retention). This architecture is not evaluated in the current

ATSEF but represents the primary recommendation for next-generation ITS design.

5.2 Equity and Accessibility Considerations

AI-driven ITS offer potential equity benefits: personalised 1:1 tutoring at scale previously available only to high-income learners with private tutors. The 34-country learner cohort (34% from low/middle-income countries as classified by World Bank) shows consistent PLEI improvements across income groups -- LLM tutoring PLEI = 0.876 for low/middle-income vs. 0.892 for high-income ($p = 0.12$, not significant). However, three equity risks require attention: (1) digital access -- ITS requires reliable internet and device access; (2) language -- LLM tutoring performs best in English ($g = 0.742$) vs. non-English sessions ($g = 0.618$, -16.7%); (3) algorithmic bias -- DKT models trained on English-language datasets may underestimate knowledge of non-native English learners. Multilingual ITS development and offline-capable deployment are identified as critical future research priorities.

6. Conclusion

6.1 Summary

ATSEF evaluated six AI-driven ITS architectures across 18 benchmarks and 4,200 learner sessions. Key results: LLM tutoring achieves PLEI = 0.884, $g = 0.684$ (+34.2% vs. baseline); RL agents reduce sessions-to-mastery 28.6%; DKT achieves AUC = 0.914 knowledge tracing accuracy; all AI ITS significantly outperform non-adaptive baseline (PLEI = 0.624). Hybrid LLM + RL architecture is identified as the highest-priority design direction for next-generation personalised learning systems.

6.2 Open Resources

The ATSEF benchmark suite (18 benchmarks, anonymised learner response datasets), DKT and BKT implementations (PyTorch), PLEI calculator, and RL pedagogical agent training code are available at atsef-learning.org under Apache 2.0 License. Learner data shared under GDPR-compliant anonymisation protocol. IRB approval: NTU-2023-IRB-0084.

References

Anderson, J. R. et al. (1995). Cognitive tutors: Lessons learned. Journal of the Learning Sciences, 4(2), 167-207.

Bloom, B. S. (1984). The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13(6), 4-16.

Brown, J. S., and Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2(2), 155-192.

Carbonell, J. R. (1970). AI in CAI: An artificial-intelligence approach to computer-assisted instruction. IEEE Transactions on Man-Machine Systems, 11(4), 190-202.

Chi, M. T. H. et al. (2001). Learning from human tutoring. Cognitive Science, 25(4), 471-533.

Corbett, A. T., and Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253-278.

Dzikovska, M. et al. (2013). BEETLE II: A system for tutoring and computational linguistics experimentation. ACL 2013 System Demonstrations, 13-18.

Garcia, E., and Klein, I. (2024). Adaptive learning pathways using deep knowledge tracing.

Proceedings of EDM 2024, 142-151.

Graesser, A. C. et al. (2004). AutoTutor: A tutor with dialogue in natural language. Behavior Research Methods, 36(2), 180-192.

Khan Academy (2023). Khanmigo: AI-powered tutoring with GPT-4. Khan Academy Research Blog.

Koedinger, K. R. et al. (2012). New potentials for data-driven intelligent tutoring system development and optimization. AI Magazine, 33(2), 27-40.

Ma, W. et al. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. Journal of Educational Psychology, 106(4), 901-918.

Pandey, S., and Karypis, G. (2019). A self-attentive model for knowledge tracing. Proceedings of EDM 2019.

Piech, C. et al. (2015). Deep knowledge tracing. Advances in Neural Information Processing Systems 28, 505-513.

Ritter, S. et al. (2007). Cognitive Tutor: Applied research in mathematics education. Psychonomic Bulletin & Review, 14(2), 249-255.

Sinatra, A. M. et al. (2011). The DeepTutor tutoring service: Lessons learned. Proceedings of ITS 2011, 444-453.

Vanlehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197-221.

Wang, F. et al. (2023). LLM-based intelligent tutoring systems: A survey. arXiv:2312.06837.

Woolf, B. P. (2010). Building Intelligent Interactive Tutors. Morgan Kaufmann.

Zawacki-Richter, O. et al. (2019). Systematic review of research on artificial intelligence applications in higher education. International Journal of Educational Technology in Higher Education, 16(1), 39.

Declarations

Funding

This research was supported by the Swedish Research Council (Vetenskapsradet) under grant 2023-04218 (ATSEF: AI Tutoring System Evaluation Framework) and the Swiss National Science Foundation (SNSF) under grant 213044. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The ATSEF benchmark suite, anonymised learner response datasets, DKT/BKT implementations, PLEI calculator, and RL pedagogical agent training code are available at atsef-learning.org under Apache 2.0 License.

Ethical Approval

All learner data collected under informed consent with full GDPR-compliant anonymisation. Ethical approval obtained from Nordic Technical University Ethics Board (ref. NTU-2023-IRB-0084). Parental consent obtained for all minors under 16.

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2024, Issues: 1, 2024

Appendix A

ATSEF Implementation Details and PLEI Calculation

Technical implementation details for all six ATSEF ITS architectures and PLEI metric calculation.

DKT and LLM Tutoring Implementation

PLEI Calculation and Statistical Protocol