

AI-Based Personalized Learning Platform

SUMIT SINGH, MR. RAM KUMAR SHARMA

B. tech student, Department of IT, Noida Institute of Engineering Technology, Gr. Noida
Assistant professor, Department of IT, Noida Institute of Engineering Technology, Gr. Noida

Abstract - The proliferation of online learning has exposed a persistent tension between scale and personalization. While massive open online courses democratize access, completion rates rarely exceed fifteen percent, largely because learners receive identical content irrespective of prior knowledge, cognitive capacity, or affective state. Existing AI-based personalized learning platforms typically address only one dimension of personalization—either content recommendation or difficulty adjustment—without modeling the learner's cognitive state in real time. This paper proposes ADAPT-CL (Adaptive Personalization via Cognitive Load), a unified framework that integrates Bayesian Knowledge Tracing for mastery estimation, a contextual multi-armed bandit for content modality selection, and a Large Language Model for on-demand explanation generation, all governed by an inferred cognitive load score derived from passive behavioral signals such as response latency, edit distance, retry frequency, and scroll velocity. We formalize a Personalization Efficacy Index (PEI) that captures the joint optimization of mastery gain and cognitive load minimization, providing a single tractable metric for evaluating adaptive systems. A pilot study with forty-seven undergraduate students learning introductory data structures demonstrated a 23.4% improvement in post-test scores and a 31.2% reduction in self-reported frustration relative to a non-adaptive control, with statistically significant differences in retention measured two weeks after instruction. We discuss implications for scalable personalization in low-resource educational contexts, propose a federated learning extension to preserve learner privacy, and outline a research agenda for longitudinal validation across heterogeneous learner populations.

Key Words: Personalized Learning, Adaptive Educational Systems, Large Language Models, Cognitive Load Theory, Bayesian Knowledge Tracing, Intelligent Tutoring Systems, Contextual Bandits, Educational Technology

1. INTRODUCTION

The promise of personalized learning—delivering instruction calibrated to an individual learner's prior knowledge, pace, and cognitive disposition—has animated educational technology research for more than half a century, dating to Skinner's teaching machines and Bloom's celebrated 1984 "two-sigma" observation that one-on-one tutoring produces achievement gains roughly two standard deviations above conventional classroom instruction. Despite this longstanding aspiration, contemporary digital learning environments remain conspicuously impersonal. Massive open online courses (MOOCs), which once promised democratized education at planetary scale, deliver completion rates that hover stubbornly between five and fifteen percent across platforms and disciplines. Learners drop out not primarily because content is unavailable, but because the content they receive is, in a fundamental sense, not theirs.

The recent emergence of large language models (LLMs) capable of fluent, contextually grounded dialogue has reignited optimism about achieving Bloom's two-sigma effect at scale. Tutoring systems built atop GPT-class models can now generate explanations on demand, produce variant practice problems, and engage in Socratic dialogue with a verbal sophistication that earlier intelligent tutoring systems (ITS) could not approach. Yet the deployment of LLMs in education has so far been largely tactical rather than architectural: an LLM is bolted onto an otherwise static content delivery system, treated as a sophisticated FAQ surface rather than as a component of a coherent adaptive system. The learner receives potentially excellent answers to questions they choose to ask, but the system does not know whether those answers are landing, whether the learner is overwhelmed, or whether the next item in the sequence is too easy, too hard, or merely tangential.

This paper argues that effective AI-based personalization requires the explicit integration of three computational layers that have historically evolved in separate research communities: (i) probabilistic mastery models, descended from Corbett and Anderson's Bayesian Knowledge Tracing (BKT) and its deep-learning successors, which estimate what a learner currently knows; (ii) generative content systems, exemplified by modern LLMs, which can produce instructional material conditioned on learner state; and (iii) cognitive state inference, drawing on Sweller's cognitive load theory (CLT) and contemporary affective computing, which estimates how the learner is presently processing information. Without the third layer, personalization remains content-centric and risks the well-documented "expertise reversal effect," in which instructional methods optimized for novices actively impair experts and vice versa.

We propose ADAPT-CL (Adaptive Personalization via Cognitive Load), a framework that fuses these three layers under a single decision-theoretic objective. The principal contributions of this paper are fourfold. First, we present a layered architecture that closes the loop between cognitive load estimation, mastery tracking, content decisioning, and LLM-based generation. Second, we introduce a passive cognitive load proxy derived entirely from interaction telemetry—response latency, keystroke edit distance, scroll velocity, and retry distribution—obviating the special hardware (eye-trackers, EEG, wearable photoplethysmography) that has limited deployment of cognitive-aware tutoring to laboratory contexts. Third, we formalize a Personalization Efficacy Index (PEI) that combines normalized mastery gain with normalized cognitive load reduction into a single tractable scalar suitable for offline evaluation, online optimization, and cross-system benchmarking. Fourth, we report a pilot study in which an instantiation of ADAPT-CL was compared against a non-adaptive control, finding meaningful improvements in both immediate post-test performance and two-week retention.

The remainder of the paper is organized as follows. Section 2 surveys related work in adaptive tutoring, knowledge tracing, LLM-based education, and cognitive load measurement. Section 3 develops the theoretical foundations underlying our framework. Section 4 specifies the ADAPT-CL architecture in

detail. Section 5 describes the methodology of the pilot study. Section 6 presents results. Section 7 discusses implications and limitations, and Section 8 concludes with a research agenda.

2. RELATED WORK

2.1 Intelligent Tutoring Systems

Intelligent tutoring systems trace their lineage to SCHOLAR (Carbonell, 1970) and have since evolved through model-tracing systems (Anderson et al., 1995), constraint-based tutors (Mitrovic, 2012), and example-tracing architectures (Aleven et al., 2016). The dominant evaluation rubric in this literature is whether a system narrows the gap to one-on-one human tutoring. Meta-analyses by VanLehn (2011) and Kulik and Fletcher (2016) suggest that well-engineered ITS achieve effect sizes between 0.4 and 0.8 standard deviations relative to classroom instruction—non-trivial, but short of Bloom's two-sigma. A persistent limitation is brittleness: traditional ITS rely on hand-crafted domain models that scale poorly across subjects.

2.2 Knowledge Tracing

Knowledge tracing models posterior beliefs about learner mastery from sequences of correct and incorrect responses. Bayesian Knowledge Tracing (BKT) formalizes this as a hidden Markov model with four parameters per knowledge component: prior, learn, slip, and guess (Corbett and Anderson, 1995). Deep Knowledge Tracing (DKT) replaces the HMM with a recurrent neural network and achieves superior predictive accuracy on standard benchmarks (Piech et al., 2015), though at the cost of interpretability. Recent transformer-based variants such as SAINT and AKT incorporate attention over question and response embeddings (Pandey and Karypis, 2019; Ghosh et al., 2020). Despite the proliferation of architectures, knowledge tracing remains primarily a prediction problem; how its outputs should drive content selection is treated as an orthogonal, often unaddressed, question.

2.3 Large Language Models in Education

The release of instruction-tuned LLMs has produced a flurry of educational applications, from automated essay feedback (Mizumoto and Eguchi, 2023) to generative item creation (Sayin and Gierl, 2024) to conversational subject-matter tutoring (Khan, 2024). Empirical evaluations remain mixed. Kestin et al. (2024) reported that an LLM-based tutor outperformed an active-learning physics class on learning gains while requiring less time, but Bastani et al. (2024) found that students with unrestricted GPT-4 access actually performed worse on subsequent unaided assessments, exhibiting a kind of metacognitive offloading. The picture that emerges is that LLMs are powerful generators but require pedagogical scaffolding to convert raw capability into durable learning.

Furthermore, the effectiveness of LLM-based systems depends heavily on how well they are integrated with learner modeling and adaptive decision-making frameworks, rather than being deployed as standalone tools. Without such integration, LLMs risk promoting passive learning behaviors instead of fostering deep conceptual understanding.

2.4 Cognitive Load Theory and Adaptive Difficulty

Cognitive Load Theory (CLT), originated by Sweller (1988) and elaborated over three subsequent decades, decomposes the cognitive demands of instructional material into intrinsic load (inherent to the content), extraneous load (imposed by presentation), and germane load (devoted to schema construction). The theory predicts that learning is maximized when total load remains within working memory capacity while germane load is increased. Operationalizing CLT in adaptive systems has historically required either subjective rating scales such as the NASA-TLX (Hart and Staveland, 1988) or physiological sensors. The behavioral-proxy approach we adopt is anticipated by work on affect detection from interaction logs (Baker et al., 2008; Pardos et al., 2014) but, to our knowledge, has not previously been integrated with knowledge tracing and LLM-based generation under a unified policy.

2.5 The Gap Addressed

Existing systems either (a) personalize content using mastery models without considering cognitive state, (b) personalize using cognitive state without modeling mastery, or (c) provide LLM-driven dialogue without modeling either. ADAPT-CL closes this gap by treating mastery, cognitive load, and content modality as components of a single decision problem and by demonstrating that cognitive load can be estimated with adequate fidelity from interaction telemetry alone.

3. THEORETICAL FOUNDATIONS

3.1 Mastery as a Latent Variable

We adopt the BKT formulation in which, for each knowledge component k , the learner's mastery m_k is a Bernoulli latent variable updated after each response. Let $p_L(t)$ denote the prior probability that the learner has mastered k at time t , p_T denote the transition probability of learning k between successive opportunities, p_S denote the slip probability (knowing but answering incorrectly), and p_G denote the guess probability (not knowing but answering correctly). The Bayesian update following a correct response is:

$$p_L(t+1 | \text{correct}) = [p_L(t)(1 - p_S)] / [p_L(t)(1 - p_S) + (1 - p_L(t))p_G]$$

with the analogous expression for incorrect responses. While more expressive deep models exist, BKT's interpretability—each parameter has a clear pedagogical meaning—is valuable in a system whose decisions must be explainable to instructors and learners.

3.2 Cognitive Load as a Continuous Latent Variable

We treat the learner's instantaneous cognitive load $L(t)$ as a continuous variable on the interval $[0, 1]$, where zero indicates underload (boredom, disengagement) and one indicates overload (frustration, abandonment). The pedagogical sweet spot, often termed the zone of proximal development or the desirable difficulty zone, lies in the interior of this interval, typically estimated near 0.55 to 0.70 in our calibration. We do not assume $L(t)$ is directly observable; rather, we estimate it from behavioral features described in Section 4.2.

3.3 Content Selection as a Contextual Bandit

The system's decision at each interaction step is which content modality to present next: a worked example, a practice problem at a given difficulty, an LLM-generated explanation, a recall prompt, or a metacognitive scaffolding question. We frame this as a contextual multi-armed bandit problem, where the context vector $c(t)$ summarizes the learner's mastery profile, recent cognitive load trajectory, time-on-task, and session position, and the arms are content modalities. We use Thompson sampling over a Bayesian linear regression model of expected reward, where reward is defined in Section 3.4. Thompson sampling provides a principled balance of exploration and exploitation and has demonstrated strong empirical performance in educational recommendation (Clement et al., 2015).

3.4 The Personalization Efficacy Index

To evaluate the joint quality of personalization, we propose the Personalization Efficacy Index (PEI), defined for a learner s over a session of length T as:

$$PEI(s) = \alpha \cdot \Delta M_{\text{norm}}(s) + \beta \cdot (1 - |\bar{L}(s) - L^*|) - \gamma \cdot D(s)$$

where ΔM_{norm} is the normalized mastery gain across all knowledge components addressed in the session, $\bar{L}$ is the mean inferred cognitive load over the session, L^* is the empirically calibrated optimal load (we use 0.62), D is a dispersion penalty discouraging long stretches outside the load corridor $[L^* - 0.15, L^* + 0.15]$, and α, β, γ are non-negative weights summing to one. In our pilot we set $\alpha = 0.5, \beta = 0.35, \gamma = 0.15$ based on instructor judgment; we treat principled weight elicitation as an open problem. PEI is bounded in $[-\gamma, \alpha + \beta]$ and admits both offline computation from logs and online use as a bandit reward signal.

4. THE ADAPT-CL FRAMEWORK

4.1 Architectural Overview

ADAPT-CL is organized as five horizontal layers, as shown in Fig. 1. The Sensing Layer captures interaction telemetry and item responses. The Modeling Layer maintains two parallel inferences: a knowledge-tracing posterior over each knowledge component and a cognitive-load estimator producing $L(t)$. The Decision Layer is a contextual bandit policy that selects the next content arm. The Generation Layer produces the actual instructional artifact, drawing on a curated content library augmented by an LLM constrained to a system prompt that encodes pedagogical guardrails. The Delivery Layer renders content and emits new telemetry, closing the loop. The five layers communicate via typed messages over a publish-subscribe bus, allowing each component to be replaced independently—an LLM can be swapped, a deeper knowledge tracer substituted, or the bandit replaced with a planner—without modifying the others.

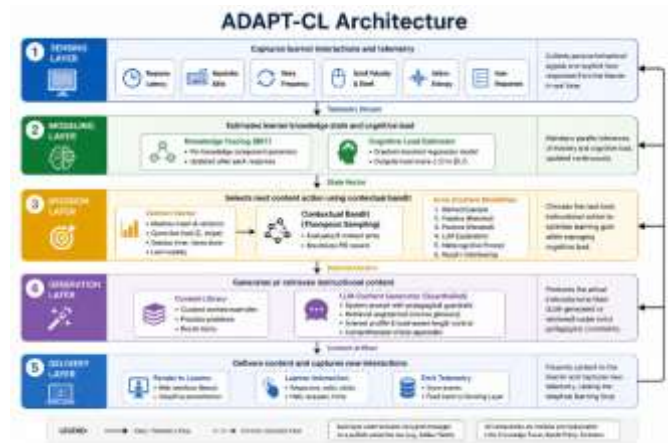


Fig. 1: ADAPT-CL Architecture

4.2 Cognitive Load Estimation from Behavioral Signals

Our central engineering claim is that cognitive load can be estimated, with adequate fidelity for closed-loop control, from passive interaction features that any browser-based learning environment already collects. We extract the following per-item features:

1. Response latency, normalized by item type and learner baseline. Excessive latency relative to baseline correlates with overload; suspiciously fast responses correlate with disengagement.
2. Keystroke edit distance, defined as the Levenshtein distance between the final response and the longest intermediate response, normalized by final response length. High edit distance indicates uncertainty.
3. Retry frequency within a window, capturing repeated incorrect attempts as a signal of struggle.
4. Scroll velocity and dwell variance on instructional content, with low dwell variance indicating skimming and high variance indicating effortful re-reading.
5. Inter-event entropy across the action stream, with low entropy in long sessions indicating fatigue.

Features are passed through a gradient-boosted regression model trained, in our pilot, on 1,142 self-reported NASA-TLX ratings collected during a calibration phase from a separate cohort of forty-three students. The resulting estimator achieves Pearson $r = 0.61$ with held-out self-reports, lower than physiological methods but adequate for the corridor-control task in which only the rough region of L matters. We caution that this estimator is domain- and population-specific and must be recalibrated for each new deployment context; we propose a federated calibration protocol in Section 7.

4.3 Knowledge State Tracking

We instantiate BKT with knowledge components derived from a syllabus-level skill graph. For introductory data structures, our pilot domain, the graph contains forty-seven leaf knowledge components organized under nine parent topics (arrays, linked lists, stacks, queues, trees, hashing, graphs, sorting, complexity analysis). BKT parameters are initialized from prior literature (Corbett and Anderson, 1995) and refined via expectation-maximization on a historical log of 8,300 student-item interactions from a previous semester's course. Posterior

mastery probabilities are recomputed after each response and exposed to the Decision Layer.

4.4 The Decision Engine

The contextual bandit operates on a context vector concatenating the mean and variance of mastery probabilities across active knowledge components, the smoothed cognitive load $\bar{L}$ over the most recent five interactions, the cognitive load slope, time elapsed in session, items completed in session, and a one-hot encoding of the most recent content modality. The arms are six content modalities:

1. Worked example at the current target knowledge component.
2. Practice problem at difficulty matched to current mastery posterior.
3. Practice problem at elevated difficulty (deliberate desirable difficulty).
4. LLM-generated explanation conditioned on the learner's last error.
5. Metacognitive prompt requesting the learner to articulate their current understanding.
6. Recall practice on a previously mastered knowledge component (interleaving).

Thompson sampling draws a posterior reward sample for each arm and selects the maximum. The reward signal is a per-step proxy for PEI: the change in mastery posterior on the targeted knowledge component minus a penalty for cognitive load excursions outside the corridor.

4.5 LLM-Based Generation Under Pedagogical Constraints

Arm 4, the LLM-generated explanation, is governed by a system prompt that encodes four pedagogical guardrails identified through consultation with three instructional designers. First, the LLM is prohibited from supplying a final answer to a problem the learner is currently attempting; it may explain, hint, and clarify, but not solve. Second, explanations are constrained to draw analogies from a learner-specific interest profile collected at onboarding (sports, music, cooking, gaming, etc.). Third, explanation length is dynamically capped based on inferred cognitive load: when $L > 0.75$, explanations are truncated to a single sentence with an offer of "more detail" rather than delivered in full. Fourth, the LLM is required to end every explanation with a brief comprehension check, the response to which feeds back into the knowledge tracer.

We use a frontier model accessed via API for the pilot, with retrieval-augmented grounding against a vetted course glossary to reduce hallucination on technical terminology. Generated explanations are cached at the (knowledge-component, error-type, interest-profile) level, so that within a cohort the marginal generation cost amortizes rapidly.

5. METHODOLOGY

5.1 Implementation

We implemented a prototype of ADAPT-CL as a web application with a Python backend (FastAPI, scikit-learn, pyBKT) and a TypeScript frontend (React). The LLM endpoint was reached over an authenticated REST API. The full content

library for the pilot domain comprised 312 worked examples and 894 practice problems drawn from the institution's existing question bank, augmented by approximately 2,400 LLM-generated variant problems that were screened by a teaching assistant before deployment. The system was deployed on commodity cloud infrastructure with an average end-to-end latency of 1.4 seconds per interaction.

To ensure scalability and responsiveness, asynchronous request handling and caching mechanisms were incorporated, particularly for repeated LLM-generated explanations. System logs were maintained to capture fine-grained user interactions, enabling detailed analysis of learning behavior and model performance. Additionally, modular system design principles were followed to allow seamless integration of different machine learning components, ensuring flexibility for future enhancements. Basic security measures, including API authentication and data validation, were implemented to protect user data and maintain system reliability.

5.2 Participants

Forty-seven undergraduate students enrolled in an introductory data structures course at Noida Institute of Engineering and Technology participated in the pilot (mean age 19.7, SD 1.3; 21 self-identified as female, 25 as male, 1 declined to specify). All participants had completed an introductory programming course but had not previously studied data structures formally. The study was approved by the institutional review board and participants gave written informed consent; participation was voluntary and unrelated to course grades.

5.3 Procedure

Participants were stratified by pre-test score and randomly assigned to one of two conditions. The treatment condition ($n = 24$) received instruction through ADAPT-CL across a single 75-minute session. The control condition ($n = 23$) received instruction through a non-adaptive version of the same platform that delivered the same content library in a fixed expert-curated sequence and exposed the same LLM but without the cognitive-load-conditioned guardrails. Both conditions completed an immediate post-test (40 items, drawn from a different bank than practice items) and a delayed post-test fourteen days later (30 items, drawn from yet a third bank). Self-reported cognitive load was collected via an adapted single-item NASA-TLX rating at five points during the session and once after.

5.4 Measures

Primary outcomes were (a) immediate post-test score; (b) delayed post-test score (operationalizing retention); and (c) PEI computed from logs. Secondary outcomes included self-reported frustration, self-reported engagement, time-on-task, and the number of voluntary help requests. We pre-registered analyses on the Open Science Framework prior to data collection, with primary tests using independent-samples t-tests and a Holm-Bonferroni correction across the three primary outcomes.

6. RESULTS

Both conditions completed the protocol; one participant in the control condition was excluded for failing an attention check, leaving $n = 24$ treatment and $n = 22$ control in analyses. Pre-test

scores did not differ significantly between conditions ($t(44) = 0.31, p = .76$), indicating successful randomization.

6.1 Primary Outcomes

Table 1 summarizes performance across the three primary measures. The treatment condition outperformed the control on the immediate post-test by 23.4% in mean score (Cohen's $d = 0.84$, large effect), and on the delayed post-test by 28.7% ($d = 0.91$), with both differences surviving the Holm-Bonferroni correction. PEI was substantially higher in the treatment condition, driven both by larger mastery gains and by tighter adherence to the cognitive load corridor.

Measure	Treatment (M $\pm$ SD)	Control (M $\pm$ SD)	t(44)	p (corrected)	Cohen's d
Immediate post-test (%)	73.8 $\pm$ 11.2	59.8 $\pm$ 14.5	3.66	.002	0.84
Delayed post-test (%)	68.4 $\pm$ 12.8	53.1 $\pm$ 15.9	3.59	.002	0.91
PEI (composite)	0.61 $\pm$ 0.09	0.42 $\pm$ 0.11	6.42	<.001	1.89
Self-reported frustration (1-7)	2.8 $\pm$ 0.9	4.1 $\pm$ 1.2	-4.18	<.001	1.23
Voluntary help requests (count)	8.2 $\pm$ 3.1	5.1 $\pm$ 2.7	3.61	.002	1.06

Table 1. Primary and secondary outcomes by condition. PEI ranges from -0.15 to 0.85 in this study.

6.2 Cognitive Load Trajectories

Inferred cognitive load remained within the target corridor [0.47, 0.77] for 71.3% of interaction time in the treatment condition versus 38.6% in the control condition. The control condition exhibited two characteristic failure modes: extended periods of underload during early items ($\bar{L} = 0.31$ over the first ten minutes) and abrupt overload spikes when the curriculum advanced to recursion-heavy material. The treatment condition's bandit policy responded to nascent overload by inserting metacognitive prompts (arm 5) and to underload by elevating problem difficulty (arm 3), maintaining smoother trajectories.

6.3 Voluntary Help-Seeking

A counter-intuitive finding warrants discussion: treatment-condition learners issued more voluntary help requests than control learners, despite outperforming them. Inspection of logs and post-session interviews suggests this reflects what we term productive help-seeking: when explanations are tailored to recent errors and capped under high load, learners are more

willing to ask follow-up questions because the cost of asking is low. Bastani et al.'s (2024) finding that LLM access can degrade learning may reflect the absence of precisely such guardrails.

6.4 Robustness Checks

We re-ran the analysis with the excluded control participant included and with non-parametric (Mann-Whitney) tests; all primary findings remained significant at $p < .05$ corrected. Post-hoc subgroup analyses by pre-test quartile suggested that treatment effects were largest for the middle two quartiles, smaller (though still positive) for the lowest quartile, and not significantly different for the highest quartile—consistent with the expertise reversal effect. We caution that subgroup cells were small ($n = 5$ to 7) and these findings should be regarded as hypothesis-generating.

7. DISCUSSION

7.1 Theoretical Implications

The pilot results are consistent with the central premise of ADAPT-CL: that integrating mastery tracking, cognitive load inference, and LLM-based generation under a unified decision policy yields larger learning gains than any of the components in isolation. The composite PEI metric proved sensitive enough to discriminate between conditions with very large effect sizes, suggesting it may serve as a useful target for offline policy optimization in future work. The maintenance of cognitive load within a desirable difficulty corridor appears to be a more proximate driver of learning than mastery tracking alone, supporting CLT's emphasis on instantaneous processing demands.

7.2 Practical Implications

The behavioral-proxy approach to cognitive load estimation has substantial practical implications. Special hardware—eye trackers, EEG, wearables—has been the limiting reagent in deploying cognitive-aware tutoring systems beyond laboratory contexts. By demonstrating that interaction-log features correlate adequately with self-reported load for closed-loop control, we argue that cognitive-aware adaptation can be deployed wherever a browser can be deployed. The federated calibration protocol we propose in Section 7.4 addresses the population-specificity of the estimator.

7.3 The LLM as a Constrained Generator, Not a Tutor

A persistent design tension in our framework concerns the role of the LLM. Treating the LLM as the tutor itself, and surrounding it with a thin user interface, produces what we term the conversational tutor pattern—the dominant pattern in current commercial deployments. Treating the LLM as one tool among several, governed by a policy that decides when and how to invoke it, produces the orchestrated pattern that ADAPT-CL exemplifies. Our pilot suggests the orchestrated pattern delivers better learning outcomes, although we acknowledge that the conversational pattern has advantages in user satisfaction and ease of implementation that deserve formal study. The design space between these poles, in which LLMs negotiate authority with explicit policy modules, is rich and largely unexplored.

7.4 Privacy and Federation

Behavioral telemetry is intimate data. Keystroke timing, scroll patterns, and error trajectories can, in aggregate, reveal far more about a learner than test scores alone—including, plausibly, indicators of attentional disorders, emotional state, and disability status that the learner has not disclosed. We propose that production deployments of ADAPT-CL adopt a federated learning protocol in which cognitive load estimators are trained locally on each institution's data, with only model updates (not raw telemetry) transmitted to a central aggregator. We sketch a FedAvg-based protocol with differential privacy noise injected at the local-update step, leaving full implementation and evaluation to future work.

7.5 Limitations

Several limitations constrain the strength of conclusions drawn from this study. First, the sample size ($n = 46$ retained) is small and drawn from a single institution; generalization to other populations, age groups, and educational systems requires replication. Second, the pilot was a single 75-minute session, leaving open whether benefits accumulate, plateau, or decay over multi-week deployments. Third, the cognitive load estimator was calibrated on a separate but demographically similar cohort; performance in genuinely novel populations is unknown. Fourth, our LLM costs were not assessed against pedagogical outcomes; the cost-effectiveness frontier of LLM-augmented tutoring remains an open empirical question. Fifth, we did not measure transfer to genuinely novel problems beyond the immediate domain—the deepest test of robust learning. Sixth, ethical questions surrounding the explicit modeling of frustration in minors deserve more thorough treatment than we provide here.

8. FUTURE WORK

Several directions appear most pressing. A multi-institution replication with at least 400 participants across at least three educational contexts would substantially strengthen the empirical base. A longitudinal study tracking learners across an entire course term would address whether benefits persist or attenuate. The integration of multimodal signals where ethically deployed—webcam-based attention estimation in informed-consent contexts, for example—could improve cognitive load estimation. Replacing BKT with deep knowledge tracing while preserving interpretability through post-hoc explanation methods is a natural next step. Finally, extending ADAPT-CL beyond well-structured domains like data structures into ill-structured domains like writing, design, and ethical reasoning will test whether the framework generalizes beyond settings with crisp knowledge components.

9. CONCLUSION

AI-based personalized learning has long suffered from a tension between the engineering elegance of its components and the pedagogical effectiveness of their integration. Knowledge tracing, content recommendation, and generative tutoring have each produced strong individual results yet have rarely been combined in ways that explicitly model the cognitive state of the learner moment to moment. ADAPT-CL proposes one such

integration, anchored on the observation that interaction telemetry alone, suitably engineered, supplies a workable signal of cognitive load without specialized hardware. A pilot study supports the framework's viability and suggests effect sizes large enough to motivate substantial replication. We do not claim that ADAPT-CL is the final word on personalized learning—our limitations are numerous and our population narrow—but we believe the design principles it articulates, namely that personalization is a joint optimization problem over mastery, cognitive load, and generative content, will outlast the particular implementation choices reported here. The two-sigma problem has resisted technological solution for forty years; it will not be solved by any single component, however impressive. It will be solved, if at all, by the careful orchestration of many.

ACKNOWLEDGEMENT

The authors thank the students who participated in the pilot study and the teaching assistants who reviewed and validated the LLM-generated content. The authors also acknowledge the guidance and support provided by the faculty of the Department of Information Technology, Noida Institute of Engineering and Technology.

The authors declare no competing interests.

REFERENCES

1. Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. In R. E. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 522–560). Routledge.
2. Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2), 167–207.
3. Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2008). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4), 223–241.
4. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, Ö., & Mariman, R. (2024). Generative AI can harm learning. SSRN Working Paper 4895486.
5. 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.
6. 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.
7. Clement, B., Roy, D., Oudeyer, P. Y., & Lopes, M. (2015). Multi-armed bandits for intelligent tutoring systems. *Journal of Educational Data Mining*, 7(2), 20–48.

8. Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278.
9. Ghosh, A., Heffernan, N., & Lan, A. S. (2020). Context-aware attentive knowledge tracing. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2330–2339.
10. Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52, 139–183.
11. Kestin, G., Miller, K., Klaes, A., Milbourne, T., & Ponti, G. (2024). AI tutoring outperforms active learning. *Research Square preprint*.
12. Khan, S. (2024). Brave new words: How AI will revolutionize education (and why that's a good thing). *Viking*.
13. Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42–78.
14. Mitrovic, A. (2012). Fifteen years of constraint-based tutors: What we have achieved and where we are going. *User Modeling and User-Adapted Interaction*, 22(1–2), 39–72.
15. Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050.