Journal of Computer Science

Research Article

Integration of Local Large Language Models, Retrieval-Augmented Generation, and Adaptive Learning

Anass Belcaid and Kamal Reklaoui

Department of Artificial Intelligence and Digitalisation, National School Applied Sciences, Tetouan, Morocco

Article history
Received: 21-05-2025
Revised: 08-08-2025
Accepted: 12-02-2026

Corresponding Author:
Anass Belcaid
Department of Artificial Intelligence and Digitaliation
National School Applied Sciences, Tetouan, Morocco
Email: a.belcaid@uae.ac.ma

Abstract: In recent years, many schools and teachers have started using closed Large Language Models (LLMs) to help with learning. These tools can be very helpful for tutoring and personal learning, but they also bring serious problems. One big issue is that they use cloud systems, which means student data is sent to outside servers. This can put privacy at risk and takes away control from students and teachers over how data is used. Also, closed LLMs often use the same method for every student. They don’t adapt to different learning styles, speeds, or needs. Because of this, many students may feel left out or unsupported especially those who need extra help or a more personal approach. In this paper, we present a novel solution that addresses these challenges by combining a local LLM with Retrieval-Augmented Generation (RAG) and adaptive learning. Our system runs entirely on the user's device, ensuring that all student data remains private and under local control eliminating reliance on external servers. RAG enhances response accuracy by retrieving relevant educational content, enabling clear explanations and context-aware questioning. To personalize learning, the system dynamically adjusts content difficulty and style based on real-time student performance, tracked using Bayesian Knowledge Tracing (BKT). We implemented our approach as a Moodle plugin, integrating it seamlessly into online learning platforms such as MOOCs. Results from a pilot study show that our system increases student success rates by +15 (from 65 to 80%), reduces response time by 20, and boosts daily student interactions by 60%. Qualitative feedback also indicates high student satisfaction and positive instructor evaluations. These improvements reflect not only technical performance but also a deeper commitment to aligning AI with the core values of education privacy, equity, and learner agency. By grounding AI support in local control and adaptive personalization, we aim to build a fairer, flexible, and trustworthy approach to educational technology, where innovation serves both pedagogical effectiveness and human dignity.

Keywords: Artificial Intelligence, Local LLMs, RAG, Adaptive Learning, Moodle, Data Privacy

Introduction

In the fast changing world of Artificial Intelligence (AI), Large Language Models (LLM) have emerged as an omnipresent tool that could deliver human-like texts and also have deep conversations with humans. And with their inclusion in the education systems, several worries and problems have been raised about their privacy. Also, an important question is how well could these models adapt to a student's needs and whether the learning experience is truly personal. While Cloud-based LLMs are useful and could handle a very large number of users simultaneously, they also come up with big risks. As an example, student data is considered sensitive and might not be safe. Also, users have to depend on external servers outside of their universities. To solve these problems, many researchers started looking for alternative solutions to use local LLM-models that run directly on campus and ensure data protection and privacy for better security and a more flexible learning experience (Bender et al., 2021; Bommasani et al., 2022).

SCIENCE Publications

Local AI models in education work best when combined with a system called Retrieval Augmented Generation (RAG). RAG mixes searching for facts with creating new text. RAG works well in education because it gives more correct answers by checking facts in trusted sources, can use a teacher's own materials, and keeps student information private by running on local computers (Lewis et al., 2021). When set up locally with RAG, an AI helper can explain things based on how much a student already knows, ask questions that match what the student understands, and get better at helping as it learns from student feedback.

Given the critical aspect of adaptability in addressing differentiation which is one of the most persistent challenges in education and the observation that most of the traditional classroom configurations have several problems to meet the diverse needs from students from the that are overwhelmed by the complexity of the content or the under-challenged which feels bored by the oversimplification (Papernot et al., 2018). These problems could be solved by using a local LLM acting as personal tutor. A local LLM has the potential to bridge the gap between the student and content by creating adequate learning pathways. Based on the student input, it can analyze their proficiency level and target questions adjustable to his proficiency level. Furthermore, Given the locality of the model, it has all the advantages of avoiding the pitfalls of transmitting personal information to a given remote system which is the heart of the privacy problem.

This problem was investigated in Floridi et al. (2018) who discuss the student concerns about their data doing to the cloud and the ethics of their exploitation.

Local LLMs with RAG do more than protect privacy and adjust to users. They also help make good education available to everyone. When these systems work without internet or expensive cloud services, advanced tutoring becomes possible even in places with few resources. This supports efforts to give everyone fair access to learning with technology (Kucirkova, 2017). Also, because these models work locally, teachers can add knowledge that matters to their culture and teaching goals in specific places.

State of the Art

In this section, we will review the current state of art of the landscape of advancements in education that form the foundation of our approach. First, we will examine the limitations of traditional e-learning architectures, which struggle to provide customized and personalized content for the learner. These systems rely on static content and predefined paths which make them lack the flexibility in delivering adaptive content that meet the student needs. Following this, we will explore the recent advances in AI and we will highlight how AI-tools like LLM and RAG could transform the student experience in interacting with an online tutor. By understanding these changes, we will highlight the opportunities and challenges that are available with our solution which aims to bridge the gap between delivering personalized and adaptive content while keeping and respecting the student privacy and data within the e-learning ecosystem.

Limitations of Traditional E-Learning Approaches

While e-learning systems offer an outstanding solution with the possibility to deliver the courses to a very large number of students, they suffer from a significant drawback which is their reliance on static content. More precisely, these systems deliver a pre-designed and static content that remains unchanged regardless of the learner's progress or needs. Consequently, this limitation to adapt to different learning styles, paces and level of understanding can reduce the efficiency of the learning experience. As an example, a student could struggle with a very fundamental concept and may find himself overwhelmed, while another has already mastered the basics and may feel disengaged due to the lack of challenge. Hence, this absence of dynamic content could reduce the effectiveness of the learning experience and will fail to foster critical thinking and deep engagement for the student. This was noted by Kucirkova et al. (2017) which presented how traditional e-learning systems often fail to address this diversity in the student need which led to suboptimal educational outcome.

Recent Advances in AI for Education

Recent advances in language models, like BERT (Devlin et al., 2019) have led to transformation in Human-Computer Interaction for learning purposes. Employing these models is the understanding itself of subtle nuances in language, thereby producing explanations, answering questions, and even carrying out dialog-based tutoring with great fluency. Integration of Retrieval-Augmented Generation (RAG), which refers to a combination of text generation and information retrieval, will ensure that educational responses are further enriched by tying their outputs to credible and relevant knowledge. Progress in adaptive learning systems has also significantly advanced. Bayesian Knowledge Tracking (BKT) and Item Response Theory (IRT) are two such approaches. BKT, coined by Koedinger et al. (2012) indicates the mastery of a student on a specific skill in time by looking into performance on various tasks so he/she can be given remediation or advancement by the system. IRT (Lee et al., 2010), on the other hand, deals with statistics to determine the difficulty of the question and the ability of the individual, ensuring that assessment is going along with individual proficiency level. Reinforcement learning-based tutoring systems (Ruan and Lu, 2025) that are another step ahead in adaptive learning are those that dynamically adapt the content according to real-time feedback and, thus, actively construct personalized learning paths while optimizing engagement (Singla et al., 2021). All of these advancements stem from the increasing potential that AI holds in addressing some of the key issues in education. These challenge-to-opportunity transformations cover the landscape of personalization all the way to adaptation, significantly marking their way into more effective and inclusive learning contexts.

Finally, many classical e-learning systems rely on cloud-based solutions to deliver content and manage interactions. While cloud-based infrastructure offers advantages such as scalability and accessibility, it also introduces significant constraints. One major concern is the reliance on internet connectivity, which can be a barrier in regions with limited or unreliable access. More critically, cloud-based systems often require the transmission of sensitive student data to external servers, raising serious privacy and security concerns. These risks are particularly problematic in educational contexts, where protecting student information is paramount. Floridi et al. (2018) emphasize the ethical challenges posed by centralized data storage in AI-driven systems calling for decentralized alternatives that prioritize user privacy. Although cloud solutions have enabled widespread adoption of e-learning, their limitations highlight the need for alternative approaches that prioritize data privacy, adaptability, and offline functionality without compromising the quality of the learning experience.

Ultimately, many classical e-learning systems are based on cloud-based architectures for content delivery and interaction management. While these cloud infrastructures proffer advantages such as scalability and accessibility, they come with enormous challenges. One such consideration is dependence on internet connectivity, which can become a barrier if access is limited or unreliable in certain locations. Even more importantly, in many cases, the cloud-based systems can lead to sensitive student data being shipped to external servers, threatening serious privacy and security. These aforementioned threats are especially serious in the educational domain, wherein protection of student data is paramount. Floridi et al. (2018) emphasized the ethical challenges posed by centralized data storage in AI-driven systems and call for decentralized alternatives with user privacy in mind. While cloud solutions have facilitated the massive uptake of e-learning, their limitations stress the need for alternatives that protect the privacy of data, ease adaptability, and allow offline use without compromising the quality of the learning experience.

Proposed Technological Approaches

The suggested solution, however, connects three essential components to create a secure, personalized, and adaptive learning experience. A local LLM with a reduced number of parameters is at its center and is meant to be lightweight and customizable for educational purposes. The local deployment ensures confidentiality and freedom from dependence on external servers, while the lesser parameter count enables its running on personal devices without compromising performance. To further enhance the model's ability to provide precise and context-sensitive answers, we introduce Retrieval-Augmented Generation (RAG), which employs the generative properties of the LLM alongside rigorous information retrieval from educational datasets. The final component is an adaptive learning system that employs state-of-the-art methods, including BKT, and IRT, and Reinforcement Learning (RL). These components together will allow a system to adapt content, measure proficiency, and optimize learning paths for each learner in real time, guaranteeing an effective educational experience for every learner.

Deployment of Local LLMs

The deployment of local large language models is a significant advancement for ethical, secure, and personalized AI-based tools in education. Top open-source models such as **(**DeepSeek et al., 2024; Bai et al., 2023; Mistral-AI et al., 2025) obtainable through resources like Ollama, impart unique advantages that make them good candidates for the fine-tuning of individual learning environments. One significant advantage of these models is their open form through which educators and developers can observe, adapt, and perfect the models to be used for educational goals. In such a scenario, openness transforms not only into a system of transparency but also empowerment for users to modify models into forms that can suit further varieties such as curricular innovations, languages, or cultural adaptations, making the content relevant and inclusive. More so, these local LLMs can still be customized by feeding them additional domain-specific knowledge bases for more accurate, context-aware information that has been adjusted for education-oriented use cases. Flexibility in scaling is another major advantage for these LLMs. The models are often offered in different sizes-from small to light forms paralleled with low-resource systems, all the way to large, very expensive forms used when applications are very advanced. For example, Deepseek has tight compact configurations power-optimized for efficiency, while Mistral provides configurations balancing performance with resources utilization requirements. All this alone ensures that it can be adopted on a wide range of different hardware configurations, thus rendering it accessible even in resource-limited environments. Among other benefits of using local LLMs such as Deepseek, Qwen, or Mistral, the major benefit is for the possibility of developing the whole learning process in a safe, private, and, above all highly adaptable way, focusing on student privacy with teaching effectiveness (Khezresmaeilzadeh et al., 2025).

Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG), endowed with the potential of retrieval and generative models, is breaking new ground for avenues of education. According to Lewis et al. (2020) RAG fuses the best of two worlds: Information retrieval systems that pull data with precision and generative language models that articulate it with fluency and creativity. Fundamentally, RAG retrieves documents or knowledge snippets from a well-curated database or corpus first, and then conditions the generation on the information retrieved. This two-phase scheme ensures that the response is not only contextually valid but also based on verifiable knowledge. This bypasses one of the major limitations of generative models that can sometimes produce incorrect outputs even if they sound plausible. For educational purposes, RAG can be implemented to provide students with precise explanations, examples, and references from an accredited educational resource, thus ensuring that the interactions made are of good quality and trustworthiness. Being modular, there is also room for educators to tailor the knowledge base that comes in RAG, which means the chosen knowledge can be aligned with specific subjects, curricula, or student needs. In this way, grounding outputs in authoritative sources while still giving flexibility to the other generative aspects of AI makes RAG a perfect blend for the generation of accurate and constructive education content, and for that reason, a crucial ingredient in the solution we propose.

In Figure 1, we see that the system starts with the user query, which is later fed into the query encoder. The retriever then proceeds to gather relevant information from the document index, which is further used by the generator to generate the final response. The modular nature of the architecture makes it both accurate and flexible for educational purposes.

Adaptive Learning Systems

Adaptive learning is an educational approach that uses technology to tailor content and feedback according to the individual needs, performance, and learning pace of each student. Unlike the standard practice of one-size-fits-all, adaptive learning systems track the learner's interactions, including responses and time taken on tasks or assessments, determine the real-time level of understanding, and use the data collected to tailor the material presented to that learner. Custom explanations, appropriate level question sets, and just the right amount of work necessary to stretch a learner's abilities while not overwhelming them are all provided at the precise time in the learning experience when the learner needs it. The intention is to create a more effective, convective, and student-centered learning culture through meeting students where they are and giving appropriate support at the right time for an ongoing improvement process. We will look at ways of making systems intelligent enough to adjust and respond continuously to the needs of the student.

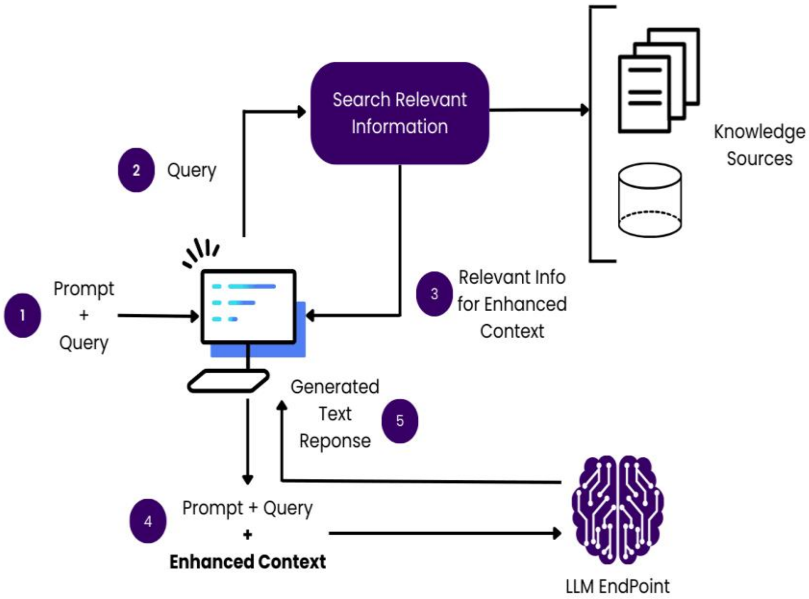
Search Relevant Information
Knowledge Sources
2 Query
1 Prompt + Query
3 Relevant Info for Enhanced Context
Generated Text Reponse 5
4 Prompt + Query + Enhanced Context
LLM EndPoint

Fig. 1: Architecture of the RAG system with a local LLM

Bayesian Knowledge Tracing (BKT) is a widely used statistical framework for modeling a learner’s mastery of specific skills over time. Introduced by Koedinger et al. (2012) BKT evaluates key parameters such as the initial probability of mastery, the probability of transitioning from non-mastery to mastery (learning rate), the likelihood of making an error despite mastery (slip), and the probability of guessing correctly without mastery (guess). By analyzing these parameters, BKT will update the probability that a learner has mastered a given skill based on their responses to tasks or exercises. Each interaction adds new evidence, which allows the model to update its estimation of the learner’s proficiency. This iterative process enables the system to tailor subsequent exercises or instructional content to the learner’s current level of understanding, ensuring that the educational experience remains both challenging and supportive.

In our system, the BKT-estimated mastery probabilities are used to modulate the prompting strategy of the LLM. Specifically, low mastery levels trigger retrieval of foundational concepts via RAG and prompt the LLM to generate more explanatory, step-by-step responses. Conversely, high mastery levels lead to more concise or advanced explanations. This adaptive control allows the LLM to deliver personalized educational content aligned with the student's evolving knowledge state.

While traditional BKT uses skill-level parameters, prior work has explored individualizing parameters to better capture student differences. The Rasch model (1PL IRT) and Additive Factors Model (AFM) incorporate student-specific proficiency terms to improve prediction and interpretability (Corbett and Anderson, 1995) and similar benefits have been observed in educational modeling.

Early attempts to individualize BKT include Corbett and Anderson (1995) who proposed fitting both student- and skill-specific parameters, combined via a merging function. While this improved alignment between predicted and actual accuracy across students, it did not significantly enhance test score prediction. Pardos and Heffernan (2010) later individualized only the initial mastery probability $p(L_0)_k$ using heuristics (e.g., first response correctness, overall performance), showing improved model fit on several datasets. Yudelsen et al. (2013) extended this by learning all four BKT parameters per student, focusing on personalized practice scheduling rather than prediction accuracy.

Despite these efforts, the benefits of individualized BKT remain mixed: Improvements in predictive power are inconsistent, optimal parameter configurations are unclear, and practical implementation in real-world ITSs is challenging. In this work, we adopt a skill-level BKT model with fixed parameters per skill, trained via standard methods (e.g. EM (Corbett and Anderson, 1995)), to ensure robustness and scalability while still enabling adaptive LLM behavior through mastery-driven prompting.

The model is defined by three core components: A prior belief over the initial state, a transition model capturing learning, and an observation model linking knowledge state to response accuracy. These can be expressed in matrix form as follows (Table 1).

In summary, Bayesian Knowledge Tracing provides a principled and interpretable framework for modeling student knowledge progression at the skill level. Despite ongoing exploration of individualized variants, the standard BKT model remains widely adopted due to its robustness, scalability, and strong empirical performance when properly calibrated. In our system, BKT serves as the main scoring tool for tracking student mastery over time. The estimated knowledge states computed using the well-defined probabilistic update rules and matrices described above, are directly integrated into our adaptive learning pipeline to inform both retrieval strategies in RAG and prompting behavior of the LLM. This tight coupling ensures that educational interventions are dynamically tailored to the learner’s evolving proficiency, enabling personalized and timely support.

Item Response Theory

Item Response Theory (IRT) is a general purpose framework which tries to model the interaction of a student's latent ability and the characteristics of an item in the test. It differs from traditional test scoring methods in that it does not assume that every question is equally informative. Rather, through IRT, one can estimate the probability of a learner correctly answering a particular item given his or her underlying skill level. That makes IRT especially suited for adaptive learning systems, where one aims to adjust the difficulty of learning tasks to the learners' proficiency.

There are mainly a few IRT models that are widely used in practical situations, and these differ in the level of complexity and number of parameters considered. The first, in 1PL, or one-parameter logistic models, referred to as the Rasch model, states that the probability of giving the correct answer is dependent only on the difference between the ability of the learner and the difficulty of the item.

Table 1: BKT Model Parameters in Matrix Form

Bayesian Knowledge Tracing Metrics		
Priors	**Transitions (A)**	**Observations (B)**
Known $p(L_0)$	[1, 0]	$1 - p(S), p(S)]$
Unknown $1 - p(L_0)$	$[p(T), 1 - p(T)]$	$[p(G), 1 - P(G)]$

The second two-parameter logistic model extends the model to allow for item discrimination, which reflects the capacity of an item to discriminate between students of varying ability levels Thirdly, the 3-parameter logistic model considers random guessing among its parameters. This set of models thus enables adaptive learning systems to select exercises with dynamic difficulty levels that would otherwise be optimally engaging or promote learning (Kabudi et al., 2021).

Reinforcement Learning

The Reinforcement Learning (RL) paradigm offers a promising methodology for adaptive learning systems where the arrangement of exercises is improved according to learner performance. Algorithms like Q-learning function by establishing a reward function that is indicative of progress-such as an improvement in task accuracy or speed-and then selecting the activity sequence that maximizes that reward (Ruan and Lu, 2025). Thus the algorithm learns continuously from the learner-interactions tremendously customizing the difficulty level or types of exercises for optimal challenge. For example, should a student perform very well in a certain task, the system might enforce more advanced problems, on the other hand, if such struggles are repeated, then more support or easier problems will be enforced. RL keeps learners motivated and progressing with their own personalized learning path by adapting dynamically to individual performance-a sure win for personalized learning.

Deep Learning-Based Recommendation Systems

Deep learning recommender systems have emerged as projections in cross-breed or hybrid adaptive environments supported by deep learning algorithms and convert the educational content to personalized learning experiences over time. These systems couple various models like auto encoders and neural networks to first analyze the content feature vectors of students and past consumption records in their recommendations of learning materials. As a result, such systems become capable of tracking how learning behavior, or learner progress, aligns to patterns in the recommendations of certain resources based on similar content defining mention or alignment, ultimately improving engagement and learning outcomes (Du Plooy et al., 2024).

System Architecture

This segment presents the architecture and design of the proposed system to interact with parts for provision of secure, adaptive, and personalized learning experiences. The structure is based on amalgamation of Local Large Language Model (LLM), Retrieval-Augmented Generation (RAG), and some advanced adaptive learning algorithms, such as Bayesian Knowledge Tracing (BKT) or Reinforcement Learning (RL). The following subsections will describe the architecture of the system in detail, the data flow and how modules interact and mechanisms regarding privacy, adaptation, and effectiveness in terms of education. This breakdown would present a clear overview of how the system's goals are achieved and obstacles discussed previously are tackled.

The architecture, as illustrated in Figure 2, orchestrates multiple components to deliver a secure, adaptive, and context-aware learning experience.

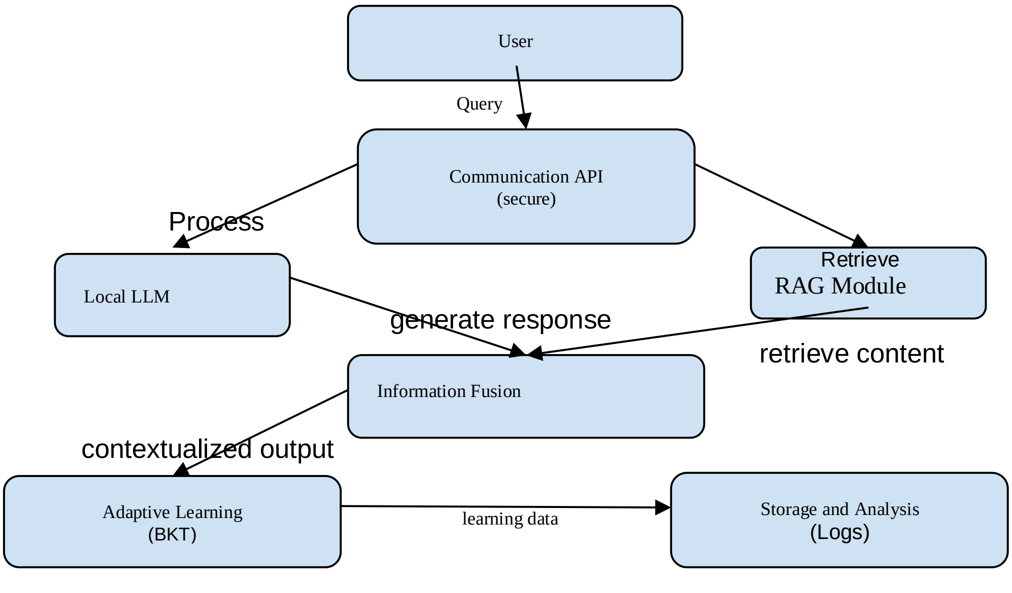

Fig. 2: Secure Adaptive Learning Framework architecture

The process begins when a user submits a query through the secure Communication API, which ensures encrypted transmission and local handling of all data. This input is then routed in parallel to two core modules: The Local LLM, which generates fluent, semantically rich responses based on its internal knowledge, and the RAG Module, which retrieves factual, up-to-date content from a curated knowledge base (e.g., textbooks, technical guides). By operating both modules simultaneously, the system combines the generative strength of LLMs with the factual reliability of retrieval-based methods, mitigating hallucinations and improving response accuracy.

The integration of these components occurs in the Information Fusion stage, where the generated response and retrieved context are combined and contextualized into a coherent, learner-appropriate output. This step involves aligning tone, difficulty, and domain specificity to ensure the final response is both accurate and pedagogically sound. The fused output is then passed to the Adaptive Learning Module, which uses BKT to assess the learner's current proficiency in relevant skills. Based on this assessment, the system dynamically adjusts the complexity, depth, and style of future responses offering step-by-step guidance for struggling learners or advanced challenges for proficient ones. Finally, all interactions and performance data are logged in the Storage and Analysis module, enabling long-term tracking, educator insights, and iterative system improvement.

This modular data flow ensures tight coupling between personalization, security, and educational effectiveness. The secure API acts as a trusted gateway, preventing data leakage and enforcing local execution. The parallel LLM-RAG design enables a balance between creativity and factual grounding, while the fusion mechanism ensures seamless integration. Contextualization is not limited to content it extends to timing, scaffolding, and feedback style, all modulated by BKT-driven insights. Together, these interactions form a responsive loop: Student input shapes retrieval and generation, which informs adaptation, which in turn influences future interactions. This closed-loop architecture enables truly personalized, privacy-preserving AI-assisted learning, where every component contributes to a cohesive, learner-centered experience.

Local LLM: Model Usage and Deployment Strategy

The system leverages the open-source DeepSeek deepseek-v3:671b model as the backbone of the LLM component. This model was selected for its strong reasoning performance, support for long-context inputs (up to 128k tokens), and compatibility with retrieval-augmented workflows. In our current implementation, the LLM is used in a zero-shot, inference-only mode and has not been fine-tuned or retrained. Instead, domain-specific accuracy and personalization are achieved through the RAG module, which retrieves relevant educational content and injects it into the prompt context.

The RAG system employs a long-chain retrieval approach, where multiple context passages (e.g., lecture notes, textbook sections) are retrieved and concatenated into an extended input prompt. This enables the LLM to generate accurate, grounded responses without requiring model updates.

During the current testing and evaluation phase, all system components including the LLM, RAG module, adaptive engine, and Moodle plugin run on a centralized backend server. This setup simplifies development, monitoring, and integration testing. Table 2 summarizes the hardware specifications of the server used for all experiments.

In future work, we plan to transition to a decentralized deployment model: The LLM and RAG system will be hosted on a private, on-premise server within the school network or run directly on student devices. The Moodle plugin will communicate securely with this local instance via a lightweight API, ensuring data remains within institutional boundaries while preserving real-time interactivity.

Pilot Study and Experiment Evaluation

To determine how effective, the proposed system is for real-life education, it has organized a pilot testing in one course under *Industrial Maintenance & Operational Safety* for Supply Chain Management students. The course selected is of a technical nature, while the student body is characterized by diverse learning needs across the board, making it an ideal ground for testing the adaptability and personalization features of the system-. This pilot study also serves to discover the effects that a combination of a local LLM with Retrieval-Augmented Generation-as well as adaptive learning techniques-would have on engaging, understanding, and performance outcomes within a local learning environment for students.

Table 2: Server Specification for Testing Environment

Component	Specification
CPU	AMD EPYC 7543P (32 cores, 64 threads)
GPU	2 NVIDIA A6000 Ada (48 GB VRAM each)
RAM	256 GB DDR5 ECC
Storage	2 TB NVMe SSD (RAID 1)
Model Quantization	GGUF Q4_K_M (4-bit)
Inference Backend	llama.cpp + custom RAG pipeline
Network	1 Gbps internal, TLS-secured API
Operating System	Ubuntu 22.04 LTS

The results would be announced through both qualitative feedback and quantitative metrics to know if the intended system would overcome the barriers of traditional e-learning and, at the same time, provide a safe entirely personalized and interactive learning experience. The next sections present the methodology, key findings, and implications of this pilot study.

Context and Implementation

This course will teach complex and fundamental concepts such as preventive maintenance, reliability analysis, and operational safety, indispensable for efficient supply chain management. These concepts demand a high level of theoretical and practical understanding from students, especially those with various degrees of prior knowledge. Beyond technical details, they will also have to correlate these concepts with real-life situations in supply chain operations because of the interdisciplinary nature of the course. Learning adaptive should then be considered a must to help students go through the material at their own pace, as well as to develop a higher degree of critical-thinking and problem-solving skills.

Materials

All experiments were conducted on a dedicated high-performance computing workstation designed for large-scale model inference and Retrieval-Augmented Generation (RAG) workloads. The system was equipped with an AMD EPYC 7543P processor featuring 32 physical cores and 64 threads, providing strong parallel performance for preprocessing, retrieval, and CPU-based inference tasks. The machine was configured with 256 GB of DDR5 ECC memory, ensuring both high bandwidth and reliability during extended experimental runs.

For accelerated inference, the platform incorporated two NVIDIA A6000 Ada GPUs, each with 48 GB of VRAM, enabling efficient execution of large language models and GPU-accelerated components of the RAG pipeline. Storage was provided by 2 TB of NVMe SSDs configured in RAID 1, offering both high I/O throughput and fault tolerance. The operating system was Ubuntu 22.04 LTS, selected for its stability and compatibility with modern machine learning frameworks.

Model inference relied on llama.cpp combined with a custom retrieval-augmented generation pipeline. To reduce memory footprint while preserving inference quality, models were quantized using the GGUF Q4_K_M (4-bit) format. All components communicated over a TLS-secured internal API on a 1 Gbps network, ensuring low-latency and secure data exchange throughout the experimental workflow.

Methodology and Metrics

As part of the current pilot study, data on student interactions with the proposed system were comprehensively collected. Metrics such as task completion success, response time to questions, and navigation paths within the platform were recorded, providing rich insights into student engagement and the quality of learning experiences. For instance, success rates highlighted areas where students struggled or excelled, response times reflected cognitive load and task difficulty, and navigation patterns revealed how learners engaged with different modules and utilized personalized recommendations.

To support the RAG component, we curated a domain-specific knowledge base consisting of a structured collection of PDFs in two key thematic areas: *Algorithms: Design and Analysis* and *Predictive Maintenance*. These materials, drawn from academic textbooks, technical manuals, and lecture notes, were preprocessed and indexed to enable accurate, context-aware retrieval during student interactions. By integrating this knowledge base with the local LLM, the system generates responses grounded in reliable educational content.

Aggregating and analyzing these interaction logs allowed us to evaluate the system’s adaptive capabilities and its effectiveness in personalizing support to individual learners, ultimately contributing to improved learning outcomes.

Assessements

Both formative assessments and satisfaction questionnaires were incorporated into the evaluation framework as an attempt to ensure that the entire system's impact is fully covered. Formative assessments have been given from time to time throughout the course to check students' understanding in key areas such as preventive maintenance or reliability analysis. The tests provided immediate feedback to learners and gave instructors the information needed to monitor progress and adjust instruction as appropriate. During and after the pilot, satisfaction questionnaires were also circulated to elicit qualitative insights from participants. The satisfaction survey included open-ended questions on issues such as the usability of the system, clarity of explanation by the local LLM, and perceived usefulness of adaptive features. The performance metrics complemented with feedback provided a balanced evaluation of the system's strengths and weaknesses.

Metrics

The evaluation of the system was guided by three primary dimensions: Engagement, academic performance, and satisfaction. Engagement was quantified by counting the interactions of the students with the system and by the time they spent in active sessions. High engagement was maintained, signifying that interest was sustained, personalization was effective, and the system captured the students' attention well. Academic performance was judged by improvement in success rates on quizzes and assignments against reduced response time, which reflected the increasing familiarity and confidence of the students with the material. Lastly, satisfaction was assessed with qualitative feedback solicited from both students and instructors. Perceived usefulness of the system in catering for varying learning needs and for nurturing a supportive educational ambience was indicated in the feedback. These metrics, together, provided reasonable grounds to gauge overall effectiveness of the system with prospects for scalability in various educational contexts.

Results and Analysis

A summary of findings is presented in Table 3 that shows the outcomes from the pilot study, which included improvements in performance, response times, engagement, and qualitative feedback. The table can give a clear view of the quantitative metrics with their specific improvement.

The results presented in Table 3 demonstrate significant improvements across multiple dimensions, underscoring the effectiveness of the proposed system. One of the most notable achievements was the 15% increase in success rates, rising from 65 to 80%. This improvement indicates that the adaptive module successfully aligned content with individual learning needs, enabling students to grasp complex concepts more effectively. Similarly, the reduction in average response times by 20, from 45 seconds to 36 seconds, reflects faster assimilation of material. These gains suggest that the system not only enhanced comprehension but also streamlined the learning process, allowing students to engage with the material more efficiently.

These numbers are complemented by qualitative feedback strengthening the effect of this system. Students were impressed by the interactive interface and the quality of the responses given by the local LLM. Emphasis was placed on clear and contextually related explanations. Also, instructors recorded positive remarks on their side, stressing students were evidently better prepared for practical sessions vital for the mastery of industrial maintenance and operational safety concepts. The 60% increase in daily interactions-from 5 to 8 interactions a day per student-further pointed toward increased engagement raised by the system.

Table 3: Summary of the Key results for the pilot study

Metric	Baseline	Improvement
Success Rate	65%	+15%
Response Time	45 seconds	-20%
Daily interaction	5 per student	+60%
Student Satisfaction	Moderate	High
Instructor Feedback	Neutral	Positive

All these findings are in agreement that the integration of adaptive learning techniques and retrieval-augmented generation did not just enhance the academic performance of students but also create a more engaging and supportive educational environment.

Discussion

Future work will focus on three directions:

(1) Transitioning to true edge deployment, where the LLM and RAG run locally on student devices or school servers
(2) Optimizing model efficiency through quantization, distillation, and caching strategies to improve response times
(3) Incorporating fairness-aware components to detect and correct for content bias, ensuring equitable support across diverse learner profiles

Foundations for the next generation of AI-powered education have been laid. By systematically addressing these challenges, we move closer to a future that is not only more intelligent and efficient but also more just, inclusive, and transformative.

Moodle Integration: Development of a Custom Block

A custom block has been created to enable the streamlined access of the LLM functionalities into the Moodle environment. Simply put, it is now possible for students to ask questions and receive real-time-generate responses like a "teacher in teaching methodology intervention". Integration into Moodle was made by developing a custom plugin integrated with the fine-grained functionality of local LLM along with RAG specifically tailored to the interface. The plugin not only maintains secure communication between Moodle and local LLM but also improves the usability for student and instructor access. We will provide the PHP code that can be used to technically demonstrate this implementation, including the block configuration, API request handling, and response display process within the Moodle environment. Such a procedure will detail possible steps that could provide a clear route map for any interested educator or developer in deploying a similar adaptive system within courses using Moodle.

Plugin Structure

The architecture provides a clear and modular architecture for the integration of its local LLM with all its functionalities into the platform. In this case, such an integration plugin is made of several components: The block interface through which students submit queries; the backend PHP scripts which take care of API requests and responses; and the configuration files, which make the whole work seamlessly fuss with Moodle's framework. To gain an insight into the inner workings of the plugin, we will first showcase the PHP codes (presented in Figure 3) that make these components, demonstrating how the data flows from the user interface into the local LLM and down to the Retrieval-Augmented Generation module. Readers will understand how this is enabled by looking closely at the structure and functioning of these blocks: Technical design choices that enable safe, real-time connections while still compatible with Moodle's extensible architecture.

The following PHP code defines the core functionality of the Moodle block that integrates the local LLM into the platform. This file, *block_llm.php*, implements the block's initialization, content generation, and interaction logic.

Listing 1: Code for the php block_llm

```
<?php
// This file is part of Moodle - http://moodle.org/
//
// Moodle is free software: you can redistribute it and/or modify
// it under the terms of the GNU General Public License.
// (c) Your Name or Institution

defined('MOODLE_INTERNAL') || die();

class block_llm extends block_base {

 // Block initialization.
 public function init() {
 $this->title = get_string('pluginname', 'block_llm');
 }

 // Generates the block content.
 public function get_content() {
 if ($this->content !== null) {
 return $this->content;
 }

 $this->content = new stdClass;
 $this->content->text = $this->generate_llm_form();
 $this->contnt->footer = '';
 return $this->content;
 }

 // Generates the input form and displays the response.
 private function generate_llm_form() {
 global $PAGE;

 // Define the form action URL: here, we submit to the same page with an 'action' parameter.
 $actionurl = new moodle_url($PAGE->url, array('action' => 'llm_submit'));

 // Create the HTML form.
 $html = html_writer::start_tag('form', array('method' => 'post', 'action' => $actionurl));
 $html .= html_writer::label('Ask your question:', 'llm_question');
 $html .= html_writer::empty_tag('input', array(
 'type' => 'text',
 'name' => 'llm_question',
 'id' => 'llm_question',
 'size' => '50'
 ));
 $html .= html_writer::empty_tag('br');
 $html .= html_writer::empty_tag('input', array('type' => 'submit', 'value' => 'Submit'));
 $html .= html_writer::end_tag('form');

 // Check if the form was submitted.
 $question = optional_param('llm_question', '', PARAM_TEXT);
 if (!empty($question)) {
 // Simulate an API call to the LLM. Replace this with a real API call.
 $response = $this->get_llm_response($question);
 $html .= html_writer::tag('div', 'Response: ' . s($response), array('style' => 'margin-top:10px;'));
 }

 return $html;
 }

 // Simulates an API call to obtain a response from the LLM.
 private function get_llm_response($question) {
 return "This is a simulated response for the question: " . $question;
 }
}
```

The sample PHP code in Listing 1 defines a Moodle block class named block_llm, which acts as the core of integrating the local LLM into the Moodle environment. The class that extends the block_base class of Moodle overrides the major innards such as: init() and get_content(). So, the init() initializes the block by giving it a heading, while the other generates its contents dynamically.

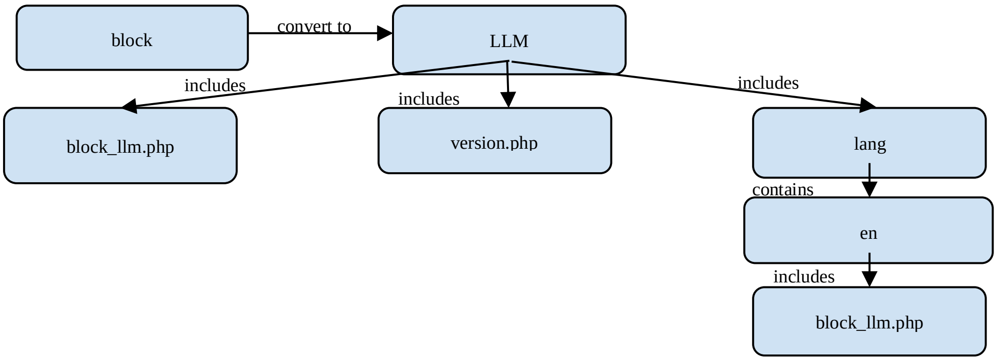

Fig. 3: Structure of the moodle plugin

Its contents contain an interactive form developed by generate_llm_form(), which allows students to submit questions and receive responses from the LLM. The form submission is done by Moodle's internal means, hence secure and seamless communication.

The generated llm_form() method creates a HTML form using Moodle's html_writer, a useful utility that helps in writing valid XHTML markup. On submission, the form captures the question posed by the user and passes it into the method *get_llm_response()*. Here, the response is simulated for demonstration; however, in a production version, this method would communicate with the real LLM API to fetch a response depending on the situation. Furthermore, the input and output were also properly filtered using *optional_param()* and *s()* functions of Moodle to prevent any kinds of possible injections. This modular approach allows best maintenance and a good opportunity for further enhancements such as when integrating sophisticated features like RAG.

Limitations and Future Works

The proposed integrated approach may have the potential for a real transformation in the setting of digital education; however, it has its limitations. These challenges pose important areas for further research and development to ensure the scalability, fairness, and ethical deployment of the project. We discuss the key limitations of the current implementation of our project as well as potential future work. Each limitation will be elaborated upon, including possible ways to overcome these challenges and thus make the solutions more effective.

Hardware Requirements

Indeed, high-end computing necessities are an important shortcoming of local LLMs; they require GPUs or other forms of processing power to run these models efficiently. Such a criterion becomes an impediment to scalability, especially in resource-constrained environments wherein access to high-performance equipment might prove to be limited. Indeed, the deployment of the models locally has its advantages in privacy, as it avoids any reliance on external services, but can become a barrier to successful large-scale adoption due to the computational overhead. Future work would be concerned with resource optimization by means of parallel processing and lightweight modeling.

Real-Time Response Optimization

Real-time generation and data fusion within the RAG module remain a technical challenge. Retrieved information needs to be merged with generated content through sophisticated algorithms that work under time constraints. Delayed response times lead to an interrupted learning experience, which necessitates further optimization of algorithms and their execution using parallel processing schemes. Going forward, there will be greater emphasis on the development of more efficient data fusion algorithms and hardware acceleration to allow for seamless interactivity and responsiveness.

Bias and Fairness in Personalization

While the system incorporates adaptive mechanisms such as BKT, IRT, and RL, skill assessment and content personalization could still be unreasonably biased. Such biases could put students on uneven scales of treatment and violate the intent of providing equitable learning experiences. Improvements in this respect will need the implementation of advanced techniques in bias detection and mitigation. In the future, we will focus on embedding fairness-aware algorithms and performing rigorous evaluations for analysing and correcting any bias that may arise within the system and hence promoting inclusivity and equity.

Data Fusion in RAG

RAG effectiveness interestingly hinges on the quality of data fusion retrieved content and generated information. In fact, if poorly integrated, the entire response will be off-mark or disconnected, a situation that eventually leads to user discontent. Future investigation will work on enhancement of data fusion techniques for improved and precise integration to overcome the limitation. This improvement includes the use of more advanced algorithms for producing better results, with how well retrieved context aligns with generative capabilities of the LLM to improve relevance and accuracy of outputs *system ends up generating.*

Future Direction

To resolve the limitation already detected, in the following work several areas should receive attention:

- Performance Optimization: Reducing latency and increasing the responsiveness of the system through exploring solutions of resource parallelization and hardware optimization
- Enhancing Data Fusion: Develop strategies that integrate the retrieved information more coherently and accurately into LLM output
- Bias Detection and Mitigation: Incorporation of fairness-aware techniques for bias detection and correction to foster an unbiased learning experience
- Scaling and Validation: Expanding the approach to more disciplines and executing full-scale evaluation programs to confirm and upgrade the model
- With these obstacles being tackled, the system proposed will be strengthened and scaled, as well as made as ethically viable a solution as possible to AI-driven education

Conclusion

This article presents an integrated architecture that advances digital education by combining local Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and adaptive learning based on Bayesian Knowledge Tracing (BKT). Together, these components enable a secure, personalized, and interactive learning experience that addresses key limitations of traditional e-learning platforms particularly regarding data privacy, one-size-fits-all content, and lack of real-time support. The integration of a custom plugin into Moodle enhances accessibility while ensuring that student data remains under local control, fostering trust and institutional ownership.

Results from a pilot study demonstrate measurable improvements in student success rates, response efficiency, and engagement, particularly in demanding ftechnical domains such as industrial maintenance and operational safety. These findings suggest strong potential for scalable, AI-enhanced education that is both effective and ethically grounded.

Nonetheless, several limitations remain. First, the current implementation relies on a centralized server for inference, which, while practical for testing, deviates from the ideal of fully on-device execution. Second, running large models like deepseek-v3:671b demands significant computational resources, limiting accessibility on low-end devices. Third, while RAG reduces hallucinations, the system still inherits some biases present in the training data and knowledge base, and currently lacks active bias detection or mitigation mechanisms.

Acknowledgment

The authors would like to express their sincere gratitude to the Laboratory of Innovative Systems Engineering and to the Directorate of the National School of Applied Sciences for granting access to their artificial intelligence computing infrastructure. The availability of this IA server was essential for the training, testing, and evaluation of the models presented in this work.

Funding Information

This research did not receive any funding.

Author’s Contributions

Anass Belcaid: Implemented the retrieval-augmented generation (RAG) system using a locally deployed large language model. Also, he wrote the first part of the paper.

Kamal Reklaoui: Designed and implemented the adaptive learning metrics used for evaluation and developed the Moodle plugin for system integration. He also wrote the second part of the paper.

References

Bai, J., Bai, S., Yang, S., Wang, Shijie, Tan, Sinan, Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *ArXiv (Computer Science > Artificial Intelligence).* https://doi.org/10.48550/arXiv.2308.12966

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., & Bernstein, M. S. (2022). On the Opportunities and Risks of Foundation Models. *ArXiv (Computer Science > Artificial Intelligence).* https://doi.org/10.48550/arXiv.2108.07258

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modelling and User-Adapted Interaction*, *4*(4), 253–278. https://doi.org/10.1007/bf01099821

DeepSeek, A. I., Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., & Dong, Kai. (2024). DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *ArXiv (Computer Science > Artificial Intelligence).* https://doi.org/10.48550/arXiv.2401.0295

Devlin, Jj., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv (Computer Science > Computation and Language)*, *1*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Du Plooy, E., Casteleijn, D., & Franzsen, D. (2024). Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement. *Heliyon*, *10*(21), e39630. https://doi.org/10.1016/j.heliyon.2024.e39630

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, *2*, 100017. https://doi.org/10.1016/j.caeai.2021.100017

Khezresmaeilzadeh, T., Zhang, J., Andreadis, D., & Psounis, K. (2025). Preserving Privacy and Utility in LLM-Based Product Recommendations. *Computer Science > Information Retrieval*. https://doi.org/10.48550/arXiv.2505.00951

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science*, *36*(5), 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x

Kucirkova, N. (2017). Is technology good for education? By Neil Selwyn. *British Journal of Educational Studies*, *65*(3), 406–408. https://doi.org/10.1080/00071005.2017.1353301

Lee, Y., Cho, J., Han, S., & Choi, B.-U. (2010). A Personalized Assessment System Based on Item Response Theory. *Advances in Web-Based Learning – ICWL 2010*, *6483*, 381–386. https://doi.org/10.1007/978-3-642-17407-0_40

Lewis, P., Perez, E., & Piktus, A. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *ArXiv (Computer Science > Computation and Language)*, *12*, 11401.

Lewis, P., Perez, E., Piktus, A., Petroni, Fabio, Karpukhin, V., & Goyal, N. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *ArXiv (Computer Science > Computation and Language)*. https://doi.org/10.48550/arXiv.2005.11401

Mistral-A.I., A. R., Jiang, Albert Q., Lo, A., Berrada, G., & Lample, G. (2025). Magistral. *Computer Science > Computation and Language*. https://doi.org/10.48550/arXiv.2506.10910

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and Privacy in Machine Learning. *Proceeding of the 2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 399–414. https://doi.org/10.1109/eurosp.2018.00035

Pardos, Z. A., & Heffernan, N. T. (2010). Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. *User Modeling, Adaptation, and Personalization*, *6075*, 255–266. https://doi.org/10.1007/978-3-642-13470-8_24

Ruan, S., & Lu, K. (2025). Adaptive deep reinforcement learning for personalized learning pathways: A multimodal data-driven approach with real-time feedback optimization. *Computers and Education: Artificial Intelligence*, *9*, 100463. https://doi.org/10.1016/j.caeai.2025.100463

Singla, A., Rafferty, A. N., Radanovic, G., & Heffernan, N. T. (2021). Reinforcement Learning for Education: Opportunities and Challenges. *ArXiv (Computer Science > Artificial Intelligence / Education)*. https://doi.org/10.48550/arXiv.2107.08828

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian Knowledge Tracing Models. *Artificial Intelligence in Education*, *7926*, 171–180. https://doi.org/10.1007/978-3-642-39112-5_18