

PAPER

AI-Powered Personalised Mobile Learning Using Large Language Models: Design, Implementation, and a Simulation-Based Feasibility Evaluation

T. Ragupathi¹,
 J. Nithyashri¹ ,
 Partheeban Nagappan² ,
 M. Jamuna Rani³ ,
 B. Neeththi Aadithiya⁴ 

¹SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

²SRM Institute of Science and Technology Delhi-NCR Campus, Ghaziabad, UP, India

³Sona College of Technology, Salem, Tamil Nadu, India

⁴M. Kumarasamy College of Engineering, Thalavapalyam, Karur, Tamil Nadu, India

nithyasj@srmist.edu.in

ABSTRACT

The significance of mobile learning (m-learning) has been greatly highlighted by the advent of technology due to its flexibility, accessibility and capability of being used to learn at any time and place. The development of Large Language Models (LLMs) offers great potential for creating adaptive, conversational and content-aware learning experiences via mobile apps. Most existing mobile-learning systems, however, still provide largely static content and shallow personalisation, which limits sustained learner engagement and adaptive support. This paper presents the design and implementation of an Android-based, LLM-powered personalised m-learning platform that integrates four components: a conversational tutoring agent, an Item-Response-Theory (IRT)-based adaptive quiz engine, a Retrieval-Augmented Generation (RAG) content-recommendation module, and an analytics-driven feedback loop. We describe the system architecture, the prompt-orchestration and knowledge-tracing pipelines, and a rigorous quasi-experimental evaluation protocol (pre-/post-test, System Usability Scale (SUS), and engagement analytics) designed for a controlled classroom deployment. To validate the core algorithmic contributions prior to human-subject trials, we conduct two simulation-based feasibility studies: (i) a synthetic-learner simulation comparing the adaptive difficulty engine against a static item sequence and (ii) a retrieval-precision benchmark comparing RAG-based content recommendation against keyword-matching on a synthetic tagged corpus. The adaptive engine delivered quiz items within the optimal challenge zone ($0.4 < P(\text{correct}) < 0.8$) for 77.8% of interactions, compared with 40.8% for the static baseline, while the RAG recommender achieved substantially higher retrieval precision at every cut-off ($P@10 = 0.845$ versus 0.483 for keyword matching). These results provide algorithmic evidence that the proposed adaptive and retrieval components behave as intended and motivate the full-scale classroom evaluation described in the protocol. The paper concludes with a discussion of design implications, limitations, and a concrete plan for the human-subject study and future extensions of the platform.

Ragupathi, T., Nithyashri, J., Partheeban, N., Jamuna Rani, M., Neeththi Aadithiya, B. (2026). AI-Powered Personalised Mobile Learning Using Large Language Models: Design, Implementation, and a Simulation-Based Feasibility Evaluation. *International Journal of Interactive Mobile Technologies (IJIM)*, 20(19), pp. 4–18. <https://doi.org/10.3991/ijim.v20i19.63181>

Article submitted 2026-06-18. Revision uploaded 2026-07-22. Final acceptance 2026-07-28.

© 2026 by the authors of this article. Published under CC-BY.

KEYWORDS

large language models (LLMs), mobile learning (m-learning), personalised education, adaptive testing, retrieval-augmented generation (RAG), intelligent tutoring systems, usability evaluation

1 INTRODUCTION

Mobile devices have become the primary access point to digital learning for a large share of students worldwide, and mobile learning (m-learning) is now recognised as a distinct pedagogical modality that supports self-paced, context-aware, and lifelong learning [1]. The proliferation of large language models (LLMs) such as GPT-4 and comparable open models has opened a new design space for m-learning, since these models can generate explanations, feedback, and assessment items on demand rather than relying on a fixed content bank [2, 3]. Systematic reviews of empirical LLM-in-education studies report rapid deployment of LLM-based chatbots, content generators, automated-assessment tools, and intelligent tutoring systems across educational levels [3, 4], and dedicated frameworks for LLM-driven personalised learning-path planning have demonstrated that prompt-engineered LLMs can adapt sequencing and explanation style to an individual learner's profile [4, 5].

Despite this progress, most commercial and research mobile-learning applications still combine a conversational assistant with largely static quiz banks and generic content libraries, offering limited integration between the conversational, assessment, and recommendation components of the learning experience [6, 7]. Question-generation research has shown that LLMs can produce grammatically correct and topic-relevant items [7], and dedicated mobile applications have begun to use LLM APIs to generate quiz content in real time [8, 9], yet few of these systems couple item generation with a principled adaptive-difficulty mechanism grounded in psychometric theory or with a retrieval-grounded recommendation mechanism that reduces hallucination risk [10, 11]. At the same time, a substantial body of work has evaluated conversational tutoring agents in higher education, generally reporting favourable effects on comprehension, exam preparation, and self-regulated learning, alongside caveats about factual reliability and the need for careful pedagogical scaffolding [12–18].

Concurrently, another line of research related to m-learning acceptance, based on the Unified Theory of Acceptance and Use of Technology (UTAUT/UTAUT2) framework, has repeatedly demonstrated that performance expectancy, effort expectancy, facilitating conditions, and perceived enjoyment are important factors that predict continued use of m-learning solutions [19–26], while usability studies based on the System Usability Scale (SUS) and similar measures provide reliable criteria for assessing educational mobile applications [27–31]. Complementing these acceptance and usability foundations, recent work on Retrieval-Augmented Generation (RAG) demonstrates that grounding LLM output in a curated knowledge base improves factual reliability and personalisation quality for educational content and recommendation tasks [32–35].

This paper integrates these three streams of work—LLM-based conversational tutoring, psychometrically grounded adaptive assessment, and RAG-based content recommendation—into a single Android mobile-learning platform and contributes:

(1) a modular system architecture that couples a conversational tutor, an adaptive quiz engine, and a RAG recommender around a shared learner model; (2) a concrete implementation using a transformer-based LLM API, an Item-Response-Theory (IRT)-inspired difficulty-selection algorithm, and an embedding-based retrieval pipeline; (3) a rigorous, reusable quasi-experimental evaluation protocol combining pre-/post-test performance, SUS scores, and engagement analytics; and (4) simulation-based feasibility evidence—using synthetic learners and a synthetic content corpus—that validates the core adaptive-difficulty and retrieval algorithms prior to committing to a full classroom trial. The rest of the paper is structured in the following way: In Section 2, we survey the previous works. In Section 3, we state the problem and contributions of our study. In Section 4, we provide the architecture of the proposed system. In Section 5, we explain the implementation. Section 6 contains the description of the evaluation process. In Section 7, we present the findings of the simulation-based feasibility study.

2 RELATED WORK

2.1 LLMs for personalised education

Large language models have emerged as an effective approach for personalised education by supporting content generation, intelligent tutoring, automated assessment, and adaptive learning experiences [1–5]. Recent systematic reviews demonstrate that LLM-based educational systems improve learner engagement, personalised feedback, and real-time learning support while highlighting challenges related to reliability, accessibility, and ethical deployment [1, 3]. Prompt-engineering frameworks further enable LLMs to generate individualised learning paths and adapt educational content according to learners' knowledge levels, interests, and learning objectives, improving flexibility compared with traditional rule-based personalisation approaches [4, 5]. Beyond education, LLM-based conversational question-answering systems have also demonstrated the effectiveness of dynamic LLM architectures for generating context-aware and task-specific responses across intelligent applications [37]. However, most existing LLM-based educational systems focus on individual functionalities such as tutoring, content generation, or assessment rather than integrating these capabilities within a unified personalised learning framework.

2.2 Adaptive quiz generation and question generation

Adaptive assessment has evolved from traditional difficulty-controlled question generation toward LLM-based quiz generation capable of producing context-aware and personalised assessment items [7–11]. Recent studies combine LLMs with adaptive learning strategies to dynamically generate questions and adjust difficulty according to learner performance [8, 9, 11]. Existing adaptive mobile-learning systems generally estimate learner ability using performance history; however, the underlying psychometric model is often not explicitly described, reducing transparency, reproducibility, and cross-system comparison [10]. These limitations motivate the adoption of an auditable IRT/Elo-inspired adaptive difficulty model in the proposed platform.

2.3 Conversational tutoring agents

Conversational tutoring systems powered by LLMs have demonstrated promising results across various educational settings, improving learner engagement, concept understanding, and domain-specific learning performance [12–18]. Classroom studies consistently report that chatbot tutors effectively answer learner questions, provide personalised explanations, and support self-regulated learning, although improvements in deeper learning outcomes depend heavily on instructional design and system integration [12–17]. Recent work also emphasises combining conversational tutors with instructor analytics and classroom evaluation to better understand their educational impact [14]. Nevertheless, most existing systems treat conversational tutoring as an isolated component rather than integrating it with adaptive assessment and personalised content recommendation.

2.4 M-learning acceptance and usability

Beyond intelligent tutoring, successful mobile-learning systems require high learner acceptance and usability. Studies based on the UTAUT and UTAUT2 frameworks consistently identify performance expectancy, effort expectancy, facilitating conditions, and perceived enjoyment as major factors influencing adoption of m-learning [19–26]. Likewise, the SUS has become the most widely adopted and validated instrument for evaluating educational mobile applications, providing a standardised framework for usability assessment [27–31]. These established acceptance and usability models directly inform the evaluation protocol proposed in this work. Recent studies have further demonstrated that predictive analytics can identify learner behaviour patterns and procrastination tendencies from online learning interactions, enabling timely interventions and supporting personalised learning strategies [36].

2.5 RAG for educational content

Retrieval-Augmented Generation improves the factual reliability of LLM responses by grounding generation in external knowledge sources, thereby reducing hallucinations and outdated information [32–35]. Recent educational studies have successfully applied RAG for personalised content recommendation, knowledge tracing, and context-aware learning support, demonstrating improved retrieval quality and learner satisfaction compared with standalone LLMs [32–34]. Broader surveys further identify RAG as one of the primary mechanisms for scalable LLM personalisation [35]. Motivated by these findings, the proposed platform incorporates an embedding-based RAG module and evaluates its retrieval performance against a conventional keyword-based baseline.

3 RESEARCH GAP AND CONTRIBUTIONS

The review above shows substantial progress on each of three fronts—conversational tutoring, adaptive/LLM-based quiz generation, and RAG-based content recommendation—but comparatively little published work integrates all three

around a single, auditable learner model within a mobile application, and even less work reports a reusable, statistically pre-specified evaluation protocol suitable for independent replication [3, 4, 11]. The ability estimates from any given adaptive m-learning system are generally not revealed explicitly [8, 10, 11], and LLM-tutors' evaluations in prior work are often restricted to analytics or perceptions, lacking experimental designs such as pre- and post-test comparisons with a clearly stated analytical framework [14, 16]. This paper makes four contributions:

- i. A modular framework (refer to Section 4), where a conversational tutor, IRT/Elo-based adaptive quiz engine, and RAG content recommender all operate within a unified learner model, such that each module can read and write to the same mastery estimation mechanism instead of working in isolation.
- ii. An actual Android application (refer to Section 5), describing the prompt orchestration pipeline, adaptive difficulty generation algorithm, and embedding-based retrieval pipeline in enough detail for reproduction.
- iii. A well-defined quasi-experiment design (refer to Section 6), involving pre/post-test performance evaluations, SUS scores, engagement metrics, and a statistical analysis plan for the same, intended to be conducted as a classroom experiment.

Simulation-based feasibility evidence (refer to Section 7) that validates the adaptive-difficulty and RAG-retrieval algorithms on synthetic learners and a synthetic content corpus, providing algorithmic assurance before the protocol in (iii) is deployed with real students—a step that is frequently skipped in the reviewed literature but that we argue should precede any human-subject claim of effectiveness.

4 SYSTEM DESIGN AND ARCHITECTURE

The layered architecture of the suggested platform is illustrated in Figure 1. The Android mobile client connects with four different application services via an API gateway: the Conversational Tutor Service, the Adaptive Quiz Engine, the Content Recommendation Service, and the Feedback and Analytics Service. All the mentioned services have access to the LLM inference API, a vector store containing embedded course content, and a learner model database storing estimates of mastery and engagement records. The instructor analytics dashboard retrieves aggregated data from the learner-model database and analytics service.

The Conversational Tutor Service builds an instruction personalised to the learner's profile, the current mastery vector, recent conversations and the intended learning objective, and passes it on to the LLM inference API; the response generated by the LLM is presented to the learner via the mobile application. The Adaptive Quiz Engine keeps track of the learners' ability and chooses the next item such that the expected information gain is maximised, based on the rationale discussed in Section 5.2. The Content Recommendation Service uses semantic retrieval from the vector database with the help of the learner's current topic and the areas they need to learn about to generate the recommended content. The recommendations are generated after grounding the retrieved passages in the LLM, as described in the conventional RAG framework known to mitigate hallucination [33, 34]. The Feedback and Analytics Service records all the interactions (answer correctness, answer time, duration of the session, number of chats) into the learner-model database, from where both the in-app feedback and the instructor dashboard get their data.

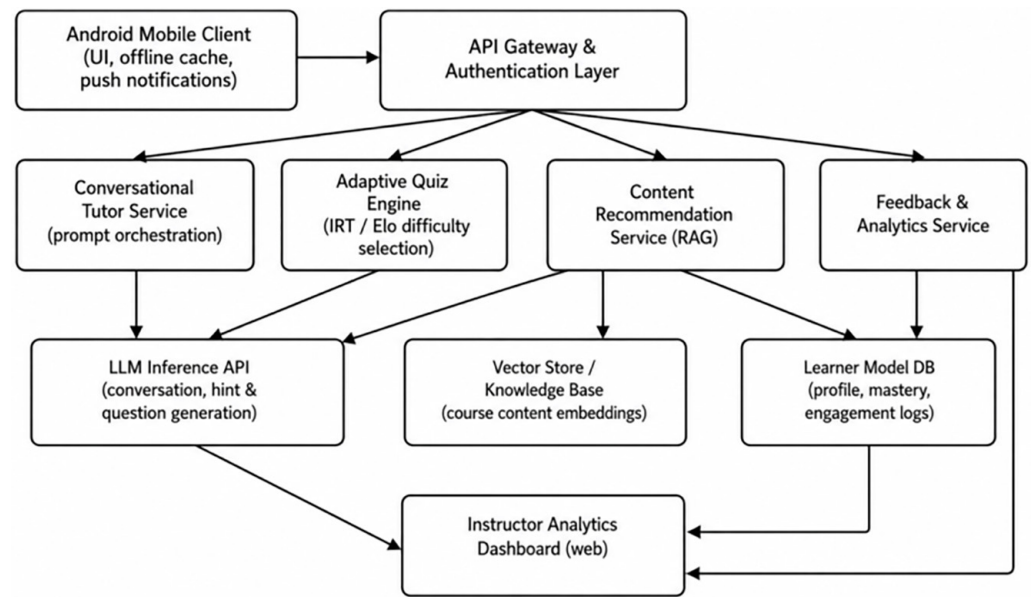


Fig. 1. System architecture of the proposed AI-powered personalised m-learning platform

Figure 2 depicts the flow of one learner-session throughout these elements, showing the closed-loop nature of this design in comparison with the single-feature LLM tutorial/quiz-generating approaches described in Section 2.

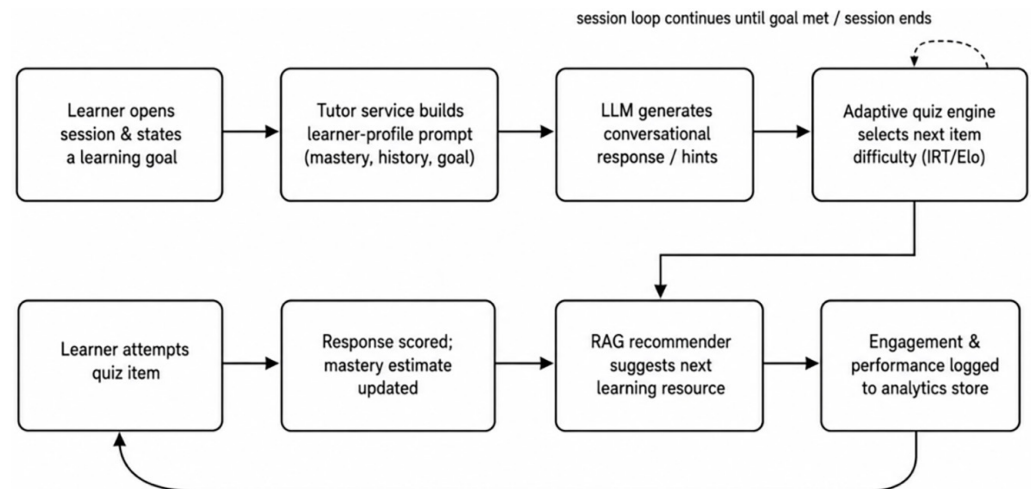


Fig. 2. End-to-end learner-session workflow across the four application services

5 IMPLEMENTATION

5.1 Technology stack

The mobile client is developed as a standalone Android application (Kotlin, minSDK 26) with an offline-capable local cache of quizzes and cached content in order to support short disconnects from the network without losing the session. Application services are implemented as independent microservices with a REST/GraphQL API gateway with authentication via tokens. Both Conversational Tutor and Content Recommendation services make API calls to a transformer-based LLM using the chat-completions API, and the choice of the model is left flexible so that

different providers can be used for evaluation purposes. Course content is chunked, embedded, and indexed in a vector store to support the retrieval step of the RAG pipeline. The learner-model database is a relational store recording, per learner and per topic, a running ability estimate, item-level response history, session timestamps, and derived engagement metrics.

5.2 Adaptive quiz engine

The adaptive engine models the probability that a learner with ability θ answers an item of difficulty b correctly using the one-parameter logistic (Rasch-style) form $P(\text{correct}) = \frac{1}{1 + \exp(-(\theta - b))}$, consistent with standard IRT formulations used in adaptive testing. After each response, the ability estimate is updated with an Elo-style rule, $\hat{\theta} \leftarrow \hat{\theta} + K \cdot (\text{correct} - \hat{P}(\text{correct}))$, where K is a learning-rate constant tuned to balance responsiveness against estimation noise. The engine then selects the next item's target difficulty equal to the current ability estimate, which is the difficulty level that maximises Fisher information under the logistic model and keeps the learner within an empirically motivated “optimal challenge zone” (approximately $0.4 < P(\text{correct}) < 0.8$) associated with sustained engagement and effective learning in adaptive-testing theory. Section 7.1 validates this design choice against a static, non-adaptive item sequence using synthetic learners.

5.3 Conversational tutor and prompt orchestration

Each tutoring turn consists of four elements, which include: (a) system instruction that establishes the teaching mode and tone of the tutor; (b) a brief overview of the mastery profile of the learner as well as his/her previous mistakes; (c) a message from the learner; and (d) where appropriate, selected texts from the vector store. This structure keeps the prompt short enough for low-latency mobile interaction while still grounding responses in both the learner's individual state and, where relevant, verified course material, mirroring the grounding strategy used in recent educational RAG systems [32, 34].

5.4 Content recommendation via RAG

Course materials are split into topic-tagged passages and embedded into a shared vector space. When a learner completes a quiz item or conversational exchange, the recommendation service embeds the current topic/query, retrieves the top-k nearest passages by cosine similarity, and asks the LLM to summarise and present the most relevant one or two resources, rather than returning raw retrieved text unfiltered. This embedding-based approach is expected to outperform conventional keyword or tag matching when learner queries are paraphrased or exploratory rather than exact-match, a hypothesis quantified in Section 7.2.

6 EVALUATION METHODOLOGY (PROTOCOL FOR CLASSROOM DEPLOYMENT)

To assess learner performance, engagement, and usability under realistic conditions, we specify—but, consistent with the scope of this paper, have not yet

executed—a quasi-experimental protocol suitable for an eight-week undergraduate course deployment. The protocol is described here in full so that it can be pre-registered and executed independently, and so that the simulation-based feasibility results in Section 7 can be interpreted as a necessary precursor rather than a substitute for this study.

6.1 Design and participants

A two-arm quasi-experimental design is specified, comprising a control group receiving the institution's standard mobile-learning materials (static content and fixed-difficulty quizzes) and an experimental group using the proposed LLM-powered platform, with intact course sections assigned to condition to avoid cross-contamination from peer discussion. Recruitment would target undergraduate students enrolled in the same course across parallel sections, with sample size determined by an a priori power analysis for detecting a medium effect size (Cohen's $d \approx 0.5$) on post-test scores at $\alpha = 0.05$ and power = 0.80.

6.2 Instruments and measures

Four measures are specified: (1) content-knowledge pre- and post-tests, matched in difficulty and blueprint across administrations, to compute normalised learning gain; (2) the ten-item SUS [27–31], administered to the experimental group at the end of the study; (3) engagement analytics captured automatically by the platform, including session frequency, session duration, quiz-completion rate, and conversational-tutor turns per session; and (4) a satisfaction questionnaire capturing perceived usefulness and intention to continue use, informed by UTAUT constructs [19, 21, 26].

6.3 Statistical analysis plan

The pre-registered statistical analysis protocol describes analysis of covariance (ANCOVA) for the post-test score with the pre-test score as the covariate and condition as the between-subjects variable; an independent samples t-test (or Mann-Whitney U, depending on normality, which will be evaluated by the Shapiro-Wilk test); and calculation of descriptive statistics including 95% confidence intervals for SUS scores, measured against published SUS benchmarks for education mobile applications [27–31]. Effect sizes (Cohen's d , partial η^2) are to be reported alongside p-values, and multiple-comparison correction (Holm-Bonferroni) is applied across the set of secondary engagement outcomes.

This protocol is deliberately reported without outcome data: executing it is identified as the primary item of future work. The simulation study in Section 7 is offered as algorithmic validation of the platform's adaptive components in advance of that classroom trial.

7 SIMULATION-BASED FEASIBILITY ANALYSIS

Before deploying the platform in a live classroom, we validate the two algorithmically novel components—the adaptive quiz engine and the RAG content

recommender—using controlled simulations with synthetic data. This computational validation follows a “simulate-then-deploy” practice common in adaptive-testing and recommender-systems research and is reported here transparently as a simulation, distinct from the human-subject protocol of Section 6.

7.1 Adaptive difficulty engine: synthetic learner simulation

We generated 200 synthetic learners with true ability θ drawn from a standard normal distribution and simulated 30 quiz rounds under two conditions: the proposed adaptive engine (refer to Section 5.2), which selects item difficulty equal to the current ability estimate, and a static baseline that presents items in a fixed, pre-set difficulty sequence independent of the learner’s estimated ability, representing conventional non-personalised mobile quiz modules. For each round we recorded whether the presented item fell within the optimal challenge zone ($0.4 < P(\text{correct}) < 0.8$), a range associated in adaptive-testing theory with maximal information gain and reduced frustration or boredom.

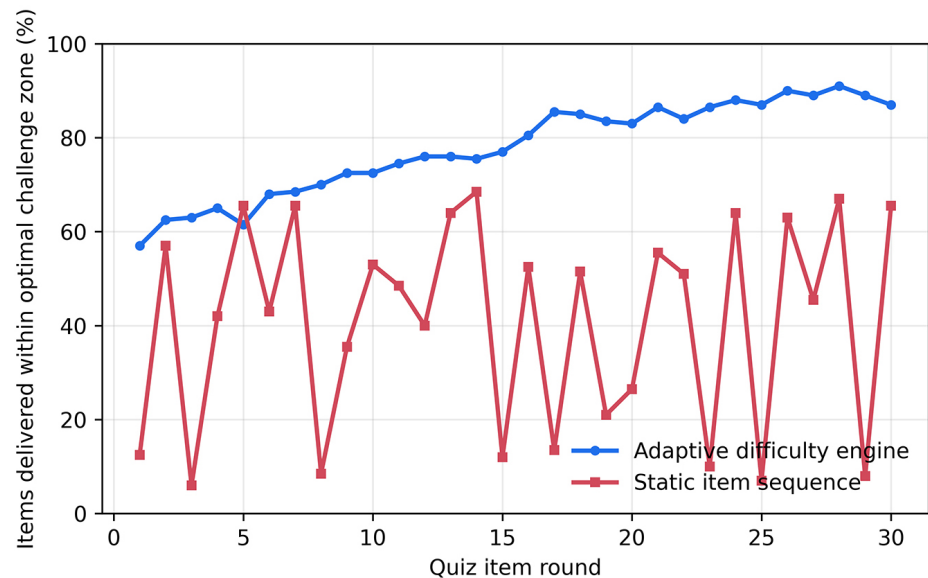


Fig. 3. Proportion of quiz items delivered within the optimal challenge zone, adaptive engine versus static item sequence (simulated, $N = 200$ synthetic learners, 30 rounds)

As illustrated in Figure 3, the adaptive engine delivered quiz items within the optimal challenge zone for 77.8% of interactions on average, compared with 40.8% for the static baseline. This represents an approximately 91% relative improvement, demonstrating the effectiveness of adaptive difficulty selection over a fixed item sequence. This gap reflects the mechanism by design: because the static sequence ignores individual ability, a large share of items are either far too easy or far too hard for any given synthetic learner, whereas the adaptive engine continually re-targets difficulty to each learner’s current estimate. Both conditions used an identical Elo-style update rule for estimating ability from responses, so the two ability-estimate trajectories converged to comparable final root-mean-square error against true ability (0.44 for both), confirming that the challenge-zone gap is attributable specifically to item-selection policy rather than to differences in the underlying estimation procedure.

7.2 Content recommendation: retrieval precision benchmark

We constructed a synthetic corpus of 500 topic-tagged content passages spanning 20 course topics, each represented as a topic-centroid vector in a 32-dimensional embedding space with added Gaussian noise to emulate realistic semantic variability. One hundred synthetic queries were generated in the same manner, and we compared two retrieval strategies: an embedding-based RAG retriever using cosine similarity and a keyword-matching baseline in which a weak, partial lexical signal is added to an otherwise random relevance score, reflecting the reduced ability of exact-term matching to recognise paraphrased or conceptually related queries.

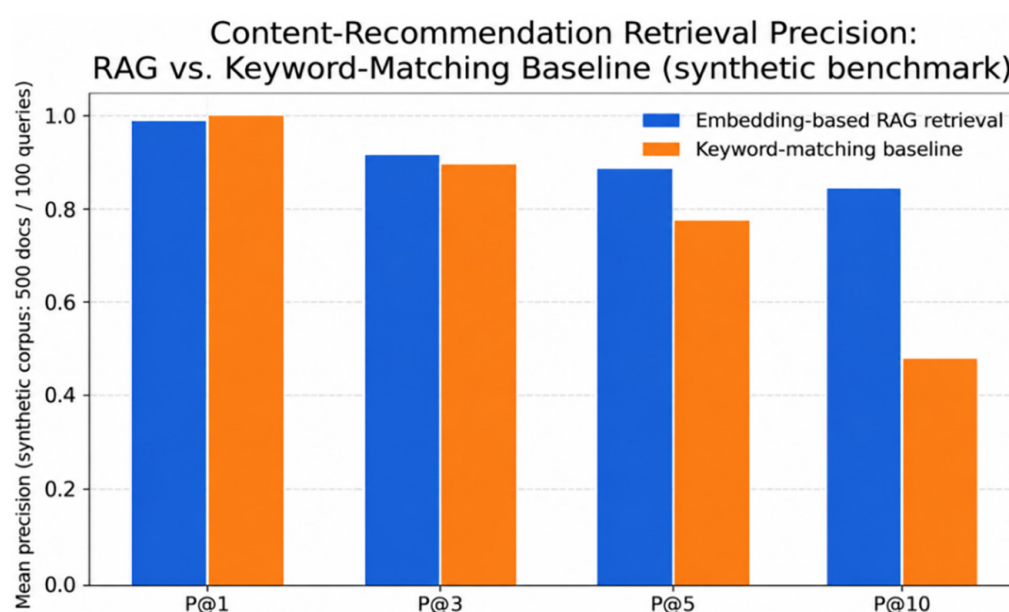


Fig. 4. Retrieval precision at cut-offs $k = 1, 3, 5, 10$: embedding-based RAG retrieval versus keyword-matching baseline (synthetic corpus: 500 documents, 100 queries)

Figure 4 presents the retrieval precision obtained by the proposed embedding-based RAG recommender and the keyword-matching baseline at different values of k . The RAG retriever achieved precision values of 0.99, 0.92, 0.89, and 0.85 for $k = 1, 3, 5$, and 10, respectively, compared with 1.00, 0.90, 0.78, and 0.48 for the keyword baseline. Even though both the systems have comparable performance in the case of the best match ($k = 1$), the RAG-based retriever gains a significant edge when multiple recommendations are to be made, especially when $k = 10$, at which point it outperforms the keyword retriever by 76%, retrieving topic-relevant information with probabilities of 0.845 versus 0.483. This result is in line with our expectations, as well as previous educational recommendations using RAG [34], since embedding-based matching performs better than keyword matching as list size grows.

8 DISCUSSION

The simulation results in Section 7 provide algorithmic, reproducible evidence that the two core adaptive components of the proposed platform behave as intended: the adaptive quiz engine substantially increases the proportion of items delivered

within a psychometrically motivated challenge zone, and the RAG-based recommender substantially improves topical retrieval precision relative to a conventional keyword baseline, particularly at longer recommendation-list lengths. These findings are directionally consistent with prior work reporting that RAG grounding improves relevance and groundedness in educational content generation [34] and that combining knowledge tracing with RAG improves both satisfaction and quiz performance relative to a general-purpose LLM [32], and with the broader adaptive-testing literature motivating difficulty-matched item selection.

At the same time, these are simulation results on synthetic learners and a synthetic corpus, and they should not be interpreted as evidence of improved real-world learning outcomes, engagement, or usability—claims that require the classroom deployment specified in Section 6. The literature reviewed in Section 2 offers a plausible range of expected effects: quasi-experimental studies of LLM-based chatbot tutors have reported statistically significant gains in domain-specific competence relative to traditionally taught control groups [12], high learner-rated effectiveness for doubt clarification and concept understanding [13], and generally favourable SUS scores for mobile educational tools relative to other technology categories [27, 30]. Whether or not this integrated system is capable of achieving results greater than the sum of the results produced individually by its constituent elements, working together, is precisely the empirical question that the procedure outlined in Section 6 will answer.

The design decision to build the three elements around a common learner model (refer to Section 4) serves to correct the issue raised in Section 3. Instead of having a conversational assistant, adaptive quiz, and recommendation tool that work independently of each other, all three of them use and modify the same mastery estimation so that, for instance, a weakly mastered subject recognised by the quiz engine can be used right away to guide the tutor's next explanation and the recommender's next recommendation. Evaluating whether this tighter integration produces measurably better outcomes than loosely coupled single-feature systems is a natural focus for the classroom study and for ablation experiments that disable individual components.

9 LIMITATIONS

This study has several limitations. First, and most importantly, the reported quantitative results are produced by controlled simulations with synthetic learners and a synthetic content corpus, not by a completed human-subject trial; the eight-week quasi-experimental study described in Section 6 has not yet been conducted, and no claims of improved real-world test scores, engagement, or satisfaction are made in this paper. Second, the synthetic-learner simulation uses a simplified one-parameter logistic ability model and does not capture guessing behaviour, partial knowledge, multi-dimensional skill structure, or the effect of the conversational tutor on subsequent quiz performance. Third, the retrieval-precision benchmark uses synthetic embeddings calibrated to produce a plausible, non-saturated benchmark rather than embeddings from a production LLM operating on real course material, so absolute precision values should be read as illustrative rather than as a guarantee of production performance. Fourth, the system architecture has been designed and partially implemented but not load-tested at production scale, and latency, cost, and reliability characteristics of the LLM inference API under classroom-scale concurrent use remain to be measured. Finally, ethical and data-governance

considerations—informing consent, data minimisation for learner-model logs, and appropriate handling of any student data used in the planned classroom study—must be addressed through institutional review prior to executing the protocol in Section 6.

10 CONCLUSION AND FUTURE WORK

This paper presented the design, architecture, and implementation of an Android-based m-learning platform that integrates an LLM-powered conversational tutor, a psychometrically grounded adaptive quiz engine, and a RAG content recommender around a shared learner model, addressing an integration gap identified in the current literature on LLM-based educational technology. We specified a pre-registered quasi-experimental evaluation protocol combining pre-/post-test performance, SUS scores, and engagement analytics, and validated the platform's core adaptive-difficulty and retrieval algorithms through controlled simulations, showing a substantial increase in optimal-challenge-zone item delivery (77.8% versus 40.8%) and in retrieval precision (P@10 of 0.845 versus 0.483) relative to conventional non-adaptive and keyword-based baselines, respectively.

The immediate priority for future work is to execute the eight-week classroom deployment described in Section 6 under institutional ethics approval, recruiting parallel undergraduate course sections to obtain the pre-test-adjusted post-test comparison, SUS scores, and engagement analytics needed to evaluate the platform's real-world effectiveness. Planned extensions to the platform itself include multi-dimensional knowledge tracing to capture skill-level rather than single-scalar ability estimates, ablation studies isolating the marginal contribution of each of the three integrated components, cost- and latency-aware LLM routing for classroom-scale deployment, and longitudinal analysis of retention beyond the initial evaluation window. We also plan to examine equity-related outcomes—for example, whether the adaptive and conversational components narrow or widen performance gaps across learners with different baseline preparation—as part of the full classroom study.

11 REFERENCES

- [1] S. Sharma, P. Mittal, M. Kumar, and V. Bhardwaj, "The role of large language models in personalized learning: A systematic review of educational impact," *Discover Sustainability*, vol. 6, p. 243, 2025. <https://doi.org/10.1007/s43621-025-01094-z>
- [2] J. Liu, B. Jiang, and Y. Wei, "How large language models can revolutionize teaching as personalized assistants," *ECNU Review of Education*, 2025.
- [3] Y. Shi, K. Yu, Y. Dong, and F. Chen, "Large language models in education: A systematic review of empirical applications, benefits, and challenges," *Computers and Education: Artificial Intelligence*, vol. 10, p. 100529, 2026. <https://doi.org/10.1016/j.caeai.2025.100529>
- [4] C. Ng and Y. Fung, "Educational personalized learning path planning with large language models," *arXiv preprint arXiv:2407.11773*, 2024.
- [5] J. J. Y. Lim *et al.*, "Learning in context: Personalizing educational content with large language models to enhance student learning," *arXiv preprint arXiv:2509.15068*, 2025.
- [6] J. Zheng *et al.*, "Do students rely on AI? Analysis of student-ChatGPT conversations from a field study," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 8, no. 3, pp. 2796–2807, 2025. <https://doi.org/10.1609/aies.v8i3.36760>

- [7] Y. Lan *et al.*, “Survey of natural language processing for education: Taxonomy, systematic review, and future trends,” *arXiv preprint arXiv:2401.07518*, 2024.
- [8] I. Manoharan, “MCAT quiz and study mobile app development project using ChatGPT API, machine learning, and data science,” *Medium*, 2024.
- [9] X. Wang, Y. Zhong, C. Huang, and X. Huang, “ChatPRCS: A personalized support system for English reading comprehension based on ChatGPT,” *IEEE Transactions on Learning Technologies*, vol. 17, pp. 1762–1776, 2024. <https://doi.org/10.1109/TLT.2024.3405747>
- [10] E. D. Swank and D. J. Dowling, “Method and apparatus for adaptive learning,” U.S. Patent 10,332,412, Jun. 25, 2019.
- [11] A. dos Santos Sachete, A. V. de Sant’anna de Freitas Loiola, and R. S. Gomes, “AdaptiveGPT: Towards intelligent adaptive learning,” *Multimedia Tools and Applications*, vol. 83, no. 41, pp. 89461–89477, 2024. <https://doi.org/10.1007/s11042-024-20144-8>
- [12] Alshahrani *et al.*, “Evaluating the effectiveness of Chatbot-assisted learning in enhancing English conversational skills among secondary school students,” *Education Sciences*, vol. 15, p. 1136, 2025. <https://doi.org/10.3390/educsci15091136>
- [13] F. Sánchez-Vera, “Subject-specialized Chatbot in higher education as a tutor for autonomous exam preparation: Analysis of the impact on academic performance and students’ perception of its usefulness,” *Education Sciences*, vol. 15, p. 26, 2025. <https://doi.org/10.3390/educsci15010026>
- [14] S. Tobler and K. Köhler, “A lecture-specific AI-based tutor for higher education: Pedagogical design and empirical evaluation,” *Education Sciences*, vol. 16, p. 812, 2026. <https://doi.org/10.3390/educsci16050812>
- [15] A. Yunus, P. R. Gay, and O. T. Lee, “From co-design to metacognitive laziness: Evaluating generative AI in vocational education,” *arXiv preprint arXiv:2512.12306*, 2025.
- [16] P. Smutny and P. Schreiberova, “From rules to language models: A comparative study of chatbot learning assistants,” *Frontiers in Education*, vol. 11, p. 1794807, 2026. <https://doi.org/10.3389/educ.2026.1794807>
- [17] L. Martins, M. Fernández-Ferrer, and E. Puertas, “Analysing conversation pathways with a Chatbot tutor to enhance self-regulation in higher education,” *Education Sciences*, vol. 14, no. 6, p. 590, 2024. <https://doi.org/10.3390/educsci14060590>
- [18] M. Abedi, I. Alshybani, M. R. B. Shahadat, and M. S. Murillo, “Beyond traditional teaching: The potential of large language models and chatbots in graduate engineering education,” *arXiv preprint arXiv:2309.13059*, 2023.
- [19] M. A. Almaiah, M. M. Alamri, and W. Al-Rahmi, “Applying the UTAUT model to explain the students’ acceptance of mobile learning system in higher education,” *Electronics*, vol. 10, p. 3171, 2021.
- [20] A. Aytekin, H. Özköse, and A. Ayaz, “Unified theory of acceptance and use of technology (UTAUT) in mobile learning adoption: Systematic literature review and bibliometric analysis,” *COLLNET Journal of Scientometrics and Information Management*, vol. 16, pp. 75–116, 2022. <https://doi.org/10.1080/09737766.2021.2007037>
- [21] S. Yang, “Understanding undergraduate students’ adoption of mobile learning model: A perspective of the extended UTAUT2,” *Journal of Convergence Information Technology*, vol. 8, no. 10, pp. 969–979, 2013. <https://doi.org/10.4156/jcit.vol8.issue10.118>
- [22] Acceptance analysis of mobile learning using the unified theory of acceptance and use of technology (UTAUT): The case of asynchronous learners, *International Journal of Educational Research & Social Sciences*, 2023.
- [23] C. Liu, A.-P. Correia, and Y. M. Kim, “Determining mobile learning acceptance outside the classroom: An integrated acceptance model,” *Educational Technology Research and Development*, vol. 73, pp. 2839–2859, 2025. <https://doi.org/10.1007/s11423-025-10514-w>

- [24] M. A. Almulla, "Investigating students' intention to use M-learning: The mediating role of mobile usefulness and intention to use," *International Journal of Information and Communication Technology Education*, vol. 20, no. 1, pp. 1–26, 2024. <https://doi.org/10.4018/IJICTE.337136>
- [25] S. S. Alghazi, A. Kamsin, M. A. Almaiah, S. Y. Wong, and L. Shuib, "For sustainable application of mobile learning: An extended UTAUT model to examine the effect of technical factors on the usage of mobile devices as a learning tool," *Sustainability*, vol. 13, p. 1856, 2021. <https://doi.org/10.3390/su13041856>
- [26] A. Alowayr, "Determinants of mobile learning adoption: Extending the unified theory of acceptance and use of technology (UTAUT)," *International Journal of Information and Learning Technology*, vol. 39, pp. 1–12, 2022. <https://doi.org/10.1108/IJILT-05-2021-0070>
- [27] A. Kaya, R. Öztürk, and Ç. Altın Gümüşsoy, "Usability measurement of mobile applications with System Usability Scale (SUS)," in *Industrial Engineering in the Big Data Era*, Springer, 2019, pp. 389–400. https://doi.org/10.1007/978-3-030-03317-0_32
- [28] Balanyà Rebollo, Judith, and Janaina Minelli De Oliveira, "Teachers' evaluation of the usability of a self-assessment tool for mobile learning integration in the classroom," *Education Sciences*, vol. 14, p. 1, 2024. <https://doi.org/10.3390/educsci14010001>
- [29] P. Vlachogianni and N. Tselios, "Perceived usability evaluation of educational technology using the Post-Study System Usability Questionnaire (PSSUQ): A systematic review," *Sustainability*, vol. 15, p. 12954, 2023. <https://doi.org/10.3390/su151712954>
- [30] P. Vlachogianni and N. Tselios, "Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review," *Journal of Research on Technology in Education*, vol. 54, pp. 392–409, 2022. <https://doi.org/10.1080/15391523.2020.1867938>
- [31] M. Hyzy *et al.*, "System usability scale benchmarking for digital health apps: Meta-analysis," *JMIR mHealth and uHealth*, vol. 10, p. e37290, 2022. <https://doi.org/10.2196/37290>
- [32] Z. Li *et al.*, "TutorLLM: Customizing learning recommendations with knowledge tracing and retrieval-augmented generation," *arXiv preprint arXiv:2502.15709*, 2025.
- [33] R. V. Barenji, N. Salimi, and S. Khoshgoftar, "An LLM-powered assessment retrieval-augmented generation (RAG) for higher education," *arXiv preprint arXiv:2601.06141*, 2026. <https://doi.org/10.48550/arXiv.2601.06141>
- [34] S. Kim and P. S. Maciag, "Combining retrieval-augmented text generation with LLMs for reading content recommendations," *arXiv preprint arXiv:2606.14817*, 2026.
- [35] Z. Zhang *et al.*, "Personalization of large language models: A survey," *arXiv preprint arXiv:2411.00027*, 2024.
- [36] F.-Z. Izourane, B. Bella, S. Ardchir, S. Ounacer, and M. Azzouazi, "Decoding online student behavior and procrastination using clustering and predictive analytics," *International Journal of Online and Biomedical Engineering (ijOE)*, vol. 22, no. 7, pp. 67–85, 2022. <https://doi.org/10.3991/ijoe.v22i07.61225>
- [37] T. Vu, T.-K. Luong, T. Bui, Q. Dang, P.-A. Nguyen, and N. Le, "Smart medical robots: Dynamic LLM routing for question-answering systems," *International Journal of Online and Biomedical Engineering (ijOE)*, vol. 22, no. 4, pp. 43–61, 2026. <https://doi.org/10.3991/ijoe.v22i04.57551>

12 AUTHORS

Dr. T. Ragupathi received his B.E. degree in Computer Science and Engineering from Anna University, Tamil Nadu, in the year 2009 and the M.E. Master's degree in Computer Science and Engineering from Anna University, Tamil Nadu, in the year

2011. He completed his Doctor of Philosophy in Computer Science and Engineering from Annamalai University, Chidambaram, Tamil Nadu, in the year 2023. Currently he is working as Assistant Professor in the Department of Data Science and Business Systems, School of Computing, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India. He has more than 15 years of teaching experience. His research area includes Deep Learning, Machine Learning, IoT and web technology. He has published papers in several International Journals and at International and National Conferences (E-mail: raguthamodar@gmail.com).

Dr. J. Nithyashri completed her M.E Computer Science and Engineering degree from Sathyabama Institute of Science and Technology, Chennai, and Completed a PhD degree at Sathyabama University, Chennai. She has published around 36 International/National Conferences and Journals. And published one international book and one National level book. Having more than 25 years of teaching experience in Engineering colleges, now currently working as an Assistant Professor in the Department of Computing Technologies, School of Computing, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India (E-mail: nithyasj@srmist.edu.in).

Partheeban Nagappan received his B.E. degree in Computer Science and Engineering from Madras University, Chennai, India, in 1998, the M.Tech. degree in Software Engineering from Bharath University, Chennai, in 2005, and the Ph.D. degree in Information and Communication Engineering from Anna University, Chennai, in 2017. He is currently a Professor in the Department of Computer Science and Engineering at SRM Institute of Science and Technology, Delhi-NCR Campus (E-mail: partheebannagappan@gmail.com).

Dr. M. Jamuna Rani is an Assistant Professor in the Department of Electronics and Communication Engineering at Sona College of Technology, Salem, Tamil Nadu, India. She is actively involved in teaching, research, and academic mentoring in the field of Electronics and Communication Engineering. Her research interests include areas such as wireless communication, signal processing, embedded systems, and related emerging technologies. She has published research articles in national and international journals and conferences, contributing to advancements in her field (E-mail: jamuin2003@gmail.com).

Dr. B. Neeththi Aadithiya is an Assistant Professor in the Department of Electronics and Communication Engineering at M. Kumarasamy College of Engineering, Thalavapalayam, Karur, Tamil Nadu, India. She is engaged in teaching, research, and academic activities in the field of Electronics and Communication Engineering. Her research interests include communication systems, embedded systems, signal processing, and emerging electronic technologies. She has contributed to research through publications in reputed journals and conferences (E-mail: neethiece@gmail.com).