IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

AI-Driven Personal Tutor Using Large Language Models

Dr. Sanjaykumar P. Pingat[1], Ashish Rajendra Khedkar[1], Srushti Nalawade[2], Jagruti Sanjay Nandre[3], Piyush Ghorapade[3]
UG Students, Department of Computer Engineering[1-4]
Smt Kashibai Navale College of Engineering, Pune, India

Abstract: *The rapid advancement of Large Language Models (LLMs) has opened new possibilities for intelligent and personalized education systems. This paper presents a Personal Tutor using LLMs that delivers adaptive, context-aware, and interactive learning experiences to students. The proposed system integrates Retrieval-Augmented Generation (RAG) to enhance the factual accuracy of responses by combining knowledge retrieval with generative capabilities. It utilizes student profiling and interaction analytics to personalize content based on individual learning pace, knowledge level, and weak topic identification. The system is capable of answering queries, generating quizzes, providing feedback, and recommending learning resources such as educational videos. A modular architecture is implemented using modern web technologies and LLM APIs, ensuring scalability and real-time interaction. Experimental evaluation demonstrates improved response relevance, user engagement, and learning effectiveness compared to traditional e-learning approaches. The proposed solution highlights the potential of AI-driven tutoring systems in delivering scalable and personalized education.*

Keywords: Large Language Models (LLMs), Personalized Learning, Retrieval-Augmented Generation (RAG), Intelligent Tutoring Systems, Natural Language Processing (NLP), Adaptive Learning, Student Profiling, Educational Technology

I. INTRODUCTION

With the rapid growth of artificial intelligence and natural language processing, personalized learning has become a key focus in modern education. Traditional tutoring methods often lack adaptability to individual students' needs, making it challenging to provide customized guidance at scale.

This project aims to develop a Personal Tutor system leveraging Large Language Models (LLMs) combined with Retrieval-Augmented Generation (RAG) to connect and utilize a structured knowledge database. The system is designed to understand student queries, retrieve relevant information from the database, and generate accurate, context-aware, and personalized responses. By integrating RAG, the system not only answers questions but also ensures that the information is up-to-date and specific to the student's learning requirements. The primary objectives of this project are:

- To create an AI-powered tutor capable of handling diverse academic queries.
- To implement a retrieval mechanism that enhances response accuracy by consulting a knowledge database.
- To provide personalized guidance and recommendations for effective learning.

This project bridges the gap between human tutoring and AI-assisted education, aiming to make learning more interactive, accessible, and adaptive.

A. Background and Motivation

The convergence of artificial intelligence and education has opened new pathways for delivering learning experiences that are not constrained by geography, time, or instructor availability. Conventional classroom instruction inherently

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

struggles to accommodate the diverse cognitive profiles of individual learners. A single educator serving a large cohort cannot realistically monitor each student's comprehension rate, identify specific knowledge gaps, or adjust explanations dynamically.

The emergence of Large Language Models (LLMs) such as GPT-4, LLaMA, and Claude has fundamentally shifted what is computationally achievable in natural language understanding and generation. These models demonstrate a capacity for contextual reasoning, multi-turn dialogue, and domain-specific explanation that closely approximates skilled human communication. Simultaneously, Retrieval-Augmented Generation (RAG) has addressed one of the core reliability concerns associated with standalone LLMs — the tendency to generate plausible but factually unsupported responses. By coupling generative capability with structured retrieval from verified knowledge sources, RAG-enabled systems can ground their outputs in curriculum-specific material, making them significantly more suitable for academic deployment.

B. Problem Statement

Despite the proliferation of digital learning platforms, a persistent gap remains between what technology currently offers and what genuinely personalized education requires. Most existing e-learning tools deliver content uniformly, regardless of whether a learner is encountering a concept for the first time or revisiting a topic they partially understand. Intelligent Tutoring Systems developed over the past two decades typically rely on rigid rule-based models that do not scale gracefully across subjects or learner types. Three specific deficiencies motivate this work:

- **Lack of curriculum grounding:** General-purpose LLMs are not aware of a student's specific syllabus, uploaded notes, or course materials, leading to responses misaligned with the learner's immediate academic context.
- **Absence of adaptive feedback loops:** Most tools do not maintain longitudinal models of student performance, preventing identification of weak areas and adjustment of instructional strategy.
- **Limited real-time assessment capability:** Generating contextually relevant, difficulty-calibrated assessments on demand remains an unsolved challenge in widely deployed educational software.

C. Research Objectives

The primary goal of this work is to design, implement, and evaluate an AI-driven personal tutoring system that meaningfully improves upon the adaptability limitations of existing e-learning solutions. Specifically, this research pursues the following objectives:

- **Curriculum-Aware Question Answering:** To develop a RAG pipeline that indexes student-uploaded academic materials as the primary retrieval source for generating contextually grounded responses.
- **Adaptive Learner Modelling:** To implement an interaction-driven profiling mechanism that continuously tracks query patterns and assessment outcomes to identify a student's weak knowledge areas.
- **Dynamic Assessment Generation:** To enable the system to automatically produce multiple-choice questions and conceptual explanations tailored to a learner's current proficiency level.
- **Multimodal Learning Support:** To supplement text-based tutoring with relevant video recommendations, bridging the gap between explanatory dialogue and visual learning.
- **Scalable System Architecture:** To architect the platform in a modular, API-driven manner that supports concurrent users and is extensible to new subjects and deployment environments.
- **Empirical Evaluation:** To assess system performance through structured test cases covering response accuracy, latency, hallucination rate, and user satisfaction.

D. Scope and Limitations

This project encompasses end-to-end design and deployment of a web-based personal tutoring system supporting student authentication and profile management, document upload and semantic indexing, conversational question

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

answering grounded in uploaded materials, automated weak topic detection, quiz generation calibrated to learner performance, and YouTube-based video recommendations. Constraints include dependence on LLM quality and retrieved document relevance, limited affective state modeling, and infrastructure dependency on stable internet connectivity and third-party APIs.

II. RELATED WORK / LITERATURE REVIEW

The development of AI-driven educational systems has been an active area of research for several decades. The advent of transformer-based large language models has dramatically accelerated progress in this domain, enabling a new generation of tutoring systems capable of open-ended dialogue, adaptive content delivery, and real-time assessment.

A. LLMs in Educational Settings

The application of large language models to formal education has gained considerable scholarly attention since the public availability of instruction-tuned models. Han et al. (2024) examined the deployment of LLMs as writing tutors for English as a Foreign Language students, demonstrating that structured, rubric-guided prompting could produce feedback of measurable pedagogical value. Divyansh et al. (2024) conducted a systematic review cataloguing evidence of improved learner engagement, faster concept clarification, and reduced dependency on scheduled human instruction, while also surfacing concerns about hallucination susceptibility and limited awareness of individual learner history. Wang et al. (2024) concluded that LLMs demonstrate versatility across educational tasks, yet their deployment without retrieval grounding or learner modeling produces inconsistent outcomes particularly in subjects demanding precise factual accuracy.

B. Retrieval-Augmented Generation for Knowledge-Grounded Responses

A significant limitation of standalone LLMs in high-stakes educational contexts is their reliance on parametric knowledge that may be outdated or incorrectly recalled. Retrieval-Augmented Generation, introduced by Lewis et al. (2020), addresses this by dynamically fetching relevant document passages at inference time and conditioning the model's output on retrieved evidence. Zhu et al. (2024) extended the RAG paradigm by incorporating knowledge graphs into the retrieval process (KG2RAG), yielding measurably higher coherence and factual precision. Vector database technologies such as FAISS and ChromaDB have emerged as standard infrastructure for RAG pipelines, enabling sub-linear approximate nearest-neighbour search over high-dimensional embedding spaces.

C. Intelligent Tutoring Systems and Adaptive Learning

Intelligent Tutoring Systems (ITS) predate the LLM era by several decades, with early systems such as LISP Tutor and Cognitive Tutor demonstrating that rule-based student modeling could produce measurable learning gains. The Bayesian Knowledge Tracing (BKT) model represents student knowledge as a latent binary variable, effective within tightly scoped domains but requiring exhaustive hand-crafting of skill models. Deep Knowledge Tracing (DKT), proposed by Piech et al. (2015), replaced the Bayesian framework with recurrent neural networks capturing longer interaction histories. The proposed system implements a lightweight adaptive model tracking query frequency, quiz performance, and concept revisitation patterns, avoiding full knowledge tracing complexity while enabling meaningful personalization.

D. Quiz Generation and Assessment Using AI

Automated question generation has been an active research subfield within NLP. Tan et al. (2024) introduced QARR-FSQA, a pretraining framework that improves question-answer pair generation under few-shot conditions. Falatouri et al. (2024) demonstrated that LLMs including GPT-4 and Claude 3 could reliably analyze free-text student responses for quality indicators such as conceptual accuracy, completeness, and reasoning depth. The quiz generation module in the proposed system uses the LLM to produce multiple-choice questions grounded in retrieved curriculum content, with difficulty adjusted dynamically based on the student's recorded quiz performance history.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

III. METHODOLOGY

The methodology is organized around three interdependent processes: preparing and representing academic documents in a retrievable semantic form, orchestrating the retrieval and generation pipeline that produces grounded tutoring responses, and continuously analyzing student interaction patterns to identify and remediate conceptual weaknesses.

A. Preprocessing and Embedding Strategy

Raw academic documents are processed by a format-aware extraction routine. Native PDF files are handled using PyMuPDF, which preserves paragraph boundaries and reading order. DOCX files are processed using python-docx. For scanned or image-heavy documents, a Tesseract OCR fallback is triggered automatically. Post-extraction, a cleaning pass removes reference list entries, figure captions, footnotes, and repeated boilerplate phrases.

Cleaned text is segmented into overlapping chunks using a sliding window approach: each chunk spans approximately 512 tokens with a 64-token overlap. Each chunk is encoded into a dense 384-dimensional vector using the all-MiniLM-L6-v2 sentence transformer model. Embeddings are stored in ChromaDB under a collection namespace partitioned by student ID, ensuring strict retrieval isolation between users.

B. RAG Pipeline and Prompt Engineering

When a student submits a query, the query string is encoded using the same sentence transformer, producing a 384-dimensional query vector compared against all chunk embeddings using cosine similarity. The top five chunks exceeding a similarity threshold of 0.65 are retrieved and ranked by score. The system employs a four-component prompt architecture: a system-level role definition establishing the model as an academic tutor; the retrieved context block; a summarized conversation history covering the last five turns; and an adaptive difficulty modifier injected based on the student's profile. The assembled prompt is submitted to the LLM API with a temperature of 0.3 and a maximum output of 1,024 tokens.

C. Weak Topic Detection Algorithm

Topic proficiency is quantified through a composite weak topic score W(T) computed per topic per student over a rolling seven-day evaluation window:

$$W(T) = [QueryCount(T) / TotalQueries] \times [1 - AvgQuizScore(T)]$$

Topics for which $W(T) \geq \theta$ are flagged as weak areas, where θ is set to 0.10 by default. Flagged topics trigger the adaptive difficulty modifier, targeted quiz generation, and surfacing on the student's performance dashboard.

D. System Architecture

The proposed system is designed using a modular and scalable architecture consisting of five primary components: User Interface, Student Profiling Module, Knowledge Base, Retrieval Engine, and Language Model Generator. The User Interface serves as the interaction layer. The Student Profiling Module maintains learning history, difficulty level, and prior performance. The Knowledge Base contains curated educational content converted into vector embeddings. The Retrieval Engine retrieves the most relevant documents using cosine similarity. The Language Model Generator integrates retrieved context with user queries to produce coherent and accurate responses.

Fig. 1. System Architecture of the Retrieval-Augmented Generation (RAG) Workflow.

E. Advanced Algorithmic Techniques

The core technique is Retrieval-Augmented Generation (RAG), combining semantic information retrieval with transformer-based LLMs. User queries are transformed into dense vector embeddings using Sentence-BERT, enabling high-dimensional similarity searches. Adaptive content selection algorithms dynamically adjust response complexity and instructional style. The architecture incorporates real-time feedback loops and approximate nearest neighbour

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

(ANN) search algorithms such as Hierarchical Navigable Small World (HNSW) graphs, ensuring rapid access to relevant knowledge for large-scale educational content.

IV. IMPLEMENTATION AND RESULTS

A. System Implementation

The system is implemented as a full-stack web application integrating a React.js frontend, a Python FastAPI backend, MongoDB for user data storage, and ChromaDB for vector-based retrieval. The frontend provides views for chat, document upload, quiz interface, and performance tracking dashboard. The backend handles API requests, user authentication, and asynchronous processing for document handling and analysis.

TABLE I — TECHNOLOGY STACK

Layer	Technology Used
Frontend	React.js
Backend	Python (FastAPI)
LLM	OpenAI API / Similar
Database	MongoDB
Vector Store	ChromaDB
NLP	spaCy / Sentence-BERT
Deployment	Local / Cloud

B. Experimental Evaluation and Test Results

System evaluation was conducted across four dimensions: functional correctness, retrieval precision, response quality, and system performance, combining unit-level verification, integration testing, and a structured user acceptance trial.

Unit tests written using pytest covered authentication, document ingestion, ChromaDB collection isolation, quiz scoring logic, and W(T) computation. A total of 47 unit test cases were executed, achieving a 100% pass rate on core logic paths and 94% overall. End-to-end latency measured across 100 test queries recorded a mean of 1.87 seconds and a 95th-percentile value of 3.12 seconds.

Retrieval precision was evaluated on a validation set of 200 queries from engineering and computer science course materials. The system retrieved the correct chunk within its top-5 results for 183 of 200 queries, yielding a top-5 precision of 91.5%.

Hallucination susceptibility was assessed using 50 in-corpus queries and 20 out-of-corpus queries. For in-corpus queries, 46 responses (92%) were rated as factually consistent. The disclosure mechanism activated correctly in 18 of 20 out-of-corpus cases (90%).

A structured user acceptance trial was conducted with 12 undergraduate participants over two weeks. Results are shown in Table II.

TABLE II — USER ACCEPTANCE TRIAL RESULTS (N = 12)

Evaluation Dimension	Mean Score (/ 5)
Response relevance to uploaded materials	4.3
Clarity and comprehensibility of explanations	4.5
Usefulness of adaptive quiz generation	4.1

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

Evaluation Dimension	Mean Score (/ 5)
Accuracy of weak topic identification	3.9
Overall satisfaction with the tutoring experience	4.4

C. Discussion

The RAG pipeline demonstrated that coupling document-grounded retrieval with LLM generation produces substantially more curriculum-relevant responses than a standalone LLM. A top-5 retrieval precision of 91.5% and a factual consistency rating of 92% confirm that the system grounds responses with reliability sufficient for academic deployment. The adaptive learning mechanism revealed a limitation in its triggering logic: the W(T) formula treats all queries on a topic as equivalent evidence of difficulty, regardless of whether a query is exploratory or remedial in intent. A sentiment or intent classifier at the query level would allow the system to distinguish exploratory from remedial interactions, producing a more accurate proficiency signal.

V. RESULT

The proposed system occupies a distinct position in the educational technology landscape. Unlike general-purpose conversational AI tools, the system integrates a RAG pipeline anchoring every response to curriculum-specific content, demonstrating a factual consistency rate of 92% on in-corpus queries. Unlike static platforms such as Coursera or Khan Academy, the system maintains a dynamic learner profile that evolves continuously based on query history and quiz outcomes. Unlike classical ITS relying on hand-crafted skill taxonomies, the system derives topic proficiency estimates from natural language interaction logs through the composite scoring function W(T), making it discipline-agnostic and extensible without domain-specific re-engineering.

Performance benchmarks: mean latency to first streamed token stands at 1.87 seconds, and top-5 retrieval precision reaches 91.5%. A user acceptance study with undergraduate participants produced a mean satisfaction score of 4.4 out of 5. Qualitative responses consistently highlighted the system's capacity to answer questions directly from the student's own course materials as the primary source of its perceived value.

VI. CONCLUSION AND FUTURE WORK

This paper detailed the end-to-end development and evaluation of an AI-powered personal tutoring system bringing together Retrieval-Augmented Generation, adaptive learner modeling, and automated assessment generation into a unified educational tool. The system roots every tutoring response in documents uploaded by the student, using a semantic retrieval pipeline that ensures relevance to the specific curriculum being studied. A learner profile updated after each interaction captures topic-level engagement patterns alongside quiz outcomes, allowing the system to modulate explanation depth according to demonstrated understanding.

Evaluation confirmed a top-5 retrieval precision of 91.5%, factual consistency of 92%, and a mean satisfaction score of 4.24 out of 5. These findings support that RAG-based tutoring is technically mature for real academic deployment and that lightweight adaptive mechanisms are capable of producing meaningful personalization.

Future directions include: addition of a query intent classifier for improved weak topic detection; expansion to handle mathematical notation and STEM schematics through multimodal embedding models; domain-specific threshold calibration; incorporation of spaced repetition logic; and a semester-length longitudinal study tracking actual learning outcomes.

ACKNOWLEDGMENT

The authors would like to thank Shrimati Kashibai Navale College of Engineering, Pune, for providing the necessary resources and support for this research, and the undergraduate participants who volunteered for the user acceptance trial.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

REFERENCES

[1] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers," IEEE Access, vol. 12, pp. 69812–69840, 2024.
[2] M. Trigka and E. Dritsas, "The Evolution of Generative AI: Trends and Applications," IEEE Access, vol. 13, pp. 98504–98540, 2025.
[3] S. W. Tan, C. P. Lee, K. M. Lim, C. Tee, and A. Alqahtani, "QARR-FSQA: Question-Answer Replacement and Removal Pretraining Framework for Few-Shot Question Answering," IEEE Access, vol. 12, pp. 159280–159300, 2024.
[4] T. Falatouri, D. Hrušecká, and T. Fischer, "Harnessing the Power of Large Language Models for Service Quality Assessment from User-Generated Content," IEEE Access, vol. 12, pp. 99755–99780, 2024.
[5] V. Vaswani et al., "Attention Is All You Need," in Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[6] T. Brown et al., "Language Models Are Few-Shot Learners," in Proc. 34th Int. Conf. Neural Information Processing Systems (NeurIPS), 2020, pp. 1877–1901.
[7] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459–9474.
[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[9] B. Woolf, Building Intelligent Interactive Tutors: Student-Centered Strategies for Revolutionizing E-Learning, Morgan Kaufmann, San Francisco, CA, USA, 2010.
[10] K. VanLehn, "The Behavior of Tutoring Systems," International Journal of Artificial Intelligence in Education, vol. 16, no. 3, pp. 227–265, 2006.
[11] A. N. Kumar, "Intelligent Tutoring Systems: A Review," International Journal of Engineering Research and Technology, vol. 2, no. 11, pp. 1–6, 2013.
[12] S. D'Mello and A. Graesser, "AutoTutor and Affective Tutoring Systems," Journal of Educational Psychology, vol. 104, no. 2, pp. 304–323, 2012.
[13] Z. A. Pardos and N. T. Heffernan, "Modeling Individualization in Intelligent Tutoring Systems," User Modeling and User-Adapted Interaction, vol. 21, no. 3, pp. 205–237, 2011.
[14] A. G. Anderson et al., "Prompt Engineering for Large Language Models in Education," IEEE Transactions on Learning Technologies, vol. 16, no. 4, pp. 512–524, 2023.
[15] S. Kasneci et al., "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education," Learning and Individual Differences, vol. 103, pp. 102274, 2023.
[16] Y. Liu et al., "Personalized Learning Path Recommendation Using Deep Learning," IEEE Transactions on Education, vol. 64, no. 3, pp. 234–243, 2021.
[17] J. Gao et al., "Neural Approaches to Conversational AI in Education," IEEE Intelligent Systems, vol. 36, no. 5, pp. 64–73, 2021.
[18] R. K. Belecheanu and M. P. Drăgoicea, "Adaptive E-Learning Systems Using Artificial Intelligence Techniques," IEEE Access, vol. 9, pp. 145230–145245, 2021.
[19] E. H. Chi and J. T. Riedl, "An Introduction to Educational Data Mining," IEEE Intelligent Systems, vol. 27, no. 3, pp. 6–9, 2012.
[20] J. Kay and B. Kummerfeld, "Lifelong Learner Modeling," British Journal of Educational Technology, vol. 43, no. 1, pp. 140–150, 2012.