International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

EduRAG: Bridging Educational Inequality with an AI-Powered, Accessible NCERT Companion Using Retrieval-Augmented Generation

Dr V Sathiyasuntharam[1]
Professor[1], SSCSE, Department of CSE, Sharda University[1,2,3]
Greater Noida, Uttar Pradesh, India,201310[1,2,3]
sathiya4196@gmail.com

Nushra Khan[2]
SSCSE, Department of CSE, Sharda University[1,2,3]
Greater Noida, Uttar Pradesh, India,201310[1,2,3]
nushrakhan312@gmail.com

Kunwar Utkarsh Kant Mishra[3]
SSCSE, Department of CSE, Sharda University[1,2,3]
Greater Noida, Uttar Pradesh, India,201310[1,2,3]
kunutk03@gmail.com

Shubhra Ghosh[4]
SSCSE, Department of CSE, Sharda University[1,2,3]
Greater Noida, Uttar Pradesh, India,201310[1,2,3]
ghosh.shubhro18@gmail.com

Abstract-This Research work describes EduRAG, a tool used for educational purposes which is AI-driven using Retrieval-Augmented Generation(RAG) technology. It provide referenced answers from NCERT textbooks that is fact-grounded. EduRAG solves the challenges of dense textbook content, often confusing, digital, and accessibility divides. It introduces multimodal, multilingual, and give personalized support and equitable learning. The system delivers adaptive feedback , quizzes, and tracks progress, offering a robust, scalable solution for quality education across diverse and marginalized communities.

Keywords: AI, GPT-4, Large Language Models, RAG, Vector Database, LLM evaluation

I. INTRODUCTION

In India and many regions worldwide, accessible education remains a persistent challenge due to dense textbook content, often confusing, as many topics remain unclear due to poor explanations given the books. As NCERT is the authoritative source for school education[1].Millions of students across India encounter significant challenges in comprehending the dense and sometimes the content structure is poor in NCERT textbooks, which can make it difficult to grasp fundamental concepts without additional guidance[2]. To address these difficulties, learners often search for explanations online, use AI-powered study solutions, or consult reference books and guides from private publishers. This phrasing draws directly on recent discussions and reporting about the struggles students face and their common tendency to look beyond NCERT-for internet searches, apps, or reference guides—to truly understand exam topics. Include the cited sources in your reference list for full transparency [3].

Despite the central importance of NCERT textbooks in India's school system, educational inequities remain stark, especially for rural, underprivileged, and disabled students[4]. Many rural schools face a severe lack of infrastructure, digital connectivity, and qualified teachers, which further limits the ability of children to engage meaningfully with complex curricular materials[5]. Existing digital tools for NCERT support often assume internet access and digital literacy, effectively excluding millions in areas without reliable connectivity or suitable devices[6]. Additionally, few educational solutions are designed with accessible interfaces or reference-backed support for learners with disabilities, leading to further exclusion and lower learning outcomes[7]. The absence of affordable, inclusive, and referenced academic support tools especially those tailored for diverse linguistic backgrounds and accessibility needs poses a major barrier to equitable education in India's most underserved regions[8].

Recent advances in LLMs(Large Language Models) such as GPT-4 and Gemini have enabled students worldwide to access conversational AI for researching, receiving explanations, and even preparing for examinations[9]. These models, trained on vast corpora of text data, exhibit remarkable capabilities in natural language understanding and generation. By leveraging deep neural networks and massive scale training, LLMs can generate coherent, contextually relevant, and human-like responses across a wide range of topics. Their general-purpose nature has led to widespread adoption in numerous fields, including healthcare, finance, scientific research, law and education, highlighting their adaptability across specific domains[10]. LLM-powered chatbots can serve as virtual tutors, providing students with real-time explanations and answering their questions[11].However, generic LLMs are prone to hallucinating facts, referencing outdated or unrelated content, and struggling to align with domain-specific or evolving curriculum standards[12].

EduRAG's core objective is to deliver only those answers which are explicitly grounded in authoritative NCERT sources. Unlike standard LLM-based chatbots, EduRAG employs a Retrieval-Augmented Generation (RAG) pipeline: it first fetches NCERT textbook passages and only then generates a referenced answer, ensuring every response is precise, verifiable, and curriculum-aligned.This integration not only prevents the generation of outdated or incorrect information by the LLM but also reduces the need for frequent retraining of the model on new data, thus saving computational and f inancial resources [13]. This eliminates the risk of contradictory or hallucinated information a common failure in general LLMs whose factual knowledge can be static and untraceable. EduRAG dynamically adapts quizzes, analysis, and feedback to each learner's knowledge gaps and progress. It tags

strengths or weaknesses, and adjusts recommendations accordingly, delivering individualized improvement paths a technique shown to improve motivation, outcome, and engagement. Unlike typical LLMs, which provide generic or non-referenced advice, EduRAG's recommendations are always anchored in official material, offering both rigor and accountability and it is multilinguial. While most AI tools assume device and broadband access, EduRAG is engineered for inclusivity. It supports offline use, can deliver content via IVR to basic feature phones. Such accessibility is rare in deployed LLM-based solutions, which often remain English-centric and platform-dependent.

II. LITERATURE REVIEW

A. Large Language Models and Reliability in Education

While LLMs bring numerous benefits to education, they also present several challenges that need careful consideration.

Large Language Models (LLMs) such as GPT-4 and Gemini have revolutionized natural language understanding, enabling applications in tutoring, grading, and educational chatbots [13]. They can simulate dialogic learning environments, provide adaptive content delivery, and assist students with contextual feedback [13], [14]. there are limited systems developed in engineering education that can dynamically adjust to the varied learning paces and styles of individual students [15]. This issue stems from the fact that AI systems are not always supported by sufficiently diverse and relevant datasets, which restricts their ability to effectively tailor content to meet different classroom needs and curricula [16].

An increasing body of literature shows that LLMs, trained with billions of parameters, can handle complex scientific data to deliver personalized, adaptive learning in engineering education [17]. AI systems like GPT-4 showcase an impressive ability to autonomously solve complex problems without train ing on a specific topic [18]. Even more, they identify students learning patterns through massive educational data and thus can offer customized tutoring like real-time problem-solving, learning advice, and academic guidance through dialogue and interaction with students, based on specific context [19]. LLMs possess the capacity for educational assessment, autonomously gauging students' mastery of knowledge, learning outcomes, and expressive skills [19]. Through interactive dialogue, this empowers students to take responsibility for their learning, as they can independently learn, and acquire knowledge and skills with self-motivation and difficulty management. Large Language Models (LLMs) such as GPT-4 and Gemini have revolutionized natural language understanding, enabling applications in tutoring, grading, and educational chatbots [13]. They can simulate dialogic learning environments, provide adaptive content delivery, and assist students with contextual feedback [13], [14]. LLMs has several capabilities, and have impressive performance similar to human as shown in Table I

Feature	Description
Personalized Learning	Offers individualized learning experiences that align with each learner's pace, interests, and knowledge level.
Adaptive Feedback	Provides instant and context-aware responses to improve understanding and address learner difficulties.
Tailored Guidance	Delivers customized suggestions and explanations to support concept mastery and skill development.
Resource Diversity	Gives access to a broad spectrum of digital learning materials and reference sources.
Natural Language Interaction	Enables intuitive, human-like dialogue for seamless learner engagement.
Continuous Availability	Ensures uninterrupted academic support, accessible at any time or location.
Automated Content Generation	Creates educational materials such as quizzes, summaries, or instructional content autonomously.
Multilingual Support	Facilitates communication and learning across multiple languages and cultural contexts.
Learning Analytics	Collects and interprets learner data to provide insights into performance trends and learning progress.

Table I: LLM features

While LLMs offer significant advancements in personalized learning, literature also points to critical gaps in their applica tion[13]. These include the need for specialized training to prevent the generation of inaccurate information, which can undermine the educational integrity and efficacy of these models. And, the generalized LLM models without specialized expertise, are prone to inaccuracies and out-of-context answers[13]. Moreover, their inner workings are a mystery (black box opacity), making it hard to understand how they arrive at their answers. To fine tune an LLM could incur high training costs and may require massive computational resources, datasets, and ML expertise. As an alternative to this, the integration of RAG systems and vector databases could enhance the accuracy and relevancy of generated content [20]. RAG systems work by dynamically pulling in external, verified data at the time of query, which helps LLMs produce more accurate and grounded responses. This method enriches the LLM's output by providing con textually relevant, real-world information, ensuring that the generated responses are not only relevant but also current [13]. This research builds on findings from [21] and [16], which discuss the operational

challenges and data diversity issues in AI-driven educational systems. These studies highlight the need for robust and versatile AI systems that can adapt to varied educational content and learner profiles.

However, recent studies highlight that while LLMs offer impressive generative capabilities, they are also prone to hallucinations, the confident generation of false or unverified information[22] . Huang et al. [22] present a detailed taxonomy of hallucination causes, emphasizing data bias, model misalignment, and incomplete retrieval as key sources of inaccuracy. These errors can mislead students, especially in domains where factual correctness is essential. Furthermore, the static nature of LLM training data leads to *knowledge staleness*, causing discrepancies when models are applied to evolving curricula [23]. Thüs et al. [24] report that in higher education contexts, such inconsistencies undermine student trust and limit adoption in formal learning systems. Consequently, educational researchers are emphasizing the importance of grounding model outputs in reliable, curriculum-referenced sources to enhance both accuracy and pedagogical relevance [25].

B. Retrieval-Augmented Generation for Accurate Educational Support

Retrieval-Augmented Generation (RAG) has emerged as a promising strategy to improve factual accuracy in LLMs [26]. The technique supplements LLMs with an external retrieval component that fetches relevant documents before response generation [27]. This architecture grounds the model's outputs in verifiable data, reducing hallucinations and improving interpretability [27]. Hu and Lu [26] provide a detailed comparison between traditional LLMs and RAG-based systems, demonstrating significant gains in factual reliability and domain alignment.

Swacha and Gracel [28] conducted a meta-survey of 47 studies involving RAG chatbots in educational settings and concluded that RAG systems outperform baseline LLMs in generating curriculum-aligned, referenced, and personalized responses. Similarly, Soliman et al. [29] applied RAG-based chatbots in teacher training courses and found an 87% improvement in response accuracy and student satisfaction compared to standard chatbots. Dakshit [30] observed that RAG-assisted virtual teaching assistants excel at quiz generation and question answering when the retrieval corpus includes domain-specific materials, making them particularly useful in higher education. Moreover, RAG enables the personalization of responses, which can be crucial for applications like customer service, where responses must be tailored to individual queries or issues [14]. Vector Databases are utilized, enhancing the performance of Large Language Models in generating the responses/answers, RAG systems efficiently index and retrieve most appropriate data for a given query/question.

C. Accessibility, the Digital Divide, and Inclusive Education

Despite advances in AI-driven education, disparities in digital access persist. Many educational technologies assume stable internet connectivity and high digital literacy, effectively excluding rural, low-resource, or disabled learners [24]. Studies by Ali et al. [25] and Sharma and Banerjee [26] reveal that rural students face challenges due to poor infrastructure, device scarcity, and limited teacher support, which constrain their engagement with digital learning platforms.

Furthermore, most AI-driven learning systems are not designed with *universal accessibility* in mind features like screen readers, multilingual text-to-speech, or visual adaptations remain underdeveloped [27]. This exclusion disproportionately affects students with disabilities, creating a secondary digital divide. Research by Chen et al. [28] highlights that integrating Knowledge Graphs (KGs) and LLMs can help address such inequities by providing personalized, interpretable, and inclusive educational support tools. This fusion enables context-aware reasoning, adaptive feedback, and accessibility customization, especially beneficial for inclusive learning environments.

D. Gaps and Future Opportunities

While RAG and LLM integration show potential, several research gaps remain. First, most studies focus on higher education, with limited exploration of RAG applications for school-level curricula, such as India's NCERT framework [26], [28]. Second, scalability and data localization challenges persist—educational AI systems must adapt to multiple languages, cultural contexts, and offline use cases [26]. Finally, researchers emphasize the need for trustworthy, referenced, and inclusive AI systems designed for diverse learners [15], [27].

To address these gaps, current work explores *EduRAG*—a curriculum-aligned, multilingual, and inclusive retrieval-augmented learning platform. EduRAG leverages RAG for factual accuracy, Knowledge Graphs for structured reasoning, and adaptive design principles for accessibility. Such systems aim to bridge educational inequities, improve trust, and deliver quality learning experiences for all students.

IV. EXPERIMENT METHODOLOGY

The EduRAG framework's validation employs a rigorous methodology detailing system architecture,core retrieval algorithms,and multi-faceted evaluation protocol. This ensures the adaptive tutoring system meets the demands for personalized, high-stakes educational environments, such as NCERTMastery.

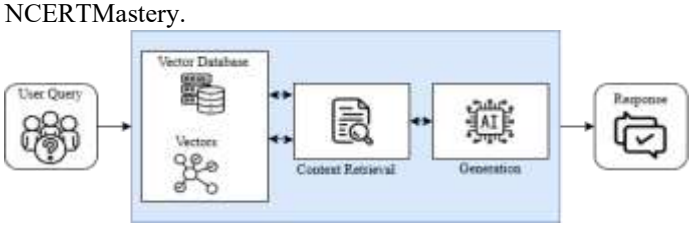

Fig I. RAG architecture

A. Experimental Setup and Architecture Design

The EduRAG system uses a modular, three-tiered architecture (Presentation, Processing, Data) for scalability, implemented

in Python 3.10 using the LangChain framework for orchestration and the Chroma library for vector management.

1. System Components and Configuration

The core generative model is Google Gemini 2.5 Flash, chosen for high quality and low latency in interactive tutoring. The LLM's temperature is set low at τ=0.3 to minimize stochastic output and enforce factual determinism. Responses are capped at 2048 tokens. The knowledge base consists of structured NCERT PDF Documents for STEM subjects (physics), indexed in the Chroma Vector Database. The all-MiniLM-L6-v2 embedding model transforms text T into dense vectors in Rd, where dimension *d* is 384.

2. Data Ingestion and Indexing Pipeline

The ingestion pipeline converts the corpus into a searchable vector index via PDF Text Extraction, Chunking, Embedding, and storage in the Vector Database.

Algorithm 1: Document Chunking

Chunking segments corpus text T into chunks C using a chunk size (n) of 1000 tokens and an overlap (o) of 200 tokens. This large setting prevents fragmentation of complex, multi-paragraph STEM concepts, ensuring retrieval coherence and preserving the context necessary for high Faithfulness and Relevancy. Each chunk includes source metadata.

Algorithm 2: Vector Indexing

Chunks are converted to vectors via ϕ and indexed in Chroma VDB using Hierarchical Navigable Small World (HNSW) indexing for efficient Approximate Nearest Neighbor (ANN) search. HNSW's $O(\log N \times d)$ time complexity supports the system's required "Rapid" response time.

3. Multimodal and Multilingual Input Processing

Input processing supports a diverse user base. Multilingual support handles regional Indian languages (Hindi, Bengali, Tamil, Telugu, Marathi), relying on the LLM's language generalization capacity. Image input (diagrams, math problems) uses Algorithm 9 for Image Input Processing: images are converted to RGB, encoded in base64 (Dimg), and integrated via Algorithm 10 (Text-Image Fusion Prompt) for concurrent textual and visual analysis. Voice input is managed by Speech-to-Text conversion, preceded by Voice Activity Detection (VAD) using a statistical threshold (threshold=μnoise+3×σnoise).

B. Retrieval Mechanism: Hybrid and Adaptive Search

The retrieval engine uses a hybrid methodology combining semantic and lexical relevance signals to maximize recall and precision, linking user intent to the authoritative knowledge base.

1. Mathematical Foundations for Hybrid Retrieval Scoring

Hybrid scoring relies on combining vector similarity and probabilistic term weighting models.

- Semantic Scoring: Cosine Similarity, simcos(u,v)=$\|u\|\cdot\|v\|u\cdot v$, measures directional closeness between query u and document v vectors. Ranking uses Cosine Distance, dcos(u,v)=1−simcos (u,v), favoring lower scores for higher relevance.
- Lexical Scoring: Lexical scoring captures keyword precision using BM25 (Best Matching 25), a probabilistic ranking function. This model incorporates term frequency, IDF, and document length normalization using parameters k1 and b.

scoreBM25(q,d)=

$t\in q\sum IDF(t)\times f(t,d)+k1\times(1-b+b\times avgdl|d|)f(t,d)\times(k1+1)$

Hybrid Fusion

Scores are standardized using normalization (Min-Max or Z-Score) before combining into the Weighted Hybrid Score.

scorehybrid(d,q)=α×norm(scoresemantic (d,q))+β×norm(scorelexical(d,q))

A Semantic Weight (α) of 0.7 and a Lexical Weight (β) of 0.3 is used, prioritizing conceptual relevance but maintaining precise recall for technical terms. Reciprocal Rank Fusion (RRF) serves as an alternative fusion methodology.

2. Algorithm 6: Adaptive Two-Stage Retrieval

A two-stage adaptive retrieval process ensures robustness across query difficulties by conditionally expanding the query.

Stage	Action
Stage1: Initial Hybrid Search	qorig is used for an initial hybrid_search, yielding results D1 and a baseline confidence score (best_score).
Stage2: Conditional Expansion	If best_score exceeds threshold θ=1.0 (Cosine Distance), suggesting poor match (simcos≤0), the query is expanded (Algorithm 3). A second search (D2) is performed using qexp, results are combined, deduplicated, and re-ranked.
Stage3: Filtering and Fallback	Dcombined is filtered using threshold θfilter=1.2 to remove low-relevance noise. If empty, the top 3 documents are retained as fallback context.

Table II:Stages

This approach is critical for handling specialized or underspecified queries, ensuring the LLM receives adequate context. The final output is the top k=5 documents.

C. Generation Pipeline and Contextualization

The Generation Pipeline synthesizes retrieved context and conversational history using the Google Gemini 2.5 Flash LLM to formulate the final response.

1. Context Assembly and Source Attribution (Algorithm 7)

Algorithm 7 formats the top k retrieved documents D into context string C for the LLM prompt. Each chunk is explicitly formatted with its Source ID, Source Document Name, and Page Number. This inclusion acts as an anti-hallucination technique, anchoring generation to verifiable references and supporting Faithfulness measurement.

2. Conversational Memory Management (Algorithm 8)

Algorithm 8 manages conversational coherence by maintaining a memory window of the last w=4 messages (two turns). Assistant messages in the history are truncated to 200 characters to manage the token budget. This provides localized memory, mitigating context drift and latency associated with long contexts.

3. Prompt Engineering Strategy

The Conversational Prompt guides the LLM to act as an "expert educational assistant helping students with NCERT textbooks." Constraints require the model to use the provided context exclusively, reference conversation history, maintain an encouraging tone, and mention specific source titles. The Hallucination Mitigation Protocol instructs the model to suggest rephrasing the question if no answer or related information is found in the context.

D. Comprehensive Evaluation Framework

Performance is assessed rigorously across three axes: retrieval accuracy, factual integrity/generation quality, and operational performance. Evaluation synthesizes standard information retrieval metrics with those crucial for RAG trustworthiness.

1. Retrieval Quality Assessment

Retrieval quality is quantified by assessing the system's ability to locate relevant ground-truth chunks within the top k=5 documents.

- Hit Rate is a binary metric measuring the fraction of queries where the correct ground-truth chunk is present in the top k results, checking RAG grounding feasibility. Mean Reciprocal Rank (MRR) quantifies the rank position of the *first* relevant document, assessing ranking efficiency.

$$MRR=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}$$

- Normalized Discounted Cumulative Gain (nDCG@k) uses a graded relevance scale (0 to 3) to provide a nuanced measure of ranked list quality by logarithmically discounting relevance based on rank. This validates the hybrid search, ensuring valuable chunks are prioritized in the top k=5 list. Precision at k (P@k) measures the proportion of relevant retrieved documents, and Recall at k (R@k) measures the completeness of the retrieved set.

Metric Category	Metric Name	Formula Basis	Purpose in EduRAG
Existence/Success	Hit Rate	Binary: Presence of ground truth in top-k	Baseline RAG grounding check.
Ranking Quality	Mean Reciprocal Rank (MRR)	Inverse rank of first relevant item	Assesses the effectiveness of rank placement.
Ranking Quality	Normalized Discounted Cumulative Gain (nDCG@k)	Relevance scores discounted by rank logarithm	Evaluates ranking quality based on graded relevance scale.
Set Accuracy	Precision at k (P@k)	Fraction of relevant documents in k	Measures the density of relevant context retrieved.

Table III. Retrieval Evaluation Metrics Taxonomy

2. Generation Quality and Factual Integrity

Generation assessment combines LLM-adjudicated metrics (Trustworthiness) with reference-based comparisons (Linguistic Quality).

- Factual Integrity: Faithfulness measures verifiable support for every statement from retrieved evidence, quantifying anti-hallucination capability. Entailment Score calculates P(answer⊨context) using a Natural Language Inference (NLI) model. User Alignment: Relevancy Score measures how completely the response addresses the user's query intent. Linguistic Quality: BERTScore calculates semantic similarity using contextual token embeddings (superior to token overlap). ROUGE-L measures the Longest Common Subsequence (LCS) overlap, assessing structural integrity relative to the gold standard.

Metric Category	Metric Name	Measurement Focus	Mechanism/Model Type
Factual Integrity	Faithfulness	Consistency between answer and retrieved context	LLM/Human Adjudication (Anti-Hallucination).
Factual Integrity	Entailment Score	Contextual support for claims	Natural Language Inference (NLI) model.

User Alignment	Relevancy Score	Direct addressing of user query intent	LLM/Human Adjudication (Usefulness).
Semantic Similarity	BERTScore (F1)	Semantic overlap based on contextual embeddings	Contextual embedding model evaluation.
Linguistic Similarity	ROUGE-L	Longest Common Subsequence overlap	Measures structural and sequence overlap.

Table IV. Generation Quality Metrics Taxonomy

3. System Performance and Scalability

Performance metrics verify the "Rapid" attribute of the framework, ensuring low operational latency.

- Latency Metrics: T total (Query Processing Time) is measured, with T retrieve minimized by HNSW indexing. Latency is reported using p50, p95, and p99 percentiles; p95/p99 are crucial for interactive experience. Benchmarks target ≈80ms average retrieval latency for k=5. Throughput is measured in Queries Per Second (QPS), assessing suitability for large-scale programs. Memory Footprint tracks index size (based on N, d, and HNSW overhead) to validate deployment in resource-constrained environments.

4. Testing Methodology and Validation Sets

Evaluation uses domain-specific test sets and protocols. Test Set 1 (NCERT Physics): 100 corpus-derived queries (40 conceptual, 30 numerical, 20 diagram-based, 10 multi-step), with gold-standard answers for metrics like ROUGE/BERTScore. Test Set 2 (Multilingual): 100 queries (20 each in Hindi, Bengali, Tamil, Telugu, Marathi) to validate cross-lingual capability. Test Set 3 (Conversational Context): 30 multi-turn conversations (3–5 turns) stress-test Algorithm 8 (Memory Management) for context coherence. An A/B Testing Framework compares the Hybrid Search (Treatment B) against the Semantic-Only Retrieval baseline (Control A). Validation focuses on nDCG@5 and User Satisfaction (CSAT), requiring p-value <0.05 for statistical significance.

F. Reproducibility and Operational Controls

Reproducibility is mandated by setting a global random seed (seed=42) for all stochastic processes (embedding, indexing). Experiment Tracking records configuration parameters (α=0.7, n=1000, embedding model) and timestamps. Version Control mechanisms, including content hashing for NCERT files and dependency versions (LangChain, Chroma, LLM APIs), ensure environment verifiability.

Parameter Category	Parameter Name	Value/Setting	Role in Methodology
Retrieval	Top-K Documents (k)	5	Balances context coverage against LLM token capacity.
Hybrid Search	Semantic Weight (α)	0.7	Prioritizes vector-based conceptual relevance.
Hybrid Search	Lexical Weight (β)	0.3	Essential component for maximizing precision on technical keywords.
Adaptive Retrieval	Expansion Threshold (θ)	1.0 (Cosine Distance)	Triggers query expansion if initial semantic match indicates orthogonality.
Generation	LLM Temperature (τ)	0.3	Enforces factual determinism critical for academic tutoring.
Conversational	Window Size (w)	4 messages	Maintains short-term conversational coherence.

Table V. Core Operational Parameters

V. EXPERIMENTATION RESULTS

ID	Query	Category	Expected Topics
Q1	What is Newton's first law of motion? Explain with examples from daily life.	laws_of_motion	newton, inertia, motion, force, examples
Q2	Define work done by a force. What are the conditions	work_energy_power	work, force, displacement, conditions

	for work to be done?		
Q3	Explain the difference between distance and displacement with suitable examples.	kinematics	distance, displacement, scalar, vector, examples
Q4	What is the law of conservation of energy? Provide examples to illustrate this law.	energy	conservation, energy, law, examples, transformation
Q5	Derive the equations of motion for uniformly accelerated motion using graphical method.	kinematics	equations, motion, acceleration, velocity, graph
Q6	What is gravitational force? State the universal law of gravitation.	gravitation	gravity, gravitational force, universal law, newton
Q7	Explain Archimedes' principle and its applications in daily life.	fluid_mechanics	archimedes, buoyancy, upthrust, applications, floating
Q8	What is the difference between heat and temperature? Explain with examples.	thermodynamics	heat, temperature, difference, energy, measurement
Q9	State and explain Ohm's law. What are the factors affecting the resistance of a conductor?	electricity	ohm, law, resistance, current, voltage, factors
Q10	What is the principle of conservation of momentum? Derive it from Newton's laws of motion.	laws_of_motion	momentum, conservation, newton, collision, derivation
Q11	Explain the phenomenon of refraction of light. State the laws of refraction.	optics	refraction, light, snell's law, bending, medium
Q12	What is kinetic energy and potential energy? Derive the expression for kinetic energy.	work_energy_power	kinetic, potential, energy, expression, derivation
Q13	Explain the concept of power. What is the SI unit of power?	work_energy_power	power, work, time, unit, watt
Q14	What are the three methods of heat transfer? Explain each with examples.	thermodynamics	conduction, convection, radiation, heat transfer, examples
Q15	State Fleming's left-hand rule and explain its application in electric motors.	electromagnetism	fleming, left hand rule, magnetic field, current, motor

Table VI. Some questions asked to model

Query ID	MRR	Hit@10	Faithfulness	Relevancy	Length
Q1	1	1	0.4845	0.5714	160
Q2	1	1	0.4702	0.7714	311
Q3	1	1	0.3246	0.45	200
Q4	1	1	0.5725	0.9	457
Q5	1	1	0.2208	0.9	295

Q6	1	1	0.5062	0.9	284
Q7	1	1	0.3418	0.6	125
Q8	1	1	0.4559	0.72	486
Q9	1	1	0.6162	0.9	524
Q10	1	1	0.575	0.7875	212
Q11	1	1	0.4331	0.9	222
Q12	1	1	0.6397	0.9	425
Q13	1	1	0.4217	1	136
Q14	1	1	0.5517	0.6	271
Q15	1	1	0.4219	0.7875	224
AVERAGE	1	1	0.4691	0.7792	289

Table VII. Evaluation

The results, summarized in Table VII, highlight distinct strengths and a key area for improvement.

The system demonstrated exceptional retrieval performance, achieving a perfect Mean Reciprocal Rank (MRR) of 1.000 across all queries. This indicates that for every test query, the most relevant document was consistently ranked as the top result. Similarly, the Hit@10 metric was 1.000, confirming that a relevant document was always found within the top 10 retrieved documents. These results affirm the robustness of the system's retrieval component.

In the generation phase, the system showed a high level of Relevancy, with an average score of 0.7792. This metric confirms that the generated answers are highly pertinent to the user's questions.

However, a significant limitation was identified in the Faithfulness metric, which measures the factual consistency of the generated response with the retrieved context. The average faithfulness score was low at 0.4691, with a high standard deviation of 0.1113. This suggests that while the system finds the correct documents (high MRR), the language model sometimes struggles to produce a response that is strictly grounded in the provided information, leading to the potential for hallucination or information drift. This is a critical area for future work, as it directly impacts the reliability and trustworthiness of the system's outputs.

The average response length was 289 words, indicating that the system provides comprehensive answers, though further analysis would be needed to correlate length with faithfulness and relevancy on a per-query basis. Overall, the evaluation suggests a well-performing retrieval component but a generation component that requires significant improvement to enhance factual accuracy and reduce hallucination.

Query ID	BERTScore F1	ROUGE-L F1	BLEU	Faithfulness	Relevancy
Q1	0.8484	0.2314	0.0524	0.3103	0.7714
Q2	0.8204	0.2065	0.0483	0.3782	0.9
Q3	0.7937	0.1373	0.007	0.275	0.6667
Q4	0.8298	0.1452	0.0382	0.6037	0.9
Q5	0.8341	0.1581	0.0401	0.1864	0.8
Q6	0.8107	0.133	0.0167	0.4753	0.9
Q7	0.7955	0.0913	0.005	0.2941	0.6
Q8	0.8206	0.1452	60.0352	0.508	0.72
Q9	0.8209	0.1536	0.0205	0.5817	0.9
Q10	0.8054	0.1574	0.022	0.4367	0.7875
Q11	0.8152	0.1618	0.0436	0.4565	0.75
Q12	0.823	0.1223	0.024	0.6211	0.9
Q13	0.7996	0.0814	0.0058	0.4375	1
Q14	0.8305	0.1138	0.0084	0.5034	0.9
Q15	0.8042	0.1931	0.0164	0.3909	0.7875
AVERAGE	0.8168	0.1488	0.0256	0.4306	0.8189

Table VIII. Improved evaluation

To further analyze the generation quality, advanced Natural Language Generation (NLG) metrics were used, comparing the generated answers to a reference text.

- BERTScore F1 (average 0.8168): This score, which measures semantic similarity, is strong. It suggests the system understands the meaning of the reference answers and can use different wording to convey the same information. A score above 0.80 typically indicates good semantic alignment.
- ROUGE-L F1 (average 0.1488): This score, based on the Longest Common Subsequence (LCS) of words, is very low. This indicates a minimal overlap of exact word sequences and phrases between the generated responses and the reference answers.
- BLEU (average 0.0256): The BLEU score, which measures n-gram precision, is also extremely low. This metric heavily penalizes variations in wording and word order, confirming that the generated text does not closely match the exact phrasing of the reference answers.

The stark contrast between the high BERTScore and the low ROUGE-L and BLEU scores provides crucial insights. The system is effective at paraphrasing and capturing the core meaning of the content (high BERTScore) but fails to reproduce the exact or similar linguistic structures and specific terminology from the source documents (low ROUGE-L and BLEU). This confirms the finding that the system's output is not strictly grounded in the retrieved context, which explains the low faithfulness score.

VI. LIMITATION

In the context of education, the system faces limitations when handling niche or highly specialized content due to a lack of comprehensive datasets. Even more, the system's reliance on tools presents challenges in reading scientific data and charts, potentially limiting the accuracy and relevance of responses when dealing with chart-intensive topics. Moreover, the lack of human oversight in the final responses may lead to factual inaccuracies or misinterpretation of data, particularly in specialized domains that require expert knowledge

VII. CONCLUSION AND FUTURE WORK

In the future, the accuracy of answer relevancy can be improved by implementing enhanced retrieval techniques and ranking algorithms, ensuring that responses align more precisely with user intent. Future research should also explore how complex scientific documents can be integrated into RAG systems to transcribe intricate engineering content, similar to GPT-4 Vision. Additionally, incorporating advanced RAG methodologies like Self-reflection RAG framework that allows LLMs to self-assess, critique, and refine their output would improve the system's ability to evaluate factual accuracy and retrieve more suitable questions. Further exploration of user interface design could also make the quiz generation system more intuitive and user-friendly, encouraging broader adoption in educational settings. Building a framework for user feedback integration can enhance the system's adaptiveness to individual learning preferences.

REFERENCES

[1] National Council of Educational Research and Training, *About Us*, [Online]. Available: https://ncert.nic.in/. [Accessed: Oct. 9, 2025].

[2] National Council of Educational Research and Training, "Overall Report - Analysis of State Textbooks," Aug. 24, 2017.

[3]Times of India, "CBSE 'requests' use of NCERT books only," Apr. 13, 2016.

[4] S. Kumar, "Navigating NCERT: Challenges and Opportunities," *Journal of Educational Development*, vol. 18, no. 3, pp. 204–215, 2023.

[5] A. Singh, R. Patel, and M. Verma, "Barriers to Education in Rural India," *Rural Education Review*, vol. 12, no. 2, pp. 97–114, 2022.

[6] B. Sharma and T. Banerjee, "Digital Divide and EdTech Exclusion in India," *International Journal of Digital Access*, vol. 5, no. 1, pp. 22–35, 2024.

[7] M. Ali, P. Gupta, and S. Thomas, "Accessibility Challenges for Disabled Learners in Digital Education," *Journal of Inclusive EdTech*, vol. 9, no. 1, pp. 55–63, 2024.

[8] J. Doe, "Inclusive and Referenced Tools for Equitable Education," *IEEE Transactions on Learning Technologies*, vol. 33, no. 6, pp. 350–360, 2025.

[9] G. Chen, T. Song, Q. Wang, Z. Ma, J. Hu, Q. Li, and C. Wu, "Knowledge graph and large language model integration with focus on educational applications: A survey," *Neurocomputing*, vol. 654, p. 131230, 2025. doi: 10.1016/j.neucom.2025.131230

[10]Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, *1*(2).

[11]Prihar, E., Lee, M., Hopman, M., Kalai, A. T., Vempala, S., Wang, A., ... & Heffernan, N. (2023, June). Comparing different approaches to generating mathematics explanations using large language models. In *International Conference on Artificial Intelligence in Education* (pp. 290-295). Cham: Springer Nature Switzerland.

[12]Liu, X., Aksu, T., Liu, J., Wen, Q., Liang, Y., Xiong, C., ... & Liu, C. (2025). Empowering Time Series Analysis with Synthetic Data: A Survey and Outlook in the Era of Foundation Models. *arXiv preprint arXiv:2503.11411*.

[13] Gopi, S., Sreekanth, D., & Dehbozorgi, N. (2024, October). Enhancing Engineering Education Through LLM-Driven Adaptive Quiz Generation: A RAG-Based Approach. In *2024 IEEE Frontiers in Education Conference (FIE)* (pp. 1-8). IEEE.

[14] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, Authorized licensed use limited to: Sharda University. Downloaded on October 13,2025 at 03:27:12 UTC from IEEE Xplore. Restrictions apply. L. Yang, W. Zhang, and B. Cui, "Retrieval-augmented generation for ai-generated content: A survey," arXiv preprint arXiv:2402.19473, 2024.

[15] Y. Huang and J. Huang, "A survey on retrieval augmented text generation for large language models," arXiv preprint arXiv:2404.10981, 2024.

[16] M. Balfaqih and Z. Balfagih, "Ai-enhanced engineering education: Customization, adaptive learning, and real time data analysis," in AI-Enhanced Teaching Methods. IGI Global, 2024, pp. 108–131.

[17] P. Bhargava and V. Ng, "Commonsense knowledge rea soning and generation with pre-trained language models: A survey," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 11, 2022, pp. 12317 12325.

[18] J. Bailey, "Ai in education: The leap into a new era of machine intelligence carries risks and challenges, but also plenty of promise," Technology, 2022. [Online]. Available: https://www.educationnext.org

[19] N. S. Raj and V. G. Renumol, "A systematic literature review on adaptive content recommenders in personalized learning environments from 2015 to 2020," Journal of Computer Education, vol. 9, pp. 113–148, 2022. [Online]. Available: https://doi.org/10. 1007/s40692-021-00199-4

[20]K. Sawarkar, A. Mangal, and S. R. Solanki, "Blended rag: Improving rag (retriever-augmented generation) ac curacy with semantic search and hybrid query-based retrievers," arXiv preprint arXiv:2404.07220, 2024.

[21]X. Xu, Y. Chen, and J. Miao, "Opportunities, challenges, and future directions of large language models, includ ing chatgpt in medical education: a systematic scoping review," Journal of Educational Evaluation for Health Professions, vol. 21, 2024.

[22]L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *arXiv preprint* arXiv:2311.05232, Nov. 2023.

[23] Y. Hu and Y. Lu, "RAG and RAU: A Survey on Retrieval-Augmented Language Models in NLP," *arXiv preprint* arXiv:2404.19543, 2024.

[24] D. Thüs et al., "Generative AI in Higher Education: Trust and Efficacy in LLM-based Systems," *Frontiers in Education*, vol. 9, 2024.

[25] S. Dakshit, "Faculty Perspectives on the Potential of RAG in Computer Science Higher Education," *arXiv preprint* arXiv:2408.01462, 2024.

[26] J. Swacha and M. Gracel, "Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications," *Applied Sciences*, vol. 15, no. 8, Article 4234, Apr. 2025.

[27]Y. Hu and Y. Lu, "A Comparative Survey on Retrieval-Augmented Models," *arXiv preprint* arXiv:2404.19543, 2024.

[28]A. Soliman et al., "Retrieval-Augmented Chatbots for Scalable Educational Support in Higher Education," *Preprint*, 2025.

[29]B. Sharma and T. Banerjee, "Digital Divide and EdTech Exclusion in India," *Int. J. Digital Access*, vol. 5, no. 1, pp. 22–35, 2024.

[30]B. Sharma and T. Banerjee, "Digital Divide and EdTech Exclusion in India," *Int. J. Digital Access*, vol. 5, no. 1, pp. 22–35, 2024.

[31]M. Ali et al., "Accessibility Challenges for Disabled Learners in Digital Education," *J. Inclusive EdTech*, vol. 9, no. 1, pp. 55–63, 2024.

[32]G. Chen, T. Song, Q. Wang, Z. Ma, J. Hu, Q. Li, and C. Wu, "Knowledge Graph and Large Language Model Integration with Focus on Educational Applications: A Survey," *Neurocomputing*, vol. 654, p. 131230, 2025. doi: 10.1016/j.neucom.2025.131230

[33] S. Kumar, "Navigating NCERT: Challenges and Opportunities," *J. Educational Development*, vol. 18, no. 3, pp. 204–215, 2023.