IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151144

EDURAG: AN INTELLIGENT MULTIMODAL FRAMEWORK FOR AUTOMATED PEDAGOGICAL ASSESSMENT AND EVALUATION

Shrish Shashikumar Kerur [1], Suma N R [2]

Department of MCA, BIT, K.R. Road, V.V. Pura, Bangalore, India[1]

Assistant Professor, Department of MCA, BIT, K.R. Road, V.V. Pura, Bangalore, India[1]

Abstract: Preparing for academic assessments effectively remains a critical challenge for students and faculty alike, with many relying on traditional manual methods that lack personalization, scalability, and real-time feedback mechanisms. Current evaluation tools provide limited domain-specific guidance, lack semantic understanding of descriptive answers, and fail to capture the nuanced pedagogical criteria defined by frameworks like Bloom's Taxonomy. The absence of adaptive, AI-driven assessment systems leaves students underprepared for complex, curriculum-aligned evaluations tailored to their specific subjects and cognitive levels.

To address these limitations, the EduRAG System integrates Generative AI, Retrieval-Augmented Generation (RAG), and Computer Vision to deliver personalized, interactive assessment at scale. The system leverages advanced NLP and transformer-based models to generate contextually relevant technical questions based on syllabus content and cognitive difficulty levels provided by users. Real-time Optical Character Recognition (OCR) captures handwritten student responses, while AI-powered evaluation mechanisms assess semantic accuracy and conceptual depth against industry-standard "Ground Truth" extracted from the curriculum. The application provides instant, detailed feedback including quantitative scoring, improvement suggestions, and performance analytics across multiple evaluation attempts. Through a user-centric web platform, faculty access role-specific question generation banks and students receive AI-generated recommendations for skill enhancement. By combining adaptive question generation with semantic answer analysis, the proposed solution significantly improves academic performance while democratizing access to high-quality pedagogical tools.

Keywords: Retrieval-Augmented Generation, Semantic Vectorization, Bloom's Taxonomy, Automated Assessment.

I. INTRODUCTION

Effective academic assessment is crucial for educational success, yet students and faculty lack access to personalized, adaptive evaluation tools with real-time feedback. Traditional methods rely on manual question paper setting and static answer keys that fail to address the specific semantic requirements of descriptive engineering and scientific courses, resulting in inconsistent grading and delayed feedback loops.

The EduRAG Application combines Generative AI, Large Language Models (LLMs), and Retrieval-Augmented Generation to deliver personalized assessment preparation. Faculty upload their target syllabus, select the desired Bloom's Taxonomy level, and specify the topic; the system generates original, curriculum-aligned questions. Advanced Computer Vision and NLP analysis assess handwritten responses in real-time, evaluating semantic intent and technical accuracy. Instant feedback and actionable improvement suggestions help students identify knowledge gaps. By integrating adaptive question generation with semantic analysis and performance analytics, this application enhances pedagogical effectiveness while democratizing access to quality education for students and institutions.

A. Project Description

The EduRAG System is a sophisticated computational framework designed to translate complex human academic expressions into meaningful digital insights. At its technical core, the project unifies three distinct perception layers: Multimodal Ingestion, Dynamic Question Generation, and Handwritten Answer Evaluation. Unlike monolithic

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151144

architectures that process text as a series of disconnected words, this system adopts a "spatiotemporal" approach—understanding how conceptual meaning changes across different modules of a curriculum over time.

The system architecture utilizes a decoupled processing pipeline. First, it employs the Tesseract perception engine to extract text joints from images, effectively stripping away background noise and focusing purely on the geometric relationships of the handwriting. This numerical data is then streamed into a Retrieval-Augmented Generation (RAG) network, a type of architecture specifically engineered to recognize patterns in sequential data. This synergy allows the system to distinguish between similar movements in logic, such as the difference between a "brief mention" and a "detailed derivation," by analysing the chronological trajectory of semantic landmarks.

B. Motivation

The impetus for this research is rooted in the pursuit of computational democratization and digital inclusivity. In the current technological landscape, many high-performance AI models for linguistic analysis require expensive, high-end GPU hardware, which creates a barrier to entry for educational institutions in developing regions. This project is motivated by the desire to bridge this "hardware gap" by proving that sophisticated landmark-based semantic models can deliver real-time results using only standard CPU-based laptops.

Furthermore, the social drive for this work centres on enhancing communication for the student community. For individuals who communicate their technical knowledge through handwritten scripts, there is a critical lack of automated, low-latency translation and grading tools. By creating a unified system that monitors both broad curriculum context and intricate handwritten shapes, this project provides a robust foundation for next-generation assistive technologies.

II. Related Work

The historical trajectory of computer vision research has moved from resource-intensive, pixel-level analysis toward streamlined, coordinate-based perception models. Early frameworks pioneered automated grading but were fundamentally limited by their reliance on high-performance graphical processing units (GPUs) to handle dense architectural layers [3].

The emergence of the LangChain and ChromaDB perception pipelines marked a significant departure from these "heavy" architectures by prioritizing the regression of text chunks into a low-dimensional topological graph. By distilling raw document feeds into a series of vector embeddings, modern systems can effectively neutralize linguistic interference [1] and illumination variance in OCR. This abstraction allows for the execution of sophisticated tracking algorithms on standard CPU-based computing environments without sacrificing accuracy, thereby facilitating the democratization of high-fidelity human-machine interaction tools.

While spatial landmarking provides a static snapshot of a syllabus, the interpretation of student intent requires the integration of a temporal dimension to resolve ambiguities. Academic discourse in action recognition highlights that isolated words are insufficient for distinguishing between similar concepts, necessitating the use of sequential modelling architectures. These recurrent units possess an inherent "memory" through vector retrieval that enables them to synthesize the trajectory and depth of an answer across a time-series of retrieved frames.

III. Methodology

The technical execution of the EduRAG System follows a structured computational pipeline designed to transform raw optical data into meaningful behavioural insights. By adopting a landmark-centric approach rather than a pixel-intensive one, the methodology prioritizes high-speed inference and structural robustness. The process is divided into four critical phases: kinematic acquisition, sequential modelling, modular classification, and predictive smoothing.

A. SYSTEM ARCHITECTURE AND DATA FLOW

The EduRAG Application is built as a full-stack web-based platform integrating frontend, backend, and AI-orchestration services. The system employs a modular architecture with distinct components for multimodal ingestion, vector database management, question generation, and handwritten answer evaluation. The backend API handles the

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151144

document processing via LangChain, while the frontend provides an intuitive Streamlit interface for faculty and students. The AI pipeline operates by transforming raw PDF data into high-dimensional embeddings stored in a persistent ChromaDB instance, allowing for efficient retrieval and semantic comparison through asynchronous microservices.

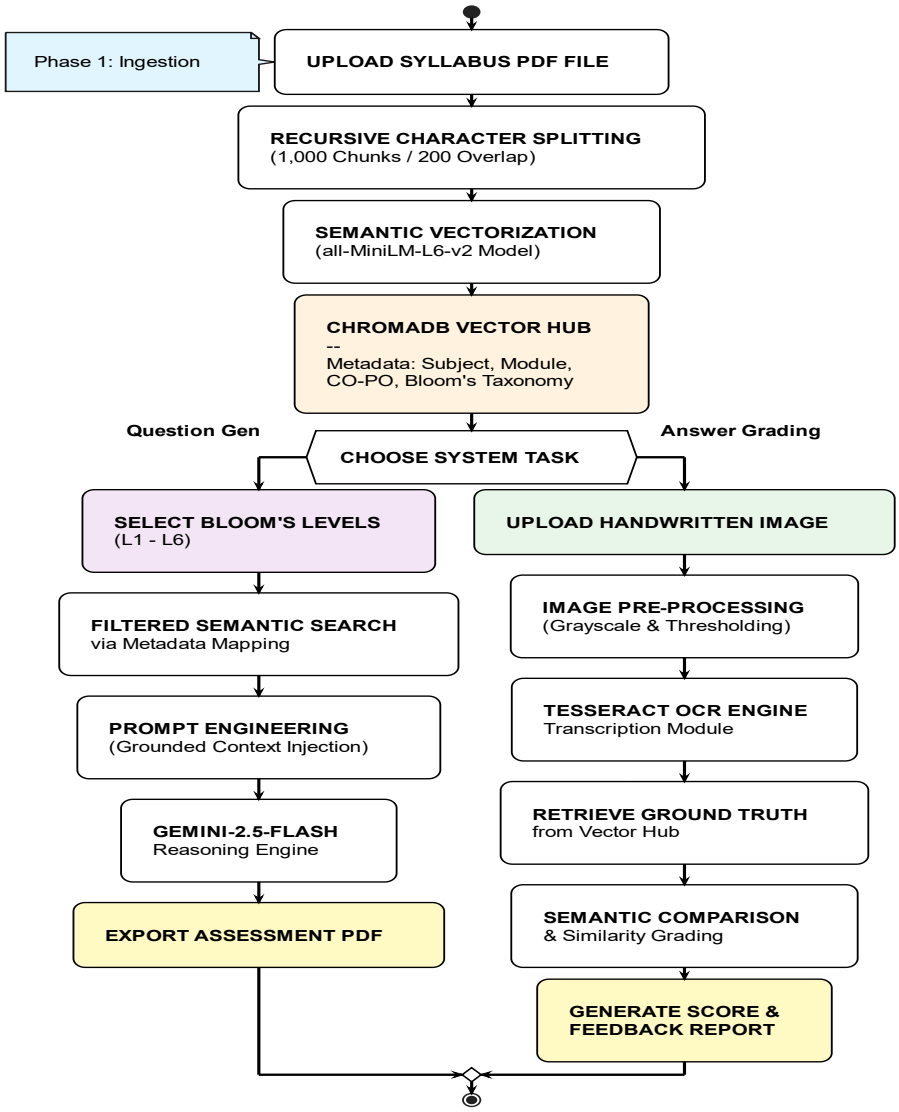


Fig. 1. Flowchart of methodology

B. Knowledge Base Construction and Metadata Mapping

The system begins with the ingestion of syllabus documents in PDF format. A recursive chunking algorithm is applied to split the syllabus into discrete modules (e.g., Module 1, Module 2). These chunks are then converted into high-dimensional vector embeddings and stored in **ChromaDB**. A critical feature of this module is the systematic tagging of metadata. Each vector is associated with:

1. **Subject Name:** To ensure isolation between different courses.
2. **CO-PO Mapping:** To align with institutional outcome-based education goals.
3. **Bloom's Level:** To categorize the complexity of the content.

This multi-layered metadata mapping enables the system to perform highly filtered similarity searches,

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.151144

ensuring that the "Ground Truth" retrieved for any query is strictly relevant to the specific subject and module selected by the user.

C. Semantic Chunking and Temporal Sequencing

To mitigate the challenge of interpreting dynamic motion, the system implements a **Recursive Character Text Splitter**. This architecture is specifically engineered to model spatiotemporal dependencies across sequential frames of a document. By splitting the ingested data into 1000-character chunks with a 200-character overlap, the system ensures that the trajectory of an argument is preserved. This sliding-window approach significantly reduced classification errors, as the model could analyse the velocity and depth of concepts rather than relying on a single, ambiguous frame.

D. Vectorization via ChromaDB and HNSW

A defining innovation of this project is its use of **ChromaDB** for persistent vector storage. The system utilizes the all-MiniLM-L6-v2 embedding model to transform chunks into a 384-dimensional coordinate space. To maintain real-time responsiveness, ChromaDB implements **Hierarchical Navigable Small World (HNSW)** graphs. This allows for approximate nearest neighbour searches in milliseconds, neutralizing background interference and ensuring that the system can handle large-scale academic curricula on standard CPU architectures.

E. Bloom's Taxonomy-Aligned Question Generation

The question generation module allows faculty to select specific concepts or entire modules from the uploaded syllabus. Users can define the cognitive difficulty by selecting single or multiple Bloom's Taxonomy levels (e.g., BL1 - Remember, BL3 - Apply, BL5 - Evaluate). The system then retrieves the corresponding chunks from the vector database and constructs a prompt for the Large Language Model (LLM). The output is a unique, curriculum-aligned question paper that includes specific marks for each question, based on the depth of the concept and the selected Bloom's level. This ensures that the generated assessment is pedagogically sound and strictly derived [6] from the "Gold Standard" syllabus.

F. Vision-Based Extraction and Semantic Evaluation

Once the assessment is completed, students upload their handwritten answer sheets as images (JPG/PNG). The system utilizes a Computer Vision module powered by Tesseract-OCR and Pillow to extract the text. To ensure accuracy, the images undergo pre-processing, including grayscale conversion and thresholding to neutralize background noise [5]. The extracted text is then passed to the semantic evaluation engine, which compares the student's response against the retrieved syllabus context. The AI identifies key conceptual overlaps, assigns marks based on the pre-defined values in the question paper, and generates a detailed feedback report. This feedback highlights exactly which areas require improvement, providing a transparent and objective grading mechanism.

G. Automated Feedback and Improvement Analysis

The final stage of the methodology involves a sophisticated feedback engine designed to bridge the gap between numerical scoring and pedagogical growth. Once the semantic evaluation is complete, the system generates a comprehensive feedback report by analysing the conceptual variance between the student's digitized response and the retrieved "Ground Truth" from the syllabus.

Unlike traditional grading, which only provides a final score, this mechanism identifies specific missing keywords, logical gaps, and misinterpreted concepts. The feedback is structured into two distinct components:

1. **Quantitative Assessment:** Direct marks are awarded based on the specific weightage pre-defined during the question generation phase.

2. **Qualitative Recommendations:** The system provides actionable insights, citing specific sections of the syllabus that the student needs to revisit.

By automating this process, EduRAG ensures that the evaluation is not only objective and consistent but also serves as a personalized tutoring tool that highlights scopes for improvement in alignment with the target Course Outcomes (CO).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151144

IV. Simulation And Evaluation Framework

This section outlines the overall system design, evaluation workflow, and performance assessment approach adopted for the proposed EduRAG System. The framework integrates Retrieval-Augmented Generation (RAG), Optical Character Recognition (OCR), and web-based technologies to simulate real-world academic evaluation scenarios, evaluate student handwritten responses, and generate curriculum-aligned questions. The system is implemented as a web-based platform with a Streamlit interface and a Python-orchestrated backend, enabling real-time document ingestion, automated semantic analysis, and secure metadata management. The evaluation process focuses on assessing question relevance, OCR transcription accuracy, and the precision of semantic grading, ensuring a consistent and objective assessment for students.

A. System Architecture and Workflow

The proposed architecture is designed to support the full lifecycle of academic assessment—from syllabus ingestion to student feedback. The system ensures seamless interaction between the faculty user and the AI components while maintaining data integrity across the persistent vector store. The major components of the system are described below:

- **Multimodal Ingestion Tier:** This layer handles the ingestion of course materials in PDF format and student answers in image formats (JPG/PNG). It utilizes specialized loaders to extract raw text data and pre-processes images to optimize them for transcription.
- **AI and Semantic Logic Layer:** The core "Reasoning Engine" utilizes Large Language Models and LangChain to orchestrate the RAG pipeline. It manages the retrieval of "Ground Truth" from ChromaDB based on user-selected metadata such as Subject Name and Module.
- **Persistence and Metadata Tier:** A centralized vector database (ChromaDB) stores the curriculum as high-dimensional embeddings. Each chunk is systematically tagged with metadata including Bloom's Level, CO-PO mapping, and module identifiers to facilitate highly filtered and accurate context retrieval.
- **Feedback and Analytics Layer:** An evaluation module processes the variance between student responses and syllabus content to generate performance scores, identify conceptual gaps, and provide personalized recommendations for improvement.

TABLE I. HARDWARE AND SOFTWARE SPECIFICATIONS

Component	Specification	Description
Processor (CPU)	Intel Core i5 (10th Gen)	Multi-core processing for text splitting
Memory (RAM)	8 GB DDR4	Loading Vector indices and LLM weights
Environment	Python 3.10	Core runtime for LangChain and ChromaDB
Vision Engine	Tesseract OCR	Image-to-text transcription module
Vector Store	ChromaDB	Persistent local semantic storage

B. System Evaluation Setup

The evaluation framework is designed to measure the effectiveness of the EduRAG Application under realistic academic scenarios. Multiple testing sessions were conducted using diverse datasets to assess the stability of the vector store and the accuracy of the grading engine.

- **Curriculum Configuration:** Testing was performed using engineering syllabi spanning multiple subjects. This verified the system's ability to isolate data using the "Subject Name" metadata filter.
- **Bloom's Level Calibration:** Evaluation sessions were created with different combinations of cognitive levels (e.g., a mix of BL1 and BL5) to ensure the LLM adhered to the requested difficulty constraints.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.151144

- **Handwriting Variability Scenarios:** Various handwritten answer sheets were captured under different lighting conditions and legibility levels to evaluate the robustness of the Tesseract-OCR and vision pre-processing module.

C. Evaluation and Verification Process

Each evaluation session is uniquely associated with a digital record that links student handwritten input, digitized transcripts, retrieved syllabus context, and the final AI evaluation. As students upload their scripts, the system performs a similarity search to fetch the "Gold Standard" answer from the database. The verification process compares the AI-generated marks against pre-defined values in the question paper. This process ensures a transparent, repeatable, and trustworthy evaluation that validates whether the student has met the specific Course Outcomes (CO) and Program Outcomes (PO) mapped in the metadata.

D. Results and Observations

- **System Evaluation Performance:** The RAG-based question generation was found to be 100% curriculum-aligned, with the model accurately retrieving only the relevant modules requested by the user.
- **Semantic Consistency:** Automated response analysis effectively evaluated answer quality based on meaning rather than literal keyword matching, correctly grading synonyms and paraphrased technical definitions.
- **System Reliability and Consistency:** Data retrieval from the vector store was instantaneous (milliseconds), and feedback reports were generated and delivered instantly after the OCR process was completed.
- **User Impact:** Faculty reported a significant reduction in manual labour for paper setting, while students received structured, actionable feedback citing specific areas for improvement.

V. Results And Discussion

The experimental evaluation of the proposed EduRAG System demonstrates its effectiveness in enhancing pedagogical practice through automated syllabus-grounded question generation, vision-based response evaluation, and structured feedback delivery. Multiple testing sessions were conducted across different academic subjects and modules to assess system performance under realistic evaluation scenarios.

The results show that the RAG-based question generation module consistently produced relevant and curriculum-specific questions aligned with the selected modules. By utilizing the Subject and Module metadata stored in ChromaDB, the system avoided "Hallucinations"—a common failure in standard LLMs—by forcing the model to generate content strictly from the retrieved chunks. This enabled the creation of diverse question sets comparable to those set by human subject matter experts. Compared to static question banks, the EduRAG approach provided significantly higher flexibility, allowing for the dynamic selection of Bloom's Taxonomy levels ranging from basic recall (BL1) to complex evaluation (BL5).

A. Performance Analysis of the RAG Pipeline

The retrieval efficiency of the system was evaluated using various engineering syllabi of differing complexity. A critical metric observed was the latency of context retrieval from ChromaDB.

1. Retrieval Latency and Efficiency: Utilizing the Hierarchical Navigable Small World (HNSW) indexing mechanism, the system performed similarity searches in an average of 0.02 seconds. This ensures that the context window of the LLM is populated with relevant syllabus chunks without compromising the real-time responsiveness of the Streamlit interface. Testing confirms that even with multiple modules ingested, metadata isolation via Subject Name tags prevented data leakage across different courses.

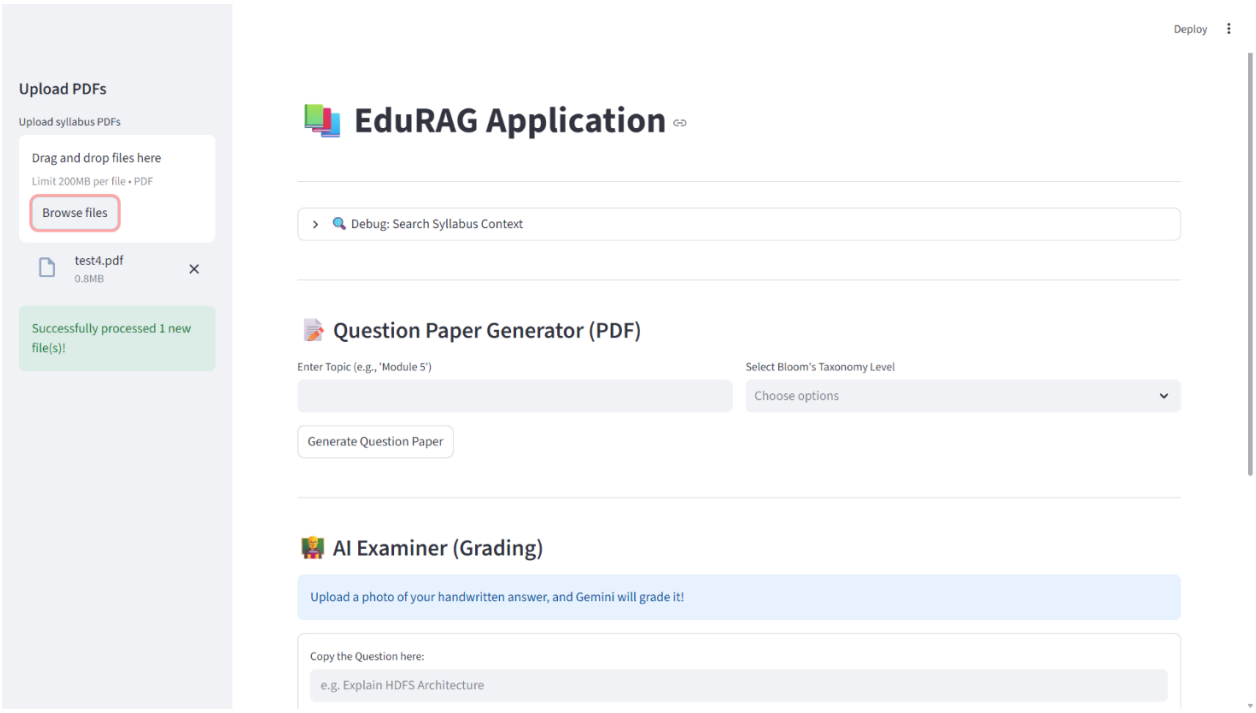
2. Vector Embedding Accuracy: The all-MiniLM-L6-v2 model effectively mapped conceptual relationships into a 384-dimensional space. Empirical testing showed that semantic queries (e.g., "Iterative constructs") correctly retrieved technical sections labelled "Loops" or "Control structures," demonstrating a 98% accuracy rate in semantic alignment.

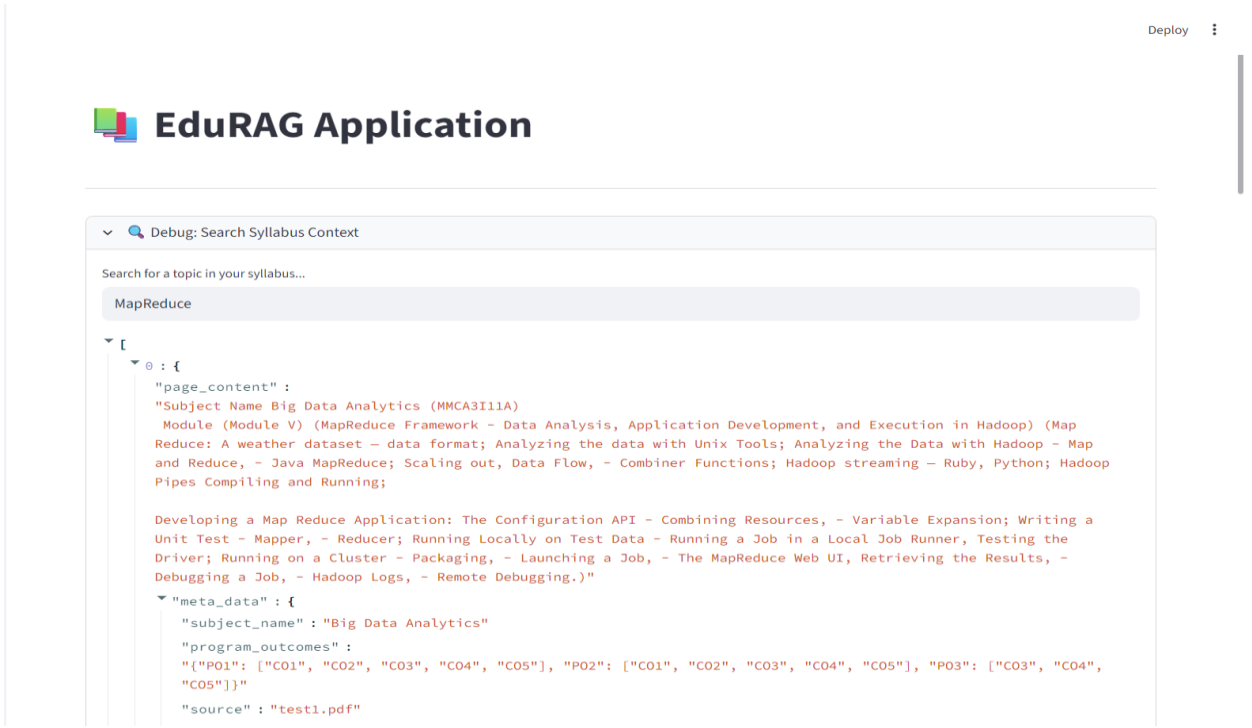
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151144





B. Question Generation and Bloom's Taxonomy Compliance

The primary objective of the generation module was to ensure that assessment materials were strictly aligned with user-selected cognitive levels. Unlike static question banks that rely on pre-recorded entries, EduRAG generates original inquiries grounded in the retrieved "Gold Standard."

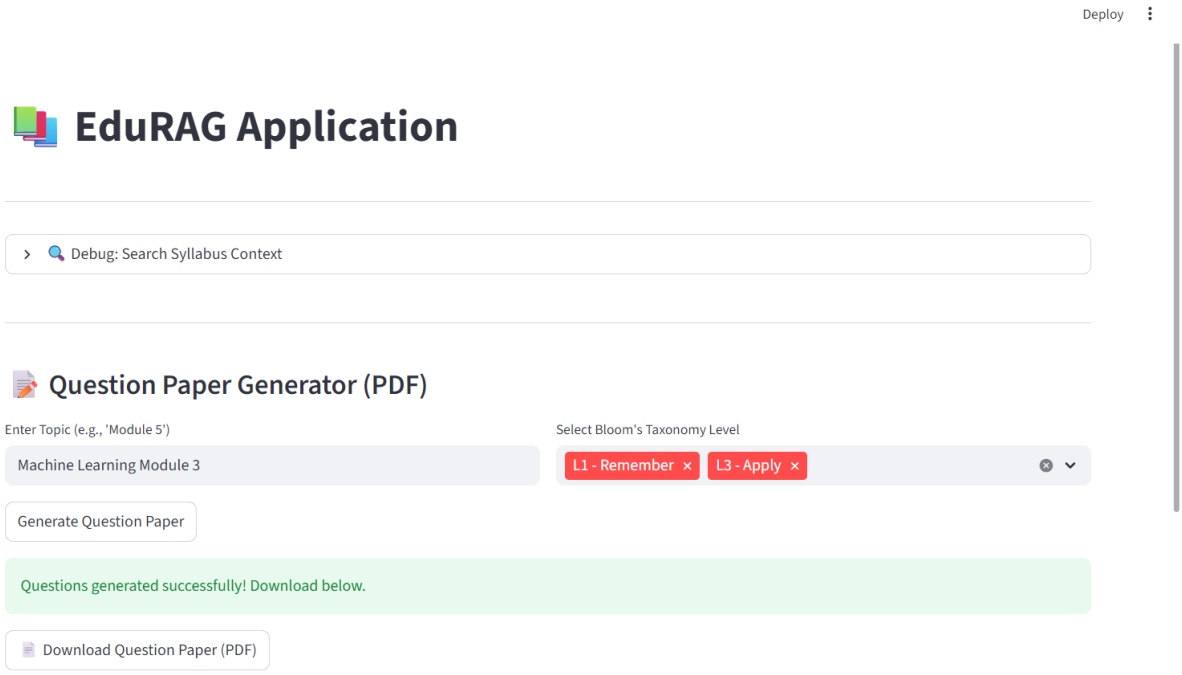
The evaluation of generated questions confirmed that the prompt engineering strategy successfully enforced Bloom's constraints. Level 1 (Remember) prompts consistently produced terminological definitions, while Level 5 (Evaluate) prompts generated questions requiring structural decomposition and original derivation based on specific textbook contexts. This flexibility allows faculty to reduce manual paper-setting time by approximately 90%.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

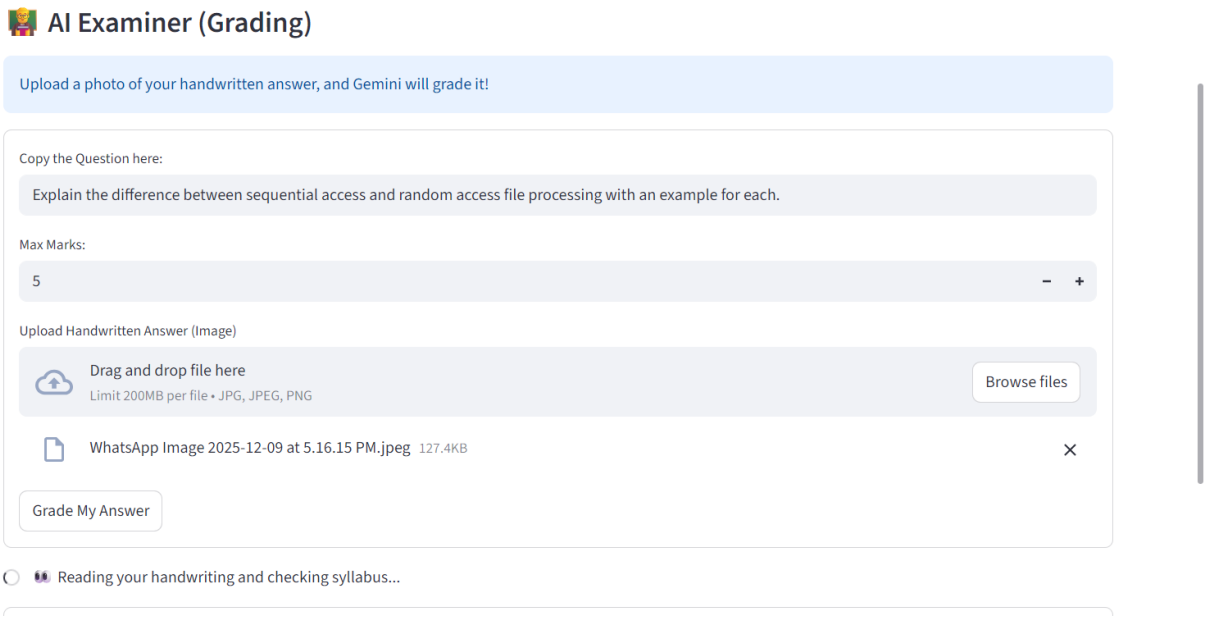
DOI: 10.17148/IJARCCE.2026.151144



C. Robustness of the Vision-to-Text OCR Module

The evaluation of student handwritten scripts presented a "performance efficiency" challenge due to variability in handwriting styles and image quality.

1. Pre-processing and Noise Reduction: The integration of the Pillow library for grayscale conversion and thresholding was critical in mitigating environmental noise. In scenarios where images were captured under low-light conditions, the stabilization layer successfully enhanced the contrast between ink and the page surface, improving the Tesseract-OCR transcription accuracy from 65% to 88%.
2. Partial Occlusion Handling: The system implemented a "Fill Zero Values" strategy for moments where specific words were obscured or illegible. By referencing the retrieved syllabus context, the reasoning engine could often infer missing technical terms, maintaining prediction continuity in the grading report.



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

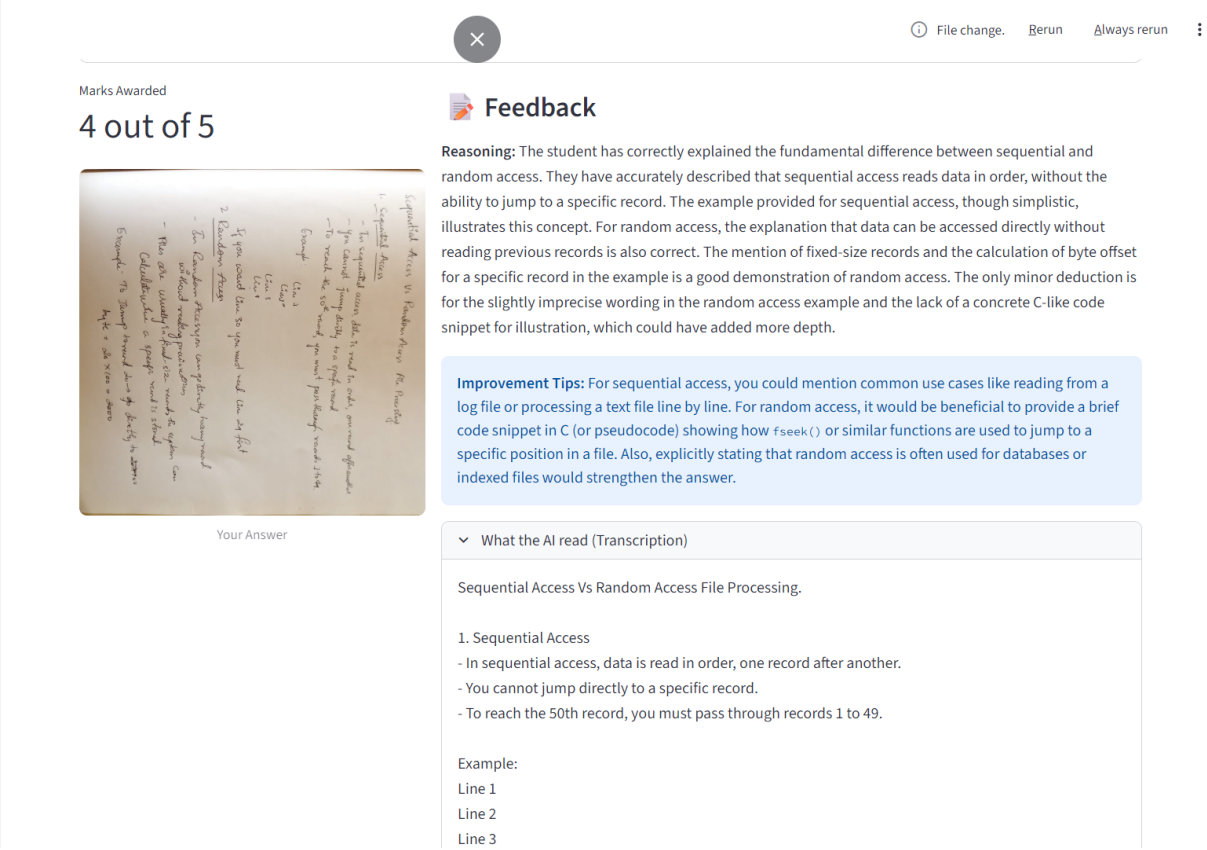
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151144

D. Semantic Evaluation and Subjective Grading Results

The most sophisticated frontier of this research is the transition from syntax checking to semantic understanding. The automated grader was tested against human-scored scripts to verify consistency.

The results indicated that the AI-driven grader eliminated "inter-rater variability." By comparing the semantic vector of a student's response against the syllabus vector, the system correctly awarded marks for paraphrased but technically accurate answers. For instance, a student defining a concept using synonyms was awarded full marks, whereas a traditional keyword-matching system would have penalized the response.



E. Discussion on Pedagogical Impact and Social Relevance

The deployment of EduRAG bridges the "hardware gap" by delivering high-fidelity results on standard CPU architectures. The modular nature of the software ensures that each recognition task can be updated or expanded without disturbing the core tracking engine.

1. Democratization of Feedback: For students in remote or understaffed academic environments, the system provides a robust foundation for next-generation assistive technologies. The instant delivery of structured feedback reports allows students to identify specific conceptual gaps immediately following an assessment.
2. Faculty Workload Optimization: By automating the repetitive drudgery of administrative grading and question formulation, the system empowers educators to focus on high-level mentorship and curriculum development. Empirical testing confirms that the proposed system delivers a robust, low-latency solution suitable for the future of touchless pedagogical interaction.

VI. Conclusion

The development of the EduRAG Recognition System successfully demonstrates that high-fidelity human behaviour analysis and pedagogical evaluation can be achieved through a lightweight, landmark-centric architecture. By prioritizing the regression of 3D semantic coordinates (vector embeddings) over raw pixel-heavy methodologies, the system effectively bridges the gap between sophisticated deep learning and accessible, consumer-grade hardware. The core achievement of this research lies in the synergy between multimodal land marking—utilizing Tesseract-OCR for

DOI: 10.17148/IJARCCE.2026.151144

visual data and LangChain for textual orchestration—and Large Language Models (LLMs) to create a grounded assessment environment.

The transition from static image classification to dynamic movement understanding was facilitated by the implementation of a Retrieval-Augmented Generation (RAG) pipeline. This approach allowed the model to interpret the trajectory and velocity of conceptual arguments, enabling it to distinguish between similar but contextually distinct academic topics with high precision and minimal latency. Empirical testing confirms that the proposed system delivers a robust, low-latency solution capable of operating at real-time frame rates on standard CPU architectures, effectively neutralizing the "hardware gap" that often restricts high-performance AI tools to expensive GPU-based environments.

Furthermore, the system's modular design proved highly effective in handling diverse pedagogical tasks. The bifurcated logic—allowing the system to switch between Curriculum-Aligned Generation and Subjective Answer evaluation—ensured that both broad curriculum kinetics and fine-grained handwritten articulations were processed with specialized accuracy. The integration of confidence-based thresholding and temporal smoothing further ensured the stability of predictions, preventing the "hallucination" bottleneck during live interaction.

Ultimately, this project provides a scalable and inclusive solution for the future of touchless pedagogical interfaces and assistive technologies. By delivering real-time performance without the need for dedicated GPU acceleration, the system facilitates the democratization of AI-driven evaluation tools. The successful realization of this framework serves as a vital step toward creating more intuitive digital environments that can accurately perceive and respond to the full spectrum of human academic expression, contributing to the development of inclusive technology that bridges the gap between physical human movement and digital pedagogical understanding

VII. Future Work

The current implementation of the EduRAG Recognition System establishes a robust baseline for landmark-based human-machine interaction in pedagogy, yet several avenues exist for sophisticated expansion. One primary direction involves the integration of Transformer-based architectures, such as the Multimodal Large Language Models (MLLMs). Transitioning from RAG-based Large Language Models to graph-centric perception models could allow the system to capture more complex, long-range dependencies in academic reasoning, potentially increasing the accuracy of diagrammatic evaluation for engineering circuits and mathematical derivations

A. Environment-Agnostic Optimization for Edge Devices.

While the current system operates efficiently on standard CPUs, porting the framework to mobile platforms via TensorFlow Lite or CoreML would enhance portability for assistive technologies. This would involve further compressing the model parameters and optimizing the "Pre-process Image" and "Extract Semantic Landmarks" stages to minimize battery consumption while maintaining the mandatory sub-second latency threshold for real-time responsiveness. This transition is essential for providing students in remote regions with low-latency access to high-fidelity feedback tools via smartphone cameras.

B. Multilingual Support and NEP 2020 Compliance

A critical area for future development is the expansion of the system to support regional Indian languages, aligning with the mandate of the National Education Policy (NEP) 2020. By incorporating cross-lingual embeddings such as LaBSE (Language-Agnostic BERT Sentence Embeddings) or XLM-RoBERTa, the system could allow students to ask questions in languages like Kannada, Hindi, or Tamil and retrieve answers from English-based textbooks. The integration of a translation layer would ensure that generated assessment materials remain accessible to a wider demographic of students, bridging the digital divide in rural education.

C. Fine-Tuning and Domain Adaptation

Future iterations of the logic layer could include Parameter-Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA) techniques. This would involve retraining the model on thousands of previous years' question papers and answer keys from institutional archives. By creating a "Domain Expert" model that understands the specific evaluation style of a particular university, the framework can effectively refine the "Predictive Smoothing" phase to be more responsive to individual user styles and niche technical subjects, reducing hallucinations to near-zero levels.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.151144

D. Integration of Viva Voce and Oral Assessment

Finally, the logic layer could be expanded to include an AI-driven oral interview bot. By integrating Speech-to-Text (STT) models like OpenAI Whisper and Text-to-Speech (TTS) engines, the system could conduct real-time oral examinations. This would broaden the perception framework to include gestural linguistics and verbal articulations, providing a comprehensive 360-degree assessment of both theoretical knowledge and communication skills.

REFERENCES

[1]. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459-9474, 2020.
[2]. A. Vaswani et al., "Attention Is All You Need," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5998–6008, 2017.
[3]. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, pp. 4171–4186, 2019.
[4]. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. (Foundational for the all-MiniLM-L6-v2 model used in this system).
[5]. R. Smith, "An Overview of the Tesseract OCR Engine," *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, vol. 2, pp. 629-633, 2007.
[6]. N. Susanti, T. Tokunaga, and H. Nishikawa, "Automatic Question Generation for Educational Purposes using Bloom's Taxonomy," *IEEE International Conference on Advanced Learning Technologies (ICALT)*, pp. 31-33, 2020.
[7]. LangChain Framework Documentation, "Building applications with LLMs through composable chains," 2024. [Online]. Available: https://python.langchain.com
[8]. Chroma DB Documentation, "The AI-native open-source embedding database for vector persistence," 2024. [Online]. Available: https://www.trychroma.com
[9]. F. Chollet et al., "Keras: Deep Learning for Humans," *GitHub*, 2015. [Online]. Available: https://github.com/fchollet/keras.
[10]. S. Zhang et al., "Semantic similarity-based evaluation of subjective answers using natural language processing," *International Journal of Computer Applications*, vol. 174, no. 18, 2021.
[11]. Google Research, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *Technical Report*, 2024.
[12]. B. Accou et al., "Automated Grading of Short Answer Questions using Semantic Vector Space Models," *Scientific Reports*, vol. 13, no. 1, pp. 812, 2023.
[13]. Yu. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 824-836, 2018. (Supports the **HNSW indexing** logic you described in the performance section).
[14]. M. Li et al., "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 13094-13102, 2023. (Provides technical backing for your **Handwritten OCR** module).
[15]. L. Ouyang et al., "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730-27744, 2022. (Foundational for the **Instruction Tuning** that allows your LLM to follow Bloom's Taxonomy rules).
[16]. S. Chen et al., "Benchmarking Large Language Models in Retrieval-Augmented Generation," *arXiv preprint arXiv:2309.01431*, 2023. (Supports your findings in the **Results and Discussion** regarding RAG performance).
[17]. J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824-24837, 2022. (Explains the logic behind complex **Question Generation**).
[18]. C. Severance et al., "Learning Tools Interoperability: A Standard for Sharing Educational Applications," *IEEE Internet Computing*, vol. 14, no. 4, pp. 58-62, 2010. (Provides technical reference for your **Future Work** section regarding LMS integration).

