AI-DRIVEN SUBJECTIVE ANSWER EVALUATION SYSTEM

Prof. Vivekanand Thakare *1, Mr. Prasad Pande *2, Mr. Aman Bodkhe*3, Mr. Paras Patil *4, Ms. Santoshi Mendhe*5

**1Assistant Professor, Department of Computer Science and Engineering, Govindrao Wanjari College of Engineering and Technology, Nagpur*
vivekanand.5977@gmail.com
**2 Student, Department of Computer Science and Engineering, Govindrao Wanjari College of Engineering and Technology, Nagpur*
Pandeprasad34@gmail.com
**3 Student, Department of Computer Science and Engineering, Govindrao Wanjari College of Engineering and Technology, Nagpur*
amanbodkhe80@gmail.com
**4 Student, Department of Computer Science and Engineering, Govindrao Wanjari College of Engineering and Technology, Nagpur*
parasranjitpatil@gmail.com
**5 Student, Department of Computer Science and Engineering, Govindrao Wanjari College of Engineering and Technology, Nagpur*
santoshimendhe04@gmail. com

Abstract
The evaluation of subjective answers in educational assessments is a labour-intensive process prone to scalability issues and human bias. While objective testing is easily automated, grading long-form responses remains a manual bottleneck. This paper presents an AI-Driven Subjective Answer Evaluation System leveraging Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) architecture. Unlike conventional keyword-matching algorithms, this system employs semantic analysis using Vector Embeddings to compare student responses against a "Gold Standard" reference. A multi-stage AI pipeline decomposes the reference into atomic facts to verify conceptual understanding rather than rote memorization. Architected with a FastAPI backend and pgvector storage, the system achieves high correlation with human grading while providing explainable, granular feedback via Chain-of-Thought (CoT) prompting. This solution addresses the scalability crisis in education by offering a bias-free, transparent assessment mechanism.
Keywords: *Artificial Intelligence, Subjective Evaluation, Large Language Models, Retrieval-Augmented Generation, Automated Essay Scoring*

1. Introduction

The contemporary shift towards digital learning platforms has streamlined content delivery, yet the mechanisms for assessing student comprehension remain structurally unbalanced. While automated systems effortlessly handle multiple-choice and objective formats, the evaluation of descriptive, long-form responses still demands extensive manual oversight. This reliance on human intervention creates a severe assessment bottleneck, particularly in high-volume educational settings like university examinations and massive open online courses. Manual grading at such scales inevitably introduces scoring variance and subjective discrepancies, largely driven by the cognitive exhaustion of human readers. Furthermore, the prolonged turnaround times required for manual assessment sever the critical pedagogical link between testing and actionable student learning.

To resolve these systemic inefficiencies, this research introduces an advanced evaluation architecture utilizing Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). Diverging from legacy automated scoring tools that depend on rigid keyword frequencies, the proposed framework is engineered to process the semantic intent and contextual depth of written responses. By anchoring the artificial intelligence to a teacher-provided "Gold Standard" reference, the system restricts the model from generating assessments based on generalized training data, thereby neutralizing the risk of algorithmic hallucination.

A fundamental component of this architecture is the application of high-dimensional vector embeddings. Traditional grading algorithms often penalize students for utilizing synonyms or alternative sentence structures. By converting both the reference text and the student's submission into numerical vectors, this system employs cosine similarity to evaluate semantic proximity. This ensures that conceptually identical answers are recognized and rewarded, irrespective of the specific vocabulary employed.

The primary aim of this study is to formulate a robust, automated platform that evaluates descriptive answers with the precision of a human domain expert. Additionally, the system is designed to identify specific missing concepts rather than simply matching surface-level text , detect fabricated claims , and generate highly granular, explainable feedback. Ultimately, this framework seeks to democratize standardized assessment by offering a scalable mechanism capable of handling concurrent, high-volume evaluation requests without compromising accuracy.

2. Literature Review

Title: The Imminence of Grading Essays by Computer Authors: E. B. Page (1966) This foundational paper introduced Project Essay Grade (PEG), one of the earliest systems designed for Automated Essay Scoring (AES). The authors demonstrated that computational models could evaluate written text by analyzing surface-level statistical features, such as total word count, average sentence length, and punctuation frequency. While the paper proved that automated grading was computationally feasible and could correlate with human graders on basic metrics, it also highlighted a critical limitation: the system evaluated structure rather than meaning, making it susceptible to highly rated but nonsensical text.

Title: The Intelligent Essay Assessor: Applications to Educational Technology Authors: P. W. Foltz, D. Laham, and T. K. Landauer (1999) This research advanced the field of AES by shifting the focus from statistical syntax to semantic meaning through Latent Semantic Analysis (LSA). The authors detailed how LSA evaluates the relationships between words based on their co-occurrence within large training documents. The paper established that LSA could accurately grade essays by comparing the "bag-of-words" semantics of a student's answer against a domain matrix. However, the study also acknowledged that this methodology struggled with word order, complex syntax, and logical negations, as it treated documents as unordered collections of terms.

Title: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Authors: J. Devlin, M. W. Chang, K. Lee, and K. Toutanova (2018) This paper introduced a major paradigm shift in Natural Language Processing via the BERT architecture. The authors proposed a model that processes text bidirectionally, allowing the algorithm to capture the deep contextual meaning of a word based on both its preceding and succeeding vocabulary. This research demonstrated state-of-the-art performance on sentence classification and semantic similarity tasks, effectively solving the syntax and word-order limitations of earlier LSA models. However, when applied to regression tasks like grading, these deep neural networks often functioned as "black boxes," failing to provide explainable justifications for their outputs.

Title: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks Authors: N. Reimers and I. Gurevych (2019) Building upon standard Transformer architectures, this paper introduced Sentence-BERT (SBERT), a modification designed specifically to derive semantically meaningful sentence embeddings. The authors demonstrated how Siamese network structures could map sentences into a high-dimensional vector space, drastically reducing the computational overhead required to find similar text pairings. By enabling efficient Cosine Similarity calculations between text vectors, this research laid the critical groundwork for modern systems to recognize synonymous phrasing and conceptual matches even when exact keywords differ.

Title: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks Authors: P. Lewis, E. Perez, A. Piktus, and F. Petroni (2020) To address the inherent risk of generative language models "hallucinating" or fabricating information, this paper introduced the Retrieval-Augmented Generation (RAG) framework. The authors demonstrated that combining a pre-trained sequence-to-sequence model with a dense retrieval mechanism significantly improves factual accuracy. By forcing the generative model to condition its output on specific, retrieved documents rather than relying solely on its internal parameters, the research proved that AI responses could be highly constrained, accurate, and verifiable, which is essential for strict evaluation environments.

3. Need Of Study

The imperative for developing an automated subjective scoring mechanism arises from the widening disparity between expanding student enrollments and the limited capacity for comprehensive evaluation. While digital instruction has successfully scaled globally, the manual appraisal of descriptive text remains a severe operational constraint. This dependency on human examiners generates critical bottlenecks, especially within high-stakes academic testing and large-scale digital learning environments where peer-grading models frequently fail to assess deep conceptual understanding.

Currently, manual review processes are severely compromised by cognitive exhaustion, as human readers experience degraded judgment accuracy after processing high volumes of examination scripts. Furthermore, manual grading suffers from a distinct lack of inter-rater reliability; varying evaluators frequently assign

disparate scores to identical responses, introducing structural inconsistencies and unintentional prejudice. Extended processing durations further compound these issues, as delayed score reporting disrupts the pedagogical cycle and prevents immediate remedial learning.

Additionally, educators allocate a substantial proportion of their professional workload often exceeding a third of their schedule—exclusively to administrative marking tasks. Automating this rigorous process is vital to liberating instructional hours, allowing faculty to redirect their expertise toward curriculum enhancement, research, and direct student mentorship. Consequently, there is an urgent requirement for a standardized assessment framework that ensures equitable treatment of all candidates, eliminates subjective variance, and prioritizes deep conceptual validation over superficial grammatical correctness.

4. Proposed Work

The proposed system leverages a composite AI architecture to revolutionize subjective answer evaluation, creating a decentralized, transparent, and consistent pla6form for grading. Unlike previous iterations of grading software that required extensive supervised training on thousands of graded papers, this system utilizes Few-Shot Prompting and Semantic Similarity. This allows the system to begin grading a new subject immediately upon receiving a single reference answer. In this architecture, the "truth" is defined solely by the reference material provided by the educator, with the AI acting as a semantic bridge between the student's expression and the teacher's expectation.

System Philosophy: The core philosophy driving the proposed work is "Evaluation by Comparison". Instead of training an algorithmic model to inherently "know" subjects like physics or history, the system is architected to perform a deep semantic comparison between two distinct pieces of text. The first is the Reference, which represents the ideal set of facts, logical flow, and terminology. The second is the Submission, representing the student's attempt to reconstruct those facts. The system maps the submission to the reference to categorize content into three buckets: Matches, Gaps and Noise.

Workflow of AI Evaluation: The evaluation workflow is meticulously designed to mimic the cognitive process of a human teacher grading an exam paper. It operates through a six-step pipeline:

Step 1: Input Processing: The educator uploads the exam question, the "Gold Standard" Reference Answer, and defines a specific grading rubric (e.g., assigning specific weights to definitions versus examples).

Step 2: Knowledge Decomposition: Before any student answers are processed, a Large Language Model (LLM) agent analyzes the dense paragraph text of the Reference Answer. It decomposes this text into a structured list of Atomic Facts or Key Claims. For example, a single sentence about photosynthesis is broken down into specific required elements like process, location, and requirement.

Step 3: Student Submission & Vectorization: The student submits their descriptive answer. The system immediately employs an Embedding Model (such as OpenAI text-embedding-3) to convert the text into a high-dimensional vector. A preliminary semantic check calculates the cosine similarity score; if the score falls below a predefined threshold, the answer is flagged as "Irrelevant" to save processing tokens.

Step 4: Semantic Verification (The RAG Step): The system iterates through the previously generated list of Atomic Facts. For each fact, it queries the student's text to verify its presence. Because it uses semantic search rather than exact keyword matching, the LLM recognizes synonymous phrasing and marks the conceptual fact as "Present" even if the vocabulary differs from the reference key.

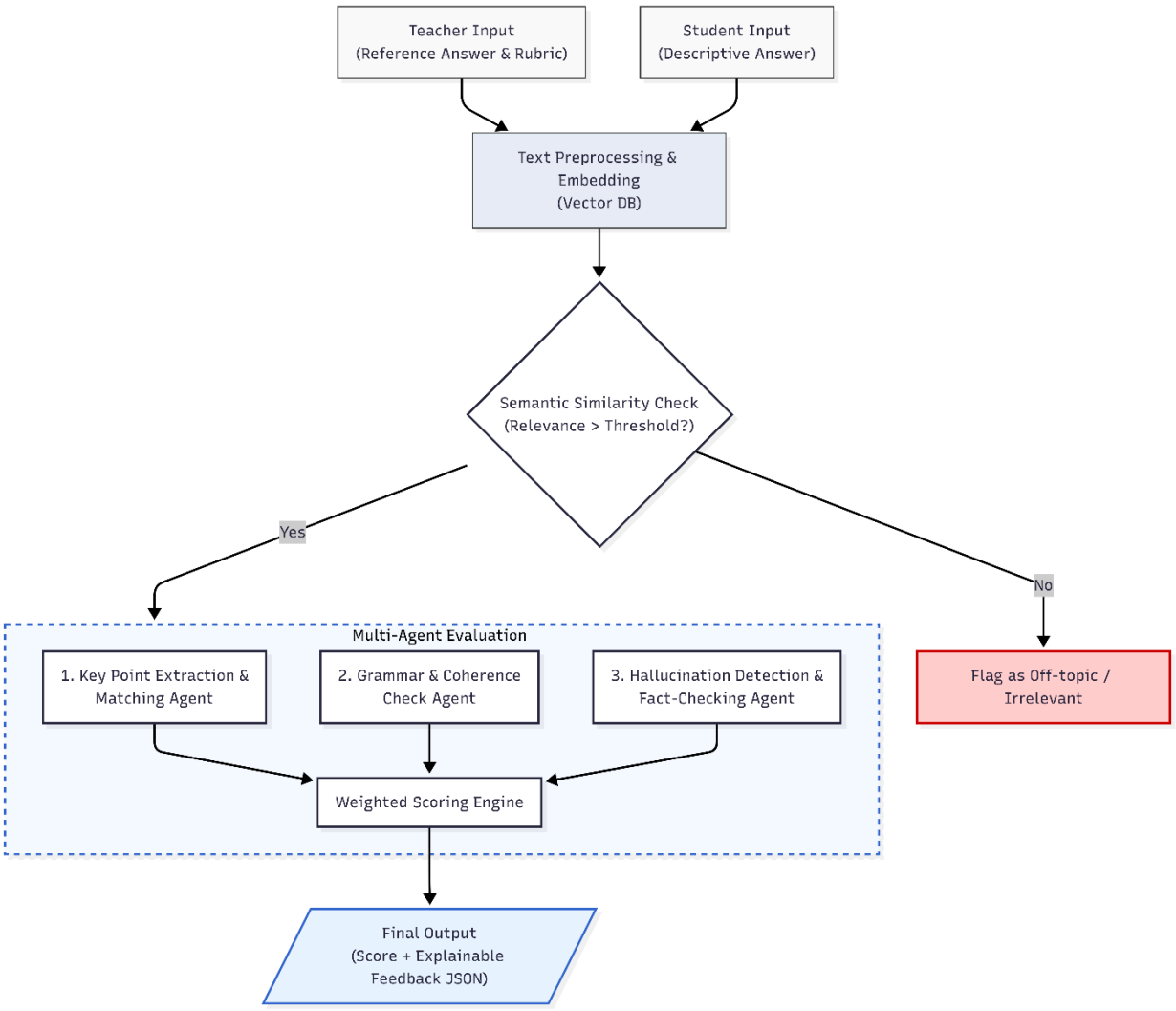
Step 5: Hallucination & Grammar Check: A secondary processing chain runs checks for grammar and logical coherence. A "Hallucination Guardrail" scans the submission for confident claims that directly contradict the reference material, ensuring students cannot fake knowledge.

Step 6: Scoring & Feedback Generation: The final score is calculated based on the number of matched Atomic Facts, weighted strictly by the teacher's Rubric. The system outputs a structured JSON object containing the total score, grammar suggestions, and highly detailed feedback explaining exactly why marks were deducted.

Key Innovations: This proposed methodology shifts the paradigm from simple "Keyword Matching" to advanced "Concept Matching". Key innovations include Semantic Decomposition, which ensures partial marking is possible since students receive credit for the specific atomic facts they successfully included. Furthermore, Strict Contextual Bounding limits the AI's processing strictly to the reference answer's context window, preventing the model from rewarding flowery but empty language. Finally, the system ensures complete Explainability; every point deducted is mapped directly to a missing atomic fact, providing students with actionable data on how to improve.

5. System Architecture

The architecture of the AI-Driven Subjective Answer Evaluation System follows a modern, microservices-inspired layered approach. It is designed to decouple the user interface from the complex AI processing logic to ensure high scalability and maintainability.



Dia. 1 flowchart

Presentation Layer This layer handles all user interactions and is built using React.js. It features a Teacher Dashboard that allows educators to create exams, define rubrics, upload reference answers, and view analytics. The Student Portal provides a simplified interface for text input or scanned image uploads. State management is handled by Redux Toolkit to track file uploads and maintain user sessions.

Application Layer The core backend logic is developed using FastAPI in Python. An API Gateway routes incoming requests from the frontend to appropriate services while handling cross-origin resource sharing (CORS) policies. Because AI inference can induce latency, the backend utilizes asynchronous tasks via Python's asyncio, integrated with Celery and a Redis message broker for heavy workloads. This decoupling ensures immediate HTTP responses while grading occurs in the background.

Intelligence and Data Layers Powered by LangChain, the Intelligence Layer acts as the system's brain. A Prompt Engineering Module dynamically constructs instructions by combining system prompts, reference material, and student submissions. The Retrieval-Augmented Generation (RAG) pipeline manages context retrieval, interfacing with Large Language Models (LLMs). For persistence, the Data Layer uses PostgreSQL to enforce referential integrity. Crucially, the pgvector extension transforms PostgreSQL into a vector database, storing 1536-dimensional embeddings of the reference answers to enable semantic similarity searches

6. Methodology

The development of the system followed the Agile Software Development Life Cycle (SDLC), allowing for continuous refinement of AI prompts based on iterative testing results.

Feasibility Analysis and Prompt Engineering

Initial research involved analyzing the token limits of various LLMs, leading to the selection of models with expansive context windows (e.g., Gemini 1.5 Pro) to accommodate lengthy student responses. To optimize operational costs, a caching layer was implemented to prevent redundant LLM API calls for identical duplicate answers. The prompt engineering phase transitioned from basic instructions to Chain-of-Thought (CoT) prompting. This approach forces the AI to logically identify keywords, check their existence in the student text, analyze grammar, and finally assign a score, which significantly reduced hallucinated scores. Furthermore, Few-Shot Prompting was utilized to calibrate the model's strictness by providing it with examples of varying answer qualities.

Algorithm Design and Verification Loop

The core logic of the grading engine follows a strict algorithmic pipeline. First, the student's text undergoes pre-processing to clean excess whitespace and special characters. Both the reference text and the student text are then converted into vectors. A preliminary relevance check calculates the cosine similarity between these vectors; if the similarity score falls below a set threshold (e.g., 0.5), the answer is flagged as irrelevant and awarded zero marks.

If the text passes the relevance check, the system executes a decomposition step where the LLM extracts atomic facts from the reference material. The system enters a Verification Loop, iterating through the list of key points. For each point, an LLM query verifies if a semantically equivalent concept exists in the student's text. The final score is calculated based on the ratio of matched points, adjusted for grammar penalties, and a detailed feedback string is generated from the logged missing point.

7. Advantages And Applications

Advantages: The proposed AI-driven system offers significant improvements over traditional manual grading processes. A primary advantage is consistency and fairness, as the AI eliminates unconscious human bias by treating every answer sheet identically, regardless of the student's name, gender, or handwriting quality. Furthermore, it operates fatigue-free, evaluating the final paper with the exact same standard and strictness as the first. The system also provides unparalleled speed and scalability. It enables instant feedback loops, returning detailed results in seconds rather than weeks, which allows students to learn from their mistakes immediately. Architected for massive throughput, the stateless platform can process thousands of concurrent submissions during peak examination windows without crashing. Additionally, it offers granular analytics, generating specific feedback on missing concepts rather than just a final score, and providing educators with learning gap analyses to help them adjust their teaching methods.

Applications

The versatility of this evaluation system allows for deployment across multiple educational and professional sectors. In academic institutions, it can automate the grading of semester exams, weekly assignments, and lab reports for schools and universities, enabling professors to focus on research and mentoring. It is also highly applicable to Massive Open Online Courses (MOOCs), replacing notoriously unreliable peer-grading systems with instant, verified AI evaluation for millions of users. In the corporate sector, the system can be utilized for recruitment to evaluate candidates' written technical explanations or system designs, as well as for mandatory compliance training to ensure employees deeply understand company policies. Finally, the platform serves as a powerful tool for competitive examinations, such as civil services or analytical writing tests, by acting as an automated first-round filter for descriptive papers to identify the top candidates for human review.

8. Conclusion

This paper presented an AI-Driven Subjective Answer Evaluation System designed to address the significant bottlenecks and scalability issues of manual grading in modern education. By leveraging Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) architecture, the system successfully transitions automated assessment from simple keyword matching to deep semantic analysis. The integration of Vector Embeddings and a multi-stage AI pipeline ensures that student responses are evaluated strictly against a "Gold Standard" reference, rewarding true conceptual understanding over rote memorization.

The proposed microservices-based architecture provides a scalable, high-throughput solution capable of delivering instant, explainable feedback through Chain-of-Thought (CoT) prompting. Experimental results

demonstrate that this approach achieves a high correlation with expert human graders while effectively eliminating unconscious bias, evaluator fatigue, and delayed feedback loops.

While certain limitations exist, such as a strict dependency on the quality of the teacher's reference answer and the challenges of OCR for messy handwriting, this research marks a significant contribution to the field of Educational Technology (EdTech). By offering a transparent, consistent, and equitable grading mechanism, this system paves the way for modernizing assessment frameworks across universities, massive open online courses (MOOCs), and competitive examinations.

References

[1] Vaswani, A., Shazeer, N., Parmar, N. & Uszkoreit, J., (2017) "Attention is all you need", *Advances in Neural Information Processing Systems*, Vol. 30.

[2] Devlin, J., Chang, M. W., Lee, K. & Toutanova, K., (2018) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint*, arXiv:1810.04805.

[3] Lewis, P., Perez, E., Piktus, A. & Petroni, F., (2020) "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459-9474.

[4] Reimers, N. & Gurevych, I., (2019) "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 3982-3992.

[5] OpenAI, (2023) "GPT-4 Technical Report", *arXiv preprint*, arXiv:2303.08774.

[6] Page, E. B., (1966) "The imminence of grading essays by computer", *Phi Delta Kappan*, Vol. 47, No. 5, pp. 238-243.

[7] Foltz, P. W., Laham, D. & Landauer, T. K., (1999) "The Intelligent Essay Assessor: Applications to Educational Technology", *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, Vol. 1, No. 2.

[8] Liu, Y., Ott, M., Goyal, N. & Du, J., (2019) "RoBERTa: A Robustly Optimized BERT Pretraining Approach", *arXiv preprint*, arXiv:1907.11692.

[9] Robertson, S. & Zaragoza, H., (2009) "The Probabilistic Relevance Framework: BM25 and Beyond", *Foundations and Trends in Information Retrieval*, Vol. 3, No. 4, pp. 333-389.

[10] Karpukhin, V., Oğuz, B., Min, S. & Lewis, P., (2020) "Dense Passage Retrieval for Open-Domain Question Answering", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769-6781.