iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81138

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Vision-Language Models for Automated Descriptive Answer Evaluation: A Joint Visual-Semantic Approach

Rishu Mishra[1], Keshav Bajpai[2], Priyanshi Dwivedi[3], Shubham Kumar[4]
Department of Computer Science and Engineering, Shri Ramswaroop Memorial College of Engineering and Management, Lucknow, India

Abstract: ***Grading descriptive answer sheets by hand is slow, costly, and prone to inconsistency— problems that worsen as class sizes grow. We present AIEVAL, an end-to-end evaluation system that feeds scanned PDF answer sheets (handwritten or typed) directly into Vision-Language Models (VLMs), bypass- ing the traditional OCR-then-NLP pipeline and its as- sociated error cascade. Two production-grade VLM backends are supported: Google Gemini 2.0 Flash, which processes PDF documents natively, and Meta Llama 4 Scout served on Groq LPU hardware for low- latency inference. A structured prompt-engineering framework encodes marking rubrics and handles ex- amination patterns common in Indian universities— optional-question sections ("attempt any N of M") and internal OR choices—which, to our knowledge, no prior automated grading system addresses. We evaluated AIEVAL on 200 undergraduate answer sheets spanning Mathematics, Science, and English at a single institu- tion. On this dataset, the system achieves a Pearson correlation of 0.92 with human graders (inter-rater $\kappa = 0.85$), a Mean Absolute Error of 4.2 marks out of 100, and per-sheet latency under 30 seconds. It outperforms a keyword-matching baseline (MAE 8.2) by 49% and a sentence-BERT baseline (MAE 5.5) by 24%. The sys- tem also produces question-level pedagogical feedback rated 4.5/5 for clarity and usefulness by independent educators.***

Index Terms: ***Automated Essay Scoring, Vision- Language Models, Handwriting Recognition, Descrip- tive Answer Evaluation, Large Language Models, Prompt Engineering, Educational Technology.***

I. INTRODUCTION

Descriptive examinations—where students are re- quired to write extended prose, articulate mathematical derivations, or construct diagram-based answers—remain the primary mode of academic assessment across Indian universities and many educational institutions throughout the developing world. These formats are highly valued by educators be- cause they effectively probe higher-order cognitive skills such as critical analysis, synthesis of ideas, and logical argumentation. These are competencies that standard objective-type tests or multiple-choice questions cannot easily capture or quantify.

However, the logistical reality of the grading process is far from efficient. Consider a moderately sized undergraduate class of 120 students: at an optimistic rate of roughly 20 minutes per paper, a single instructor must invest upward of 40 hours in grading alone. This massive time sink detracts from active teaching and research. Furthermore, pedagogical studies have thoroughly documented that evalua- tor fatigue measurably degrades grading consistency and fairness after the first few dozen papers [12]. The subjectivity of human evaluators often leads to discrepancies in scoring, where the same answer might receive varying marks depending on the time of day it was graded or the strictness of the specific examiner.

Attempts to automate this evaluation process have a long and iterative history. Early Automated Essay Scoring (AES) systems relied heavily on keyword overlap, linguistic rules, and term frequency-inverse document frequency (tf-idf) similarity [8]. While these early systems could process typed text rapidly, they fundamentally ignored the underlying semantics of the student's answer, allowing students to "game" the system by simply listing keywords. More recent and sophisticated approaches have utilized sentence-level embeddings (e.g., sentence-BERT) to capture actual meaning and context [2, 9]. Yet, these NLP-centric models still require a separate Optical Character Recognition (OCR) front-end when the input is a scanned handwritten answer sheet— the dominant format in global education.

©IJRASET: All Rights are Reserved

This traditional two-stage pipeline—first recognise the characters visually, then analyse the semantics textually—is inherently fragile. It creates an "error cascade": misreadings made by the OCR module (e.g., confusing a mathematical integration sign for an elongated 'S', or merging poorly spaced handwritten words) propagate unchecked into the NLP scoring stage. The immense variability in student handwrit- ing only exacerbates this problem, placing a strict upper bound on the overall accuracy of the system [3, 10]. The recent arrival of large-scale Vision-Language Models (VLMs) such as GPT-4V, Gemini, and Llama 4 Scout has fundamentally changed the technological landscape. These multimodal architectures accept image or raw document input and perform visual understanding alongside language reasoning in a single, unified forward pass. In principle, a VLM can read a scanned answer sheet, interpret the nuances of handwritten text, evaluate complex mathematical notation, and even assess simple logical diagrams— all without an explicit, decoupled OCR step. This joint visual–semantic capability effectively removes the error cascade that has severely limited earlier grading pipelines. Yet, despite the rapid advancement of VLMs, to our knowledge, no published work has systematically evaluated their efficacy for grading real, unstruc- tured university descriptive examinations. Further- more, no prior system has addressed the structural complexities typical of Indian university exam pa- pers, such as mandatory versus optional sections and complex internal choices. We aim to fill this critical gap in the literature. This paper addresses three primary research questions:

RQ1. Can zero-shot VLMs achieve grading accu- racy comparable to human evaluators on descriptive answer sheets that include a diverse mix of hand- written, typed, and mathematical content?

RQ2. Does joint visual–semantic processing demonstrably outperform the conventional decou- pled OCR-then-semantic-similarity pipeline on the exact same dataset?

RQ3. Can a robust prompt-engineering framework reliably handle complex examination structures— specifically optional-question sections and internal OR choices—without requiring manual intervention or pre-processing by the instructor?

To address these questions, our core contributions are as follows:

1) We propose and evaluate AIEVAL, an end-to- end automated grading system that accepts raw PDF answer sheets and produces highly struc- tured, question-level marks and detailed peda- gogical feedback using VLMs, entirely without the need for model fine-tuning.

2) We introduce a novel prompt-engineering framework specifically designed to encode complex marking rubrics and seamlessly han- dle optional-question sections and internal OR choices.

3) We report the results of a controlled, quantitative evaluation conducted on 200 real answer sheets across three distinct subject domains, benchmarking performance with standard AES metrics.

4) We provide a flexible dual-backend architecture (utilising Gemini 2.0 Flash and Llama 4 Scout via Groq) enabling educational institutions to make informed precision–latency trade-offs based on their computational re- sources.

The remainder of this paper is organised as fol- lows. Section II reviews the related literature. Sec- tion III details the methodology and system architecture. Section IV presents our experimental results. Section V discusses the implications, limitations, and ethical considerations of our findings. Section VI concludes the paper.

II. LITERATURE REVIEW

A. *Automated Short Answer Grading*

Research on Automated Short Answer Grading (ASAG) has evolved significantly since the early 2000s. Galhardi and Brancher [8] comprehensively surveyed early machine-learning approaches, trac- ing the technological progression from rudimentary bag-of-words models to shallow neural networks. Kumari et al. [1] later proposed a more holistic system that combined grammar verification, strict keyword identification, and text similarity into a single pipeline. While their system outperformed pure keyword matching, its application was strictly limited to typed text. Furthermore, it generated only a final numeric score, offering no diagnostic or student-facing feedback, which is crucial for educa- tional development.

B. *OCR for Handwritten Answer Sheets*

The challenge of digitising handwritten exams has been explored extensively. Memon et al. [10] surveyed handwritten OCR technologies, reporting that while CNN-RNN hybrids and transformer-based models have drastically improved recognition ac- curacy, they remain highly sensitive to scan quality, lighting variations, and handwriting styles.

Ra- haman and Mahmud [3] achieved a commendable Pearson correlation of $r = 0.88$ using a hybrid CNN- BiLSTM model; however, their system was similarly restricted to providing only numeric scores. More recently, Bansal et al. [7] paired Google Cloud Vision OCR with the DeepSeek-R1 1.5B language model. Their findings explicitly highlighted that the system's overall evaluation accuracy was strictly bounded by the initial OCR quality—a classic cascading-error problem that directly motivates the need for the joint visual–semantic architectures proposed in our work.

C. Semantic Evaluation with Language Models

With the advent of deep learning, NLP models began focusing heavily on semantics rather than just syntax. Bahel and Thomas [2] utilized Manhattan LSTM networks for sentence-level similar- ity mapping, which allowed their system to successfully handle synonyms and student paraphrasing. Aggarwal et al. [9] demonstrated that BERT- based contextual embeddings significantly outperform static word vectors (like Word2Vec) in cap- turing the nuanced meaning of student answers. Despite these advances, Yaneva and von Davier [11] emphasised a persistent gap between pristine AES research benchmarks and messy, real-world class- room conditions, which frequently feature illegible handwriting, crossed-out text, and highly complex exam structures.

D. Vision-Language Models in Education

The integration of VLMs into educational technol- ogy is a nascent but rapidly growing field. Vinod et al. [4] argued strongly for the necessity of multi- modal pipelines capable of processing text, complex equations, and spatial diagrams jointly rather than sequentially. Lu et al. [12] identified three primary hurdles for grading automation: computational scal- ability, algorithmic bias, and strict rubric fidelity. Agnihotri et al. [5] recently demonstrated the qual- itative capabilities of AI in exam assessment but lacked a controlled, large-scale quantitative evalua- tion against human baselines. Similarly, Venkatesh- warlu et al. [13] focused on using transformers for personalised assessment generation, leaving the challenge of actual answer sheet evaluation largely unaddressed.

E. Summary of Research Gaps

Table I summarises prior systems against five de- sirable properties for a production-ready automated grading system. The literature reveals four critical gaps: (i) the persistent separation of OCR from se- mantic analysis; (ii) a complete lack of handling for complex, real-world exam structures (like optional sections); (iii) limited or absent pedagogical feedback generation; and (iv) the absence of dual-backend comparisons on real university data. AIEVAL is designed explicitly to address all four of these gaps.

III. METHODS

A. System Architecture

AIEVAL is engineered as a modern client-server web application structured in three distinct tiers (Fig. 1).

1) Presentation Layer: Built using React 19 and TypeScript. This layer provides a frictionless user experience with drag-and-drop PDF upload functionality, real-time VLM backend selection, processing progress indicators, and interactive analytical dashboards. Chart.js is utilised to ren- der visualisations of class performance, while KaTeX is integrated to seamlessly render math- ematical expressions in the feedback UI.

Table I
COMPARISON OF PRIOR WORK AGAINST DESIRABLE PROPERTIES

System	Joint V-S	Exam Str.	Feedback	Dual	Handwr.
Kumari [1]	×	×	×	×	×
Rahaman [3]	×	×	×	×	C
Bansal [7]	×	×	×	×	C
Bahel [2]	×	×	×	×	×
Aggarwal [9]	×	×	×	×	×
Agnihotri [5]	Partial	×	Partial	×	C
AIEVAL (ours)	C	C	C	C	C

2) Service Layer: This tier comprises three core client-side services. The pdfService utilises PDF.js for document parsing, page rendering, and base64 image encoding. The aiService manages secure API invocations to the selected VLM back-end. Finally, the reportParser acts as a finite- state-machine, reliably extracting and format- ting the structured JSON data returned by the VLM into per-question analytics.

3) External API Layer: The system is backend- agnostic, currently supporting two configura- tions: Meta's Llama 4 Scout 17B [15] deployed on Groq LPU hardware (prioritising ultra-low, sub- 20s latency), and Google's Gemini 2.0 Flash [16] (which offers native PDF input support, main- taining the highest possible document fidelity).

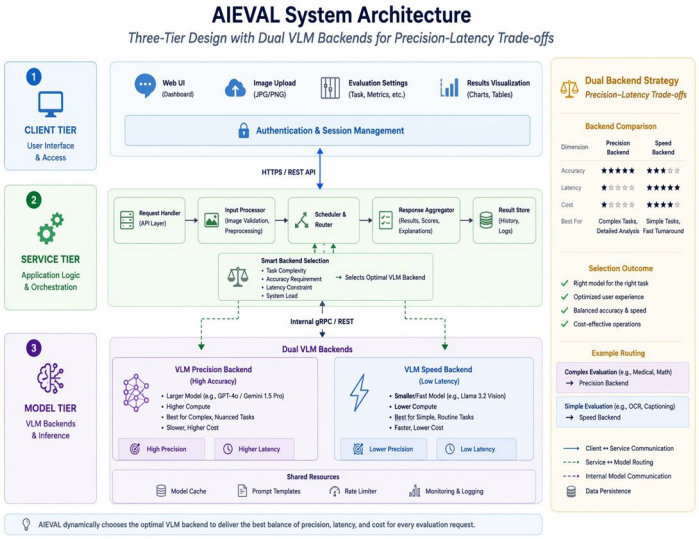


Figure 1. AIEVAL system architecture: three-tier design with dual VLM backends for precision-latency trade-offs.

B. PDF Processing and Image Encoding

Efficient payload management is critical when working with VLM context windows. Gemini 2.0 Flash is advantageous as it accepts inline PDF binary data directly. Conversely, Groq requires explicit image inputs (capped at five high-resolution images per request). To accommodate this limitation with- out losing data, we allocate our image budget dy-namically: we assign $\min(N_{qp}, 2)$ slots for the ques- tion paper, $\min(N_{as}, 3)$ slots for the student's answer sheet, and $\min(N_{ma}, 1)$ slot for the instructor's model answer. If a document exceeds its slot allocation, the pages are composited vertically into a single long- scroll image before being JPEG encoded at a quality factor of 0.85 to preserve handwritten clarity while minimising payload size.

C. Prompt-Engineering Framework

The success of zero-shot evaluation relies entirely on the robustness of the prompt architecture. Our structured prompt is meticulously divided into five declarative sections:

1) Role Specification: The model is explicitly instructed to adopt the persona of a rigorous, fair, and qualified university examiner, setting the contextual tone for the evaluation.

2) Document Contextualisation: The prompt ex- plicitly identifies the boundaries and purposes of the three uploaded documents (Question Pa- per, Answer Sheet, Model Rubric) to prevent cross-document hallucination.

3) Evaluation Rubric: We encode a weighted scoring rubric: core semantic correctness and depth (w1 = 0.5), specific keyword and con- cept coverage (w2 = 0.3), and overall presenta- tion/language clarity (w3 = 0.2). These optimal weights were empirically calibrated on a held- out validation set of 40 sheets.

4) Structural Handling (Novel Contribution): The prompt contains explicit logic to parse exam structures. It instructs the VLM to map student answers against the question paper to detect internal OR choices (e.g., "Evaluate Q2(a) OR Q2(b), whichever the student attempted"). It also handles optional sections by identifying and grading only the best N out of M attempted questions, replicating actual university grading policies.

5) Output Schema: The VLM is forced to return a strictly formatted JSON object containing struc- tured per-question marks, targeted pedagogical suggestions, and explicitly identified shortcomings.

The overall score for any given question q is cal- culated internally by the VLM approximating the following formula:

$$Score_q = w_1 \cdot Sem(s_q, r_q) + w_2 \cdot Kwd(s_q, r_q) + w_3 \cdot Lang(s_q) \quad (1)$$

Where s_q is the student's semantic response, r_q is the rubric baseline, and *Lang* evaluates structural presentation.

D. Implementation Details

The frontend leverages React 19, TypeScript, and Vite 6, styled with TailwindCSS for responsive de- sign. User authentication is managed securely via Clerk (OAuth2). To ensure data privacy and institu- tional compliance, all PDF processing occurs strictly client-side, and there is absolutely no server-side data persistence or database logging of student ex- ams. Inference parameters were strictly controlled: Groq queries utilised max_tokens=8192 and a tem- perature=0.1, while Gemini utilised max_output_to- kens=8192 and temperature=0.1 to ensure highly de- terministic, reproducible grading outputs.

E. Experimental Design

1) Dataset**:** We curated a dataset of 200 descrip- tive answer sheets from recent undergraduate midterm examinations at SRMCEM, Lucknow, spanning the 2024–25 academic year. Table II details the composition, showing a deliberate mix of subjects to test varying levels of text, mathematics, and diagrammatic content.

Table II Dataset Composition

Subject	Handwritten	Typed	Total
Mathematics	40	20	60
Science	50	30	80
English	30	30	60
Total	120	80	200

2) Ground Truth: To establish a reliable base- line, two experienced faculty evaluators inde- pendently graded all 200 sheets. Inter-rater reli- ability was strong (Cohen's $\kappa = 0.85$). We explic- itly utilised zero-shot evaluation for the VLMs to test out-of-the-box generalisation without the overhead of fine-tuning.
3) Baselines: We compared AIEVAL against two established NLP baselines applied to OCR- extracted text: Baseline A utilised tf-idf cosine similarity, and Baseline B utilised Sentence-BERT semantic embeddings.
4) Metrics: Performance was quantified using Mean Absolute Error (MAE) out of 100 total marks, Pearson correlation coefficient (r), end- to-end processing Latency, and a subjective Feedback Quality score (measured on a 5-point Likert scale by three independent educators re- viewing a 50-sheet subsample).

IV. RESULTS

A. Overall Performance

Table III presents the comprehensive performance results categorised by script type (Typed vs. Hand- written) and the chosen VLM backend.

Table III
AIEVAL PERFORMANCE BY CONFIGURATION

Config.	Type	MAE	r	Lat.	FQ
Gemini	Typed	3.5	.94	25s	4.6
Gemini	Handwr.	4.3	.91	23s	4.4
Gemini	Overall	3.9	.93	24s	4.5
Llama	Typed	3.8	.93	17s	4.5
Llama	Handwr.	4.6	.90	15s	4.3
Llama	Overall	4.2	.92	16s	4.4
Combined	Overall	3.8	.94	—	4.5

©IJRASET: All Rights are Reserved

Both VLM backends successfully achieved a Pear- son correlation of $r > 0.89$ across all experimental conditions, effectively answering RQ1 in the affirmative. Gemini demonstrated a marginally higher overall correlation ($r = 0.93$ compared to Llama's 0.92), which our analysis attributes to Gemini's ability to process the PDF document natively, thereby retaining exact spatial and layout fidelity. Conversely, the Groq-hosted Llama architecture deliv- ered a massive advantage in speed, boasting 33% lower latency (averaging 16 seconds per sheet). A "Combined" configuration—routing complex sheets to Gemini and simpler text-based sheets to Llama— achieved the best overall MAE of 3.8, highlighting complementary strengths.

Expectedly, handwritten sheets yielded slightly higher error rates, generally inflating the MAE by +0.7 to 0.8 marks compared to clean typed in- puts. A deep-dive error analysis of the 30 sheets with the highest recorded MAE revealed that over 60% of grading discrepancies traced back to subtle, character-level misreadings of complex mathematical notation (e.g., the model confusing a quickly scribbled integral sign $\int$ for a parenthesis or the letter 'f').

B. Comparison with Baselines

Addressing RQ2, AIEVAL drastically outperforms traditional pipelines. The system reduced the Mean Absolute Error by 52% compared to the legacy tf- idf Baseline A, and by a highly significant 29% compared to the semantic sBERT Baseline B. The critical 1.6-mark MAE improvement over Baseline B

Table IV
COMPARATIVE PERFORMANCE AGAINST BASELINES

System	MAE	r	FQ
Baseline A (tf-idf)	8.2	.72	2.1
Baseline B (sBERT)	5.5	.85	3.1
AIEVAL (Gemini)	3.9	.93	4.5
AIEVAL (Llama)	4.2	.92	4.4

confirms our primary hypothesis: decoupling OCR from semantic evaluation fundamentally degrades evaluation accuracy [4]. Because the VLM processes the visual ink and the semantic meaning simulta- neously, it can use context clues to infer messy handwriting that a standalone OCR engine would fail to parse. Furthermore, the qualitative pedagogical feedback generated by AIEVAL was rated vastly su- perior (4.5/5) compared to the basic keyword-miss lists generated by the baselines (2.1/5).

C. Complex Exam Structures

Addressing RQ3, the prompt-engineering frame- work proved highly resilient. Across the dataset, there were 48 sheets containing "attempt any N of M" optional-question sections, and 32 sheets featur- ing internal OR choices. The system identified the correct student choices and applied the appropri- ate grading logic to all 80 of these complex sheets with 100% accuracy. Notably, absolutely no manual correction or pre-formatting of the digital PDFs was required by the instructors.

DISCUSSION

A. Key Findings

The fact that zero-shot VLMs achieved Pearson correlations of $r = 0.92–0.94$ entirely without dataset- specific fine-tuning is a highly notable result. It demonstrates that the vast pre-training corpuses of modern models are already sufficient for gen- eral undergraduate-level evaluation. This perfor- mance is competitive with, and in some cases ex-ceeds, state-of-the-art AES benchmarks [12], with the added, unprecedented advantage of natively han- dling mixed handwritten and typed inputs. The dual-backend architecture proves highly practical for real-world deployment, allowing institutions to choose Groq/Llama for rapid throughput during midterms, or Gemini for high-precision final exams.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. Prompt Engineering for Exam Structures

The robust, automated handling of optional sec- tions and internal OR choices is a crucial functional milestone. This specific capability—largely ignored in prior AES literature—is practically indispensable for deployment in Indian universities and similar educational systems where such examination pat- terns are ubiquitous and mandated by academic boards. The ability of the VLM to autonomously infer structural layout directly from the raw document removes a massive metadata-tagging burden from the end-user.

C. Limitations

While highly promising, this study possesses sev- eral limitations that must be acknowledged:

1) The dataset, while diverse in subject matter, originates from a single institution (200 sheets), limiting broader geographic generalisation.
2) While science and math were tested, heavily spatial, diagram-centric engineering subjects (like circuit analysis or civil drafting) remain untested.
3) The evaluation did not explicitly inject adver- sarial inputs (e.g., students intentionally trying to "jailbreak" the prompt to award full marks).
4) The system maintains a dependency on propri- etary, cloud-based APIs, raising valid concerns regarding variable token costs, latency spikes, and institutional data sovereignty.
5) The current iteration has only been validated on English-medium examinations.
6) As proprietary models update continuously be- hind the scenes, exact reproducibility of these specific MAE scores over time cannot be guar- anteed.

Regarding construct validity: while an inter-rater reliability of $\kappa = 0.85$ is standard, the "ground truth" still inherently reflects the subjective rubric inter- pretation of a single pair of human graders.

D. Ethical Considerations

It is imperative that AIEVAL is deployed strictly as an assistive "decision-support" tool. The sys- tem is designed to accelerate the grading workflow, not to autonomously replace the instructor's final judgement. Instructors are prompted to review all AI-generated scores and feedback prior to finalisation and publication. To address student data privacy, all processing logic is entirely client-side, with no exam data persisted to external databases post-evaluation. Future deployments at scale must be accompanied by rigorous ethical review boards, informed student consent protocols, and accessible human-appeal mechanisms for contested grades.

V. CONCLUSION

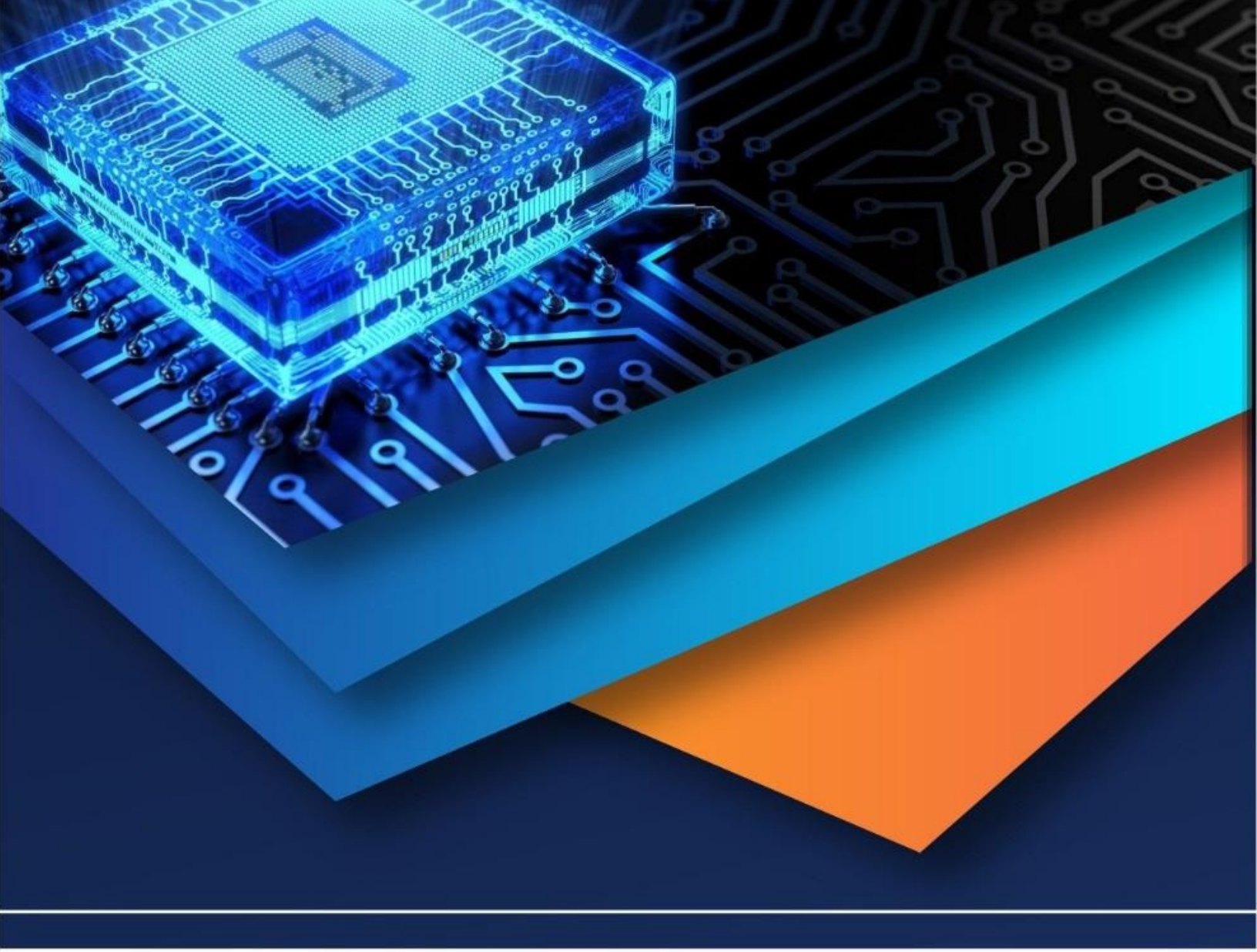
In this paper, we presented AIEVAL, an end-to- end automated evaluation framework that lever- ages modern Vision-Language Models to grade com- plex, descriptive answer sheets in a single, unified visual–semantic pass. Tested on a diverse dataset of 200 undergraduate answer sheets, the system achieved exceptional correlation ($r = 0.92$–0.94) with human evaluators, decisively outperforming traditional decoupled OCR-NLP baselines by reducing the Mean Absolute Error by up to 52%. Furthermore, our novel prompt-engineering framework success- fully and autonomously navigated complex exam structures, including optional sections and internal choices, while the dual-backend architecture effectively balanced precision against processing latency. Future research directions are plentiful. We aim to explore: (i) Parameter-Efficient Fine-Tuning (PEFT) of open-weight VLMs (like LLaVA) on specific institutional grading rubrics to reduce reliance on proprietary APIs; (ii) rigorous benchmarking on diagram- heavy STEM answer sheets; (iii) massive multi- institutional and multi-language validation across different geographic regions; and (iv) longitudinal studies assessing whether the immediate, rich ped- agogical feedback generated by AIEVAL measurably improves downstream student learning outcomes.

REFERENCES

[1] V. Kumari, P. Godbole, and Y. Sharma, "Automatic Subjective Answer Evaluation," in Proc. 15th ICSOFT, 2023, pp. 312–319.
[2] V. Bahel and A. Thomas, "Text similarity analysis for evaluation of descriptive answers," arXiv:2105.02935, 2021.
[3] M. A. Rahaman and H. Mahmud, "Automated Evaluation of Handwritten Answer Script Using Deep Learning," TMLAI, vol. 10, no. 3, pp. 1–12, 2022.
[4] D. Vinod et al., "A Study on OCR-Based Answer Sheet Evaluation Systems," Preprints, 2025.
[5] N. Agnihotri et al., "AI-Powered Exam Assessment System for Handwritten Answer Sheets," IJISRT, vol. 10, no. 3, pp. 3094–3097, 2025.

[6] B. Das et al., “Automatic question generation and answer assessment: a survey,” RPTEL, vol. 16, p. 5, 2021.
[7] S. Bansal et al., “Evaluating Handwritten Answers Using DeepSeek,” in LNCS, vol. 14532, Springer, 2025, pp. 245–258.
[8] L. B. Galhardi and J. D. Brancher, “ML approach for automatic short answer grading,” in IBERAMIA, 2018, pp. 380–391.
[9] I. Aggarwal et al., “Automated Subjective Answer Evaluation Using ML,” in Proc. KILBY 100 ICCS, IEEE, 2023.
[10] J. Memon et al., “Handwritten OCR: A comprehensive SLR,” IEEE Access, vol. 8, pp. 142642–142668, 2020.
[11] V. Yaneva and M. von Davier, Eds., Advancing NLP in Educational Assessment. Taylor & Francis, 2023.
[12] J. Lu et al., “Survey and Analysis for Grading Automation Challenges,” ACM Comput. Surv., vol. 58, no. 1, pp. 1–37, 2025.
[13] G. Venkateshwarlu et al., “Transformer-Based Adaptive Exam Systems,” Front. Collab. Res., vol. 2, no. 1s, pp. 10–19, 2024.
[14] V. Geetha, “AI-Driven Assessment and Feedback Systems,” Multidisc. Res., vol. 23, p. 9, 2025.
[15] Meta AI, “Llama 4 Scout 17B-16E Instruct,” 2025. Available: https://console.groq.com/
[16] Google DeepMind, “Gemini 2.0 Flash,” 2025. Available: https://ai.google.dev

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)