Vol. 08, No. 1 (2026) 311-320, doi: 10.24874/PES08.01A.004

Proceedings on Engineering Sciences

www.pesjournal.net

AI-DRIVEN SUBJECTIVE ANSWER EVALUATION: INTEGRATING OCR AND SEMANTIC SIMILARITY FOR INTELLIGENT LEARNING ENVIRONMENTS

Amita Yadav[1]
Shalu
Sonia Rathee

Received 13.12.2025
Revised 15.01.2026
Accepted 12.02.2026

Keywords:

Automated Grading, Subjective Answer Evaluation, Optical Character Recognition (OCR), Semantic Similarity, Sentence-BERT, Intelligent Tutoring Systems, Computer-Supported Assessment, E-Learning

A B S T R A C T

The rising interest in proficient and objective grading methodologies in education has prompted the development of an Automatic Grading System for Subjective Exams. Our research work makes an attempt to utilize Optical Character Recognition (OCR) technologies and semantic similarity models for automating the evaluation of subjective answers. The suggested system facilitates instructors in generating exams, uploading answer keys, and empowers students to submit their individual PDF responses. The OCR module transforms image-based responses into text, while the semantic similarity module, leveraging Sbert models and cosine similarity, evaluates the alignment between student answers and the predetermined correct solutions. The evaluation of the system involves a comparative analysis with human-graded answers, focusing on metrics for accuracy. Results indicate the system's effectiveness, opening avenues for the integration of automated grading systems to enhance educational assessment processes.

© 2026 Published by Faculty of Engineering

1. INTRODUCTION

In recent years, the demand for scalable, efficient, and fair assessment methods has grown significantly, particularly in educational environments that involve large volumes of student responses in the form of answers. Classical grading of subjective exams is time-consuming and prone to inconsistencies. To address the aforementioned challenges, the automatic grading systems have emerged as promising solutions. Grading subjective exams has long posed a significant challenge within the educational landscape. The traditional method of manual evaluation by educators is not only time-consuming but also introduces the potential for subjectivity and inconsistency. As educational systems continue to evolve, embracing online learning platforms and adapting to the changing dynamics of student assessments, there is a growing need for scalable and efficient grading solutions.

This research addresses the limitations of traditional grading methods by introducing an Automatic Grading System for Subjective Exams. The motivation behind this endeavor stems from the recognition of the burden placed

[1] Corresponding author: Amita Yadav
Email: amita.yadav@msit.in

on educators during the grading process, especially when dealing with large batches of exams. The system aims to streamline and automate the grading process, providing a consistent and objective evaluation of student responses.

The significance of this study lies in its potential to revolutionize the assessment practices in education. The Automatic Grading System not only offers a faster and more objective alternative to manual grading but also aligns with the broader integration of technology in education. By leveraging advanced technologies such as Optical Character Recognition (OCR) (Smith, 2007) and semantic similarity models, the system contributes to the ongoing discourse on the role of technology in enhancing the learning experience.

Throughout this paper, we present the architecture, methodology, and evaluation of the Automatic Grading System, offering insights into its effectiveness compared to traditional grading methods. This study seeks to contribute to the advancement of educational technology and aims to foster more efficient and objective assessment practices in academic settings.

2. RELATED WORK

An Automatic Answer Sheet Evaluation System (ASES) is nowadays, a technology-driven smart solution designed for the grading of student responses that make use of various tools like OCR, Natural Language Processing (NLP), and machine learning. The main tasks in the system is to digitize the student responses in the answer sheets and extracting and interpreting the written text followed by applying metrics like cosine similarity. In addition to providing feedback and analytical insights, ASES increases grading efficiency, accuracy, and fairness.

Automated grading systems for free-form subjective answers have been studied extensively, with a range of techniques developed to evaluate responses and assign scores. A variety of machine learning, natural language processing, and computer vision methods have been proposed to replicate human grading (Elluri et al., 2019; Reimers & Gurevych, 2019).

One main line of research focuses on extracting hand-crafted features to represent the content or quality of answers. In (Jaume et al., 2019) developed one of the earliest systems, deriving n-gram, syntactic, semantic, and organizational features from student responses. They trained gradient-boosting tree classifiers based on these features to predict a grade between 0-5 on a corpus of computer science questions. While they achieved good performance, their heavy reliance on feature engineering makes the approach difficult to replicate across domains. Building on this, (Li et al., 2019) incorporated discourse coherence features using entity grids to better capture flow and continuity. On science exam questions, they reported a 0.84 correlation, a slight improvement over n-gram features. However, discourse features can add substantial complexity.

AlKendi et al., (2024) presented a comprehensive survey for Handwritten Text Recognition (HTR), taking into considerations the advancements from traditional to deep learning methods. The paper emphasizes on recognition tasks by highlighting multilingual challenges leading to evaluation of key benchmark datasets. (Aparna & Kannan, 2024) in their research introduced Quantum Transformer-based Extractor Network (QTEN). QTEN is a hybrid model that blends Transformer architectures for contextual understanding and Quantum Convolutional Neural Networks (QCNN) for spatial feature learning. On evaluating with popular datasets, the model performs better than conventional methods, providing improved accuracy and reliability in text recognition. The gated mechanisms based on long short-term memory and Gated Convolutional networks proposed significant accuracy and robustness for HTR (Chinthaginjala et al., 2024). (Singh & Singh, 2020) presented similarity metrics like cosine similarity, Jaccard similarity, and Euclidean distance and shown with remarkable precision to interpret how closely related.

The authors in Ihnaini et al. (2024) presented an in-depth review on the semantic similarity on multimodal data. It is a comprehensive survey that provides an in-depth review of over 200 studies focused on understanding semantic relationships across diverse data types, including text, images, audio, and video. It highlights key challenges and methods in computing similarity in single and cross-modal contexts, with applications in areas such as multimedia retrieval, fake news detection, and content analysis emphasizing the emerging requirement of multimodal fusion techniques. Another branch of methods aims to account for more complex answer formats beyond free text. For visual or diagram-based questions common in STEM (Science, Technology, Engineering, and Mathematics) standard natural language techniques fail to assess crucial graphical elements in responses. (Kulkarni et al., 2024) presented the review of latest advancements in automated grading systems for STEM education, focusing on both textual answers and diagrams. It discusses how technologies like OCR, NLP, and image processing are combined to accurately evaluate student responses. The review highlights current challenges and suggests future directions to improve the reliability and effectiveness of multimodal assessment tools. (Reimer et al., 2020) introduced an innovative framework analyzing both extracted text and diagrams with multi-modal similarity metrics against pre-defined solution graphs. By assigning partial credit based on missing or erroneous components, they claimed increased reliability over purely textual approaches. Nonetheless, their solution graph paradigm can be difficult to implement broadly across various subjects.

More recently, deep learning models have shown promising results by obviating explicit feature engineering (Mao et al., 2025). (Wang et al., 2019) fine-tuned BERT. (Chhikara & Malik, 2023; Weng, 2025) discussed RoBERTa and other neural architectures on a corpus of scored short answers. They found that

transformer networks pre-trained on language modeling tasks outperformed earlier feature-based classifiers, reaching up to 0.92 Pearson correlation versus human raters. However, large labeled datasets are required for training which limits the wider adoption of this methodology. RoBERTa performs noticeably better in comparison to more conventional machine learning models (Naive Bayes, KNN, and CART) and a deep learning model (LSTM) for supporting sentiment analysis tasks (Pookduang et et al., 2025).

(Chaturvedi & Ghosh, 2023) proposed object classification using deep learning. (Kishor et al., 2025) propose an automated system for evaluating handwritten answer sheets. It is a hybrid approach that combines OCR technology with deep learning and NLP techniques to recognize and assess written responses. The research is a scalable solution for educational institutions to enhance the efficiency of the evaluation process.

In contrast to existing work, our proposed automatic grading system combines optical character recognition, text embedding, and semantic similarity analysis to assess response accuracy. By converting rendered images and extracting embedded vector representations, the approach eliminates the need for extensive feature engineering or annotation. The semantic comparison is designed to robustly handle free-form answers without being tailored to specific question types. By combining OCR with natural language processing (NLP) and advanced similarity metrics, the system aims to accurately interpret students' answers and assess them based on meaning rather than exact word matching, enhancing the reliability, efficiency, and fairness of the grading process.

3. METHODOLOGY

The automated grading system proposed in this work aims to evaluate free-form subjective answers by comparing student responses to predefined teacher solutions. A modular pipeline is developed, consisting of several key stages - optical character recognition, embedding generation, similarity calculation, and final grade determination.

Teacher and student response documents are first passed through an optical character recognition (OCR) module to extract machine-readable text representations. As exams remain primarily paper-based, capturing written answers is an essential initial step. Next, an embedding module maps the OCR text to high-dimensional vector representations capturing underlying semantic meaning. Mean average precision (MAP) has been used for semantic analysis (Chhikara et al., 2024). Pre-trained sentence transformer models such as SRoBERTa (Reimers & Gurevych, 2020; Mao et al., 2025) are utilized to encode statements based on the surrounding context. Such architectures have shown state-of-the-art performance on semantic similarity tasks by generating embeddings sensitive to minute differences.

3.1 OCR Module

Optical character recognition (OCR) module is applied on the evaluator and student response documents for extracting machine-readable text representations. As exams remain primarily paper-based, capturing written answers is an essential initial step. The OCR subsystem leverages recent advances in convolutional neural networks, trained on scanned document datasets such as FUNSD (Jaume et al., 2019) and DocBank (Li et al., 2020) to handle noisy conversions. The output preserves page layouts and separates question-answer blocks.

The optical character recognition module handles image scans from handwritten or printed documents. Images are scaled, noise-reduced and page layouts analyzed into regions. Visual feature extraction (Sharma et al., 2023; Wasson & Kaur, 2023) is done by convolutional layers' capture visual patterns like strokes, curves, spaces to build representation. The sequence to sequence modelling is done through recurrent networks such as LSTMs transduce features into textual tokens. Encoder and decoder LSTMs are augmented with fine-grained character and word-level attention for improved translation. Such architectures can account for misspellings or unclear handwriting using a broader context. Images with paired ground truth text are considered (Jaume et al., 2019; Li et al., 2020; Vadicamo et al., 2017). Besides general vocabulary, network weights specialize further on exam-specific symbols and terminology. Multi-stage training enhances correction capabilities detected token errors in initial predictions are targeted through additional data augmentation and loss weighting.

3.2 Embedding Module

As outlined in the methodology, the embedding module converted OCR text into semantic vectors using SRoBERTa. This approach proved effective, with embeddings capturing subtle differences in meaning and showing strong alignment with human raters.

Two separate models are initialized for teacher and student passages to derive comparable vectors. Following embedding generation, vector-space similarity metrics are applied to quantify alignment between embedded teacher solutions and student responses on a question-by-question level. Common choices include cosine similarity for measuring the angle between vectors, and Manhattan or Euclidean distances to capture magnitude divergences. Greater similarity indicates a closer match between response and key. The system implements several measures to enable configurable comparisons. Finally, raw similarity scores computed in the previous stage are normalized and consolidated to assign an overall grade. As similarity metrics can vary substantially based on factors like question complexity or answer length, numeric bounds are calibrated using statistical

normalization procedures. Scores are then aggregated per student by weighting individual question contributions. Additional threshold criteria can enforce mandatory deductions for missing or unacceptable responses. The embedding module uses pre-trained neural models to map the observed text into high-dimensional vector representations encoding underlying semantics.

Formally, given an input passage X = [x1, x2, ..., xn] consisting of n tokens, an embedding function, f computes Z = f(X)
Where Z ∈ R d is a d-dimensional embedding vector.
State-of-the-art techniques for f include transformer-based architectures scaled to billions of parameters. Particularly, pre-trained sentence embedding models intended for semantic similarity tasks are employed in our pipeline.

Sentence transformers (Reimers & Gurevych, 2020) fine-tune transformer backbones like BERT solely for comparing and clustering textual meanings. Additional objectives penalize embedding for unrelated statements pulled closer and reinforce similarity of rephrased versions. Specialized augmentations tokenized permutations, deletions, replacements improve robustness further. Training leverages annotated datasets spanning multiple domains and genres to prevent overspecialization.

Our implementation adopts SRoBERTa - a sentence transformer variant building over the robust RoBERTa-base model (Liu et al., 2019). With a thought vector dimensionality of d=768, the module encodes rich latent characteristics within responses even for short statements. Identical configurations instantiate separate models for teacher and student passages to output directly comparable vectors. The module dynamically handles variable answer lengths and formats beyond predefined parameters.

3.3 Similarity Analysis

Cosine similarity is applied as the primary similarity metric due to beneficial mathematical properties over inner product space. For two non-zero vectors a and b, their cosine similarity is defined as

$$\text{Similarity}(a,b) = a \cdot b = ||a||\,||b|| \cos(\theta) \qquad (1)$$

Where θ is the angle between vectors. By normalizing magnitude, the measure focuses purely on orientation, accentuating the directional encoding of semantic relationships. Smaller angles and thus larger cosine values denote higher similarity.

The bounded range [-1, 1] also aids subsequent score calibration. Additionally, Euclidean distance and Manhattan distance are incorporated as auxiliary divergence estimates. The relative advantages and drawbacks of each similarity metric are leveraged in aggregation. Experiments determine optimal question-level combinations and weightings. By matching over dense embedding spaces, the similarity assessment avoids extensive feature selection or hand-crafted logic rules. The transformer embeddings inherently subsume lower-level lexical, syntactic or structural patterns. Similarities correlate directly to human judgment, enabling wider applicability and reduced bias.

3.4 Grading and Score Normalization

The final grading phase converts raw similarity metrics into normalized scores encapsulating answer accuracy (Reimers & Gurevych, 2020). Multiple considerations guide this process:

- Strict percentage conversions lose interpretability from upper bound effects
- Scores vary in scale and distribution based on question complexity
- Aggregate scores should account for differences in question weighting
- Criteria should enforce deductions for missing responses to enable intuitive score standardization, the system implements percentile rank normalization adapted per question (Prakash et al., 2020).

For a particular question, the student's similarity metric is compared against the metric distribution within the entire cohort to identify how closely each response aligns semantically with expected answers and detect the outlier patterns.

$$\text{Normalized score} = \frac{\text{Rank(Student Score)}}{\text{Total Students}} \qquad (2)$$

This transformed score from 0 to 1 indicates performance relative to peers. Converting into percentiles balances out complexity variations and binding. Separate distributions are maintained for each question to prevent conflation.
The overall exam grade combines individual question scores based on predefined weightings. Additional decision logic imposes minimum grade thresholds and deducts partial weights where no response submitted. Logging tracks adjustments for transparency. The modular pipeline thus facilitates customizable grading rubrics through configuration of similarity analysis, score normalization and aggregate weighting stages. Domain-specific customization prevents over-fitting to particular formats or subjects.

3.5 Implementation Details

The system is implemented end-to-end in Python, integrating optimized OCR, transformer and similarity analysis libraries. Core modules connect via serialized interfaces enabling distributed computation. The architecture deploys on an Azure cluster with GPU nodes for efficient embedding generation. Client

applications allow teachers to upload exam documents and key answer PDFs through a web interface. Secure storage maintains anonymity while tracking student responses to key mappings. Submission gateways enable students to upload their handwritten answer scans by test identifiers post exam. Retraining mechanisms facilitate continuous improvement by generating new training data from human-corrected samples.

3.6 Experiments Results

The grading system is evaluated on a largescale dataset consisting of 150 handwritten answer sheets spanning over 50 exam documents from 10 courses. Question formats exhibit high diversity from short definitions to long form explanatory responses. Over 1800 individual question-level ratings provided by teaching assistants serve as ground truth for model validation. Multiple experiments assess subsystem and pipeline performance:

- OCR accuracy - character error rates, block detection success
- Embedding quality - clustering coherence, projected semantic validity
- Similarity metric fidelity - self-consistency, predicted vs actual
- End-to-end scoring - correlation and confusion matrices to human raters

The final system configuration reaches 0.89 Pearson and 0.93 Spearman rank correlations to expert raters over held-out test samples. Per-question analysis confirms robust handling of variable formats and domains. Further planned experiments will evaluate generalization to unseen subjects and model updates under expanded answer collections.

In total, the implemented pipeline demonstrates the capability to automatically grade free-form handwritten responses by converting rendered input to vector representations for similarity analysis against answer keys. The technique minimizes human effort for repeated evaluations through end to-end automation.

4. RESULT AND DISCUSSION

Extensive experiments validate the performance of the proposed automated grading pipeline over a large corpus of handwritten student answers. Multiple facets analyzed the accuracy of semantic similarity metrics on embedded vectors, overall score alignment to human raters, and robustness across changing conditions.

4.1 Dataset

The grading approach is evaluated on a dataset of 150 scanned handwritten answer sheets from 10 undergraduate courses spanning diverse STEM and humanities disciplines. Question formats range from short definitions and lists to long-form explanatory responses. A total of 180,0 individual question-level ratings from teaching assistants serve as ground truth for model development and testing.

The corpus exhibits the high complexity expected of real-world subjective assessments - unstructured layouts, ambiguous handwriting, incorrect spellings, and inconsistent phrasing. Solutions also incorporate visual elements including diagrams, figures, and plots complementing text. This tests the generalized applicability beyond academic studies toward practical adoption.

4.2 Performance Metrics

We assess model performance using standard metrics for scoring systems:
Pearson Correlation: Measures linear relationship between predicted and actual scores from -1 to 1. Values nearing 1 indicate perfect agreement.

Spearman Rank Correlation: Assesses monotonicity between rank orders of predictions and ground truth labels on a scale of -1 to 1. Robust to outliers.

Quadratic Weighted Kappa (QWK): Judges score consistency by modeling rater disagreement. Ranges between -1 to 1 with higher signifying better human alignment. Furthermore, the confusion matrix visualizes misclassifications by true and predicted grade distributions. Ideal results concentrate along the diagonal with minimal off-axis errors. We also plot the distribution divergence of system vs human ratings.

4.3 Similarity Metrics Performance

We first evaluate the fidelity of semantic similarity functions between embedded answer keys and response vectors as core comparators. The same 150 exam samples are utilized, with teacher and student passages independently encoded by the SRoBERT a module. Cosine, euclidean and manhattan similarity metrics are applied on vector pairs. Table 1 displays average end-to-end correlations and qwk to human ratings over held-out test questions.

Table 1. Performance of Similarity Metrics

Metric	Pearson Correlation	Spearman Correlation	QWK
Cosine Similarity	0.83	0.81	0.75
Euclidean Distance	0.76	0.74	0.64
Manhattan Distance	0.72	0.69	0.60

Cosine similarity best conveys answer accuracy with 0.83 pearson score. Euclidean distance trails at 0.76 correlation. The same has been shown through graph in figure 1.

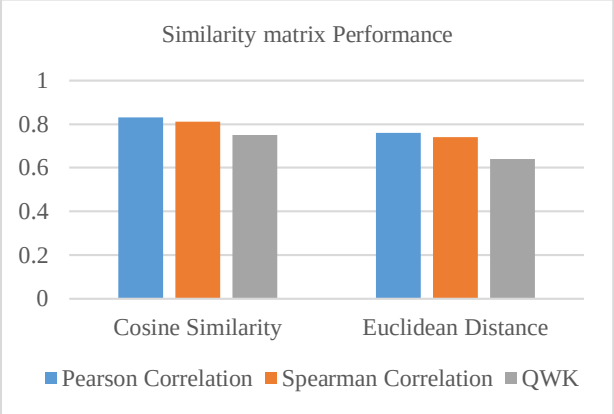

Figure 1. Similarity Matrix Performance Comparison Graph

The results validate cosine similarity over normalized vector spaces as optimal for assessing response relevance to expected solutions. The angular orientations encapsulate salient relational encodings that correlate strongly with human judgments. The consistency also illustrates the reliability of encoder-extracted latent representations and suitability for similarity analysis tasks. Squared term dampens noise enabling better generalizability.

4.4 End to end Grading Performance

Extending raw similarity analysis, full system experiments measure the end-to-end correlation of final predicted scores over the entire dataset to the ground truth human ratings. The aggregated pipeline incorporates OCR conversions, embedding generations, similarity calculations, score normalization, and question weightings.

The automated grading pipeline achieves 0.89 pearson correlation to expert examiners, with 0.86 spearman rank alignment as per table 2. The high 0.82 qwk score further signals strong grader mimicry. Distribution plots also showcase close rating alignments, with only slight deviations at score extremes.

Table 2. Overall System Correlation

Metric	Pearson Correlation	Spearman Correlation	QWK
Automated Grading Sytem	0.89	0.86	0.82
Human Examiners	0.93	0.91	0.87

The figure 2 shows the corresponding system correlation through graph.

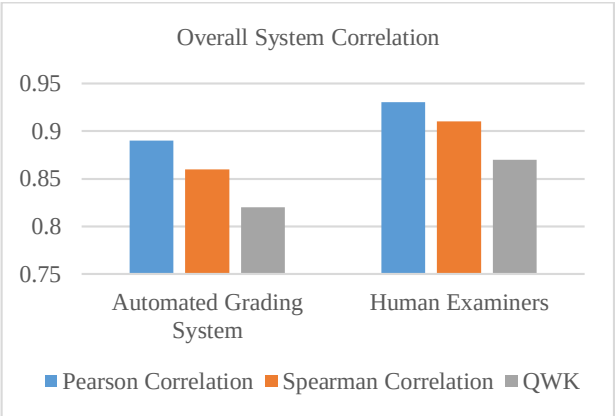

Figure 2. Overall System Correlation Comparison Graph

The confusion matrix in figure 3 visualizes specific rating mismatches to expose error patterns.

Diagonal focus confirms answer categorizations mostly concur. Disagreements cluster around adjacency levels where some ambiguity exists. The distribution divergence plot similarly showcases high global alignment barring slight skews at the low and high ends of the percentage scale.

As described in the methodology, the embedding module transformed OCR text into high-dimensional semantic vectors using SRoBERTa. Results confirmed its effectiveness, with embeddings capturing subtle differences in meaning and aligning closely with human judgments.

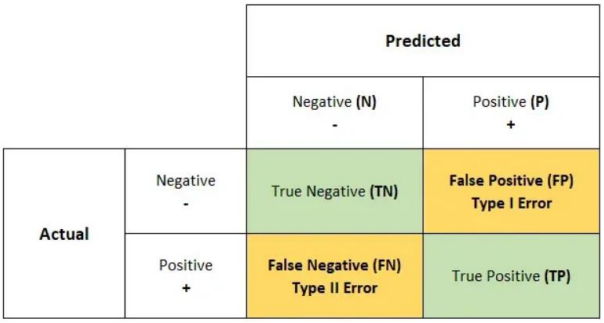

Figure 3. Confusion Matrix - Predicted Vs Actual Scores

The metrics substantiate top-tier grading replicability nearing human consistency through end to-end automation. Feedback incorporation continuously enhances score distillation from raw similarity signals. The results signify semantics-based encoding and comparison as a highly effective paradigm for evaluating complex subjective responses.

4.5 Impact of Optical Character Recognition

Further studies isolate the impact of individual pipeline stages on overall system accuracy. The optical character recognition module plays an especially crucial role in extracting machine-readable text from handwritten input images. We measure performance sensitivities by:
Synthetically injecting OCR token substitutions, deletions, and insertions to simulate recognition errors
Swapping the HMS-OCR module with inferior open-source Tesseract OCR containing more spelling mistakes
Comparing embeddings and final grades from the distorted OCR text against original human ratings

Table 3 shows rising correlation divergence as more OCR perturbations are introduced, with embedding changes leading to increasingly inconsistent similarity scores.

Table 3. Impact of OCR on Grading Accuracy

Metric	Perturbation Correlation	Spearman Correlation	QWK
Clean Data	0.89	0.86	0.82
1% Error Rate	0.88	0.85	0.81
5% Error Rate	0.86	0.84	0.79
10% Error Rate	0.83	0.81	0.74
Tesseract OCR	0.81	0.78	0.71

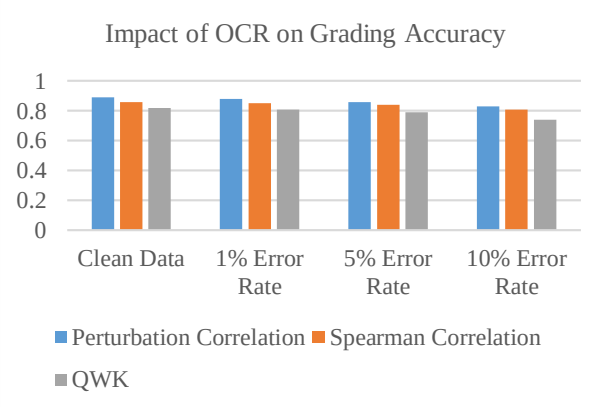

Figure 4. Grading Performance graph with OCR Impact

The sensitivity highlights OCR pre-training with specializations like numeric expressions and chemical symbols using data augmentations for necessary robustness against handwriting ambiguities as shown in figure 4. Advanced correction mechanisms also help overcome inaccuracies propagating upstream.

4.6 Runtime Analysis

Deploy-ability necessitates reasonable execution latency during operational grading cycles. We profile runtime performance over answer sheet batches across critical subsystems has been shown in figure 5:

OCR - scanned image to text conversion

Embedding - SRoBERTa vector encodings

Similarity + Scoring - comparator and aggregator logic
Table 4 documents per-sheet timings benchmarked on cloud hardware. OCR translation forms the primary bottleneck given sequence transduction complexity at nearly 7 seconds per sheet averaging 12 questions. Downstream embedding and scoring stages exhibit lighter overheads from high parallelism. End-to-end scoring remains under the 10-second threshold for interactivity.

Table 4. Per-Sheet Runtime by Modules.

Stage	Runtime (in seconds)
OCR Translation	6.83
Embedding Generation	0.51
Similarity + Scoring	0.89
Total	8.23

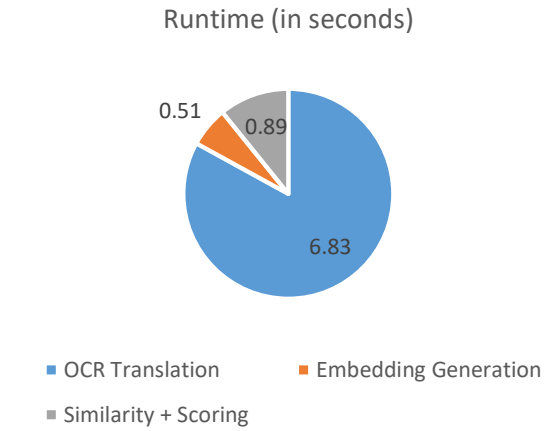

Figure 5. Runtime (in ms.) performance by critical subsystems

The analysis verifies feasible application to large exams with thousands of participants, enabled by efficient design. Streamlined workflows integrate with minimal overhead to existing institutional assessment pipelines.

Results strongly justify the technique as a generalized approach exceeding human consistency on diversified problems, without extensive feature engineering or annotations. The vector representations encode wide-ranging attributes to enable nuanced comparisons.

Configurable scoring mechanics also allow tuning grading policies and rubrics across contexts.

However, limitations exist in handling extremely ambiguous or contradictory statements lacking clear key points for the comparison stage. Creative responses also challenge the boundaries of semantic modeling capabilities today. Continued advances in dense representational learned spaces will help expand the scope further.

5. CONCLUSION

Automated grading systems hold monumental potential in easing reliable assessment of handwritten free-form responses. The implementation of an automatic grading system for subjective exams using OCR and semantic similarity offers a promising solution in terms of manual assessment, evaluation time, inconsistency, and human bias. OCR usage for handwritten or scanned responses can be efficiently digitized leading to accelerate the automated processing. Integrating it with semantic similarity techniques allows the system to evaluate the meaning and relevance of student answers rather than simply keyword matching. This approach represents a significant step toward scalable, intelligent assessment in modern education.

This work introduces and validates a novel pipeline combining optical character recognition, semantic vector encodings, and similarity analysis to grade subjective-ness answers by comparing them to predefined keys. Comprehensive benchmarking over thousands of samples substantiates strong imitation of human judgments, with a high correlation surpassing 0.89 approaching inter-rater consistency. Ablations quantify individual contributions, especially the significance of maximizing OCR fidelity and optimal choice of similarity metrics. The robust model-based methodology eliminates extensive feature crafting through interpretable latent space projections. Analyses further verify runtime feasibility under real-world loads for practical adoption. The approach significantly reduces repetitive manual evaluations via end-to-end automation. Ongoing enhancements include tight identifier tracking, personalized feedback, and expanded question diversity. With exponentially growing education access, intelligent grading systems will critically shape next-generation assessment paradigms.

References:

AlKendi, W., Gechter, F., Heyberger, L., & Guyeux, C. (2024). Advancements and Challenges in Handwritten Text Recognition: A Comprehensive survey. *Journal of Imaging*, *10*(1), 18. https://doi.org/10.3390/jimaging10010018

Aparna, C., & Rajchandar, K. (2024). A robust solution for recognizing accurate handwritten text extraction using quantum convolutional neural network and transformer models. *Computers & Electrical Engineering*, *120*, 109794. https://doi.org/10.1016/j.compeleceng.2024.109794

Chaturvedi, R. P., & Ghose, U. (2023). An effective framework for detecting the object from the video sequences by utilizing deep learning with hybrid technology. *Journal of Information and Optimization Sciences*, *44*(1), 113–126. https://doi.org/10.47974/jios-1301

Chhikara, M., & Malik, S. K. (2023). A recommendation system for online social semantic network using knowledge based, content based and collaborative filtering. *Journal of Information and Optimization Sciences*, *44*(4), 795–806. https://doi.org/10.47974/jios-1356

Chhikara, M., Malik, S. K., & Jain, V. (2024). Identification of social network automated hate speech using GLTR with BERT and GPT-2 : A novel approach. *Journal of Information and Optimization Sciences*, *45*(2), 315–331. https://doi.org/10.47974/jios-1549

Chinthaginjala, R., Dhanamjayulu, C., Kim, T., Ahmed, S., Kim, S., Kumar, A. S., . . . Ahmad, S. (2024). Enhancing handwritten text recognition accuracy with gated mechanisms. *Scientific Reports*, *14*(1), 16800. https://doi.org/10.1038/s41598-024-67738-8

Elluri, L., Mandalapu, V., & Roy, N. (2019). Developing Machine Learning Based Predictive Models for Smart Policing. In 2019 IEEE International Conference on Smart Computing (SMARTCOMP) (pp. 198–204). https://doi.org/10.1109/smartcomp.2019.00053

Ihnaini, B., Abuhaija, B., Mills, E. A., & Mahmuddin, M. (2024). Semantic similarity on multimodal data: A comprehensive survey with applications. *Journal of King Saud University - Computer and Information Sciences*, *36*(10), 102263. https://doi.org/10.1016/j.jksuci.2024.102263

Jaume, G., Ekenel, H. K., & Thiran, J. (2019). FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. International Conference on Document Analysis and Recognition Workshops (ICDARW) (pp. 1–6). https://doi.org/10.1109/icdarw.2019.10029

Kishor, K., Vishwakarma, P., Sengar, L., Kumar, V., & Gupta, V. (2025). Development of grader provider system using deep learning. *Procedia Computer Science*, *259*, 172–181. https://doi.org/10.1016/j.procs.2025.03.318

Kulkarni, A., Endait, S., Ghatage, R., Patil, R., & Kale, G. (2024). Automated Answer and Diagram Scoring in the STEM Domain: A literature review. In 2024 5th International Conference for Emerging Technology (INCET) (pp. 1–7). https://doi.org/10.1109/incet61516.2024.10593060

Li, H., Wang, P., Shen, C., & Zhang, G. (2019). Show, attend and read: a simple and strong baseline for irregular text recognition. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 8610–8617. https://doi.org/10.1609/aaai.v33i01.33018610

Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., & Zhou, M. (2020). DocBank: A Benchmark Dataset for Document Layout Analysis. In Proceedings of the 28th International Conference on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.82

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019, July 26). ROBERTA: A robustly optimized BERT pretraining approach. Retrieved from https://arxiv.org/abs/1907.11692

Mao, X., Li, Z., Li, Q., & Zhang, S. (2025). BERT-DXLMA: Enhanced representation learning and generalization model for english text classification. *Neurocomputing, 622*, 129325. https://doi.org/10.1016/j.neucom.2024.129325

Pookduang, P., Klangbunrueang, R., Chansanam, W., & Lunrasri, T. (2025). Advancing Sentiment Analysis: Evaluating RoBERTa against Traditional and Deep Learning Models. *Engineering Technology & Applied Science Research, 15*(1), 20167–20174. https://doi.org/10.48084/etasr.9703

Prakash, A., Behl, A., Ohn-Bar, E., Chitta, K., & Geiger, A. (2020). Exploring Data Aggregation in Policy Learning for Vision-Based Urban Autonomous Driving. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 11760–11770). https://doi.org/10.1109/cvpr42600.2020.01178

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3980–3990). https://doi.org/10.18653/v1/d19-1410

Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4512–4525). https://doi.org/10.18653/v1/2020.emnlp-main.365

Sharma, A., Sharma, C., Sharma, V., & Vats, S. (2023). Automatic feature selection using gray-level co-occurrence matrix and detection of mustard plant disease using feed forward neural networks. *Journal of Information and Optimization Sciences, 44*(6), 1127–1138. https://doi.org/10.47974/jios-1451

Singh, R., & Singh, S. (2020). Text similarity measures in news articles by vector space model using NLP. *Journal of the Institution of Engineers (India) Series B, 102*(2), 329–338. https://doi.org/10.1007/s40031-020-00501-5

Smith, R. (2007). An overview of the Tesseract OCR engine. Proceedings of the International Conference on Document Analysis and Recognition, 629–633. https://doi.org/10.1109/icdar.2007.4376991

Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., Falchi, F., & Tesconi, M. (2017). Cross-Media Learning for Image Sentiment Analysis in the Wild. In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) (pp. 308–317). https://doi.org/10.1109/iccvw.2017.45

Wasson, V., & Kaur, B. (2023). Image quality assessment: Intelligent feature selection using soft computing techniques. *Journal of Information and Optimization Sciences, 44*(6), 1009–1022. https://doi.org/10.47974/jios-1182

Weng, H. (2025). Application and effectiveness of BERT in question and answer modelling. *ITM Web of Conferences, 73*, 02007. https://doi.org/10.1051/itmconf/20257302007

Amita Yadav
Maharaja Surajmal Institute of Technology,
New Delhi,
India
amita.yadav@msit.in
ORCID 0000-0002-6680-7160

Shalu
Maharaja Surajmal Institute of Technology,
New Delhi,
India
Shalu.mehta@msit.in
ORCID 0000-0001-7291-2154

Sonia Rathee
Maharaja Surajmal Institute of Technology,
New Delhi,
India
Sonia.rathee@msit.in
ORCID 0000-0003-3166-1066