Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on Integrating Science, Arts, Commerce and Technology for Future Innovations 7(12s)

ACADEMIC ASSESSMENT BASED ON FINE TUNED LLMS AND NEURAL FEATURE BASED DIAGRAM EVALUATION

Rupa Rani [1], Sachin Jain [2], Mukulit Goel [3], Anuj Kumar [4]

[1] Assistant Professor, Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh, India
[2] Associate Professor, Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh, India
[3] Assistant Professor, Department of MCA, Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh, India
[4] Associate Professor, Department of Computer Science and Engineering, JB Institute of Technology, Dehradun, Uttarakhand, India

Received 14 March 2026
Accepted 13 May 2026
Published 27 May 2026

Corresponding Author
Sachin Jain, sachincs86@gmail.com

DOI
10.29121/shodhkosh.v7.i12s.2026.8237

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

The process of evaluation traditionally carried out on the answer sheets is laborious, inconsistent, and error prone. Automation of this process will significantly improve both speed and accuracy. Compared to the existing automated answer sheet evaluation solutions; while exploring subjective and essay-type answers, they did not capture the deep semantics of the language. Effective analysis of Diagrams remains neglected. This paper proposes a multimodal approach for answer sheet evaluation, which integrates both textual and visual elements. The contribution of the proposed approach is twofold: firstly, the evaluation of textual responses is improved by state-of-the-art natural language processing and fine-tuned large language model Llama 3.2; secondly, diagram evaluation has been enhanced with a neural network-based feature matcher, LightGlue, further complemented by a custom image preprocessing pipeline, integration of OCR, and NLP metrics to improve diagram feature evaluation accuracy and thus allow for the precise extraction and analysis of diagram labels. Experimental results reveal that our system achieves very good accuracy and consistency comparable to those of human evaluators. However, system performance may degrade due to digitization quality, such as poor handwriting or an unclear image. In conclusion, the proposed system overcomes the existing gaps in automated evaluation methods and hence provides a holistic solution to assess answer sheets.

Keywords: Automated Academic Assessment, Large Language Models (LLMs), Neural Network, Finetuning, Optical Character Recognition, Diagram Detection and Matching, Prompt Tuning

1. INTRODUCTION

Educational institutions worldwide assess their respective candidates' skills and knowledge through various methods, primarily through objective and subjective questions. Objective questions, which require candidates to select an option from a list are easily evaluated through automation. In contrast, automating the evaluation of subjective question answers is a bit trickier. For one thing, subjective answers are much more flexible and open to interpretation.

They tend to vary widely in their structure, tone of language, vocabulary and presentation. People often use synonyms, descriptive language and convenient abbreviations making each of their responses unique. The length of the answer is also a flexible attribute. Therefore, keyword extraction and matching cannot be the sole basis for assessment as we are dealing with Natural Language. The traditional way, i.e. the manual grading of answer sheets can be slow-going and exhaustive. But this is just one of the issues. Often the marks assigned to students on their exam sheets depend on the mental state and concept clarity of the examiner rather than on the answers written. In addition, there is no consistency in evaluation of subjective or long answer type questions. Different examiners can assign different marks for the same answer. The same examiner can also assign different marks for the same answer at different timeframes. These variations in grading highlight the lack of a defined set of rules or rubric for evaluation of long, subjective and essay type answers. Sometimes potential biases like halo effect, anchor bias or logical fallacy may also be introduced by manual grading. And there is the added issue of transparency where students have no way of knowing on what basis marks were assigned to them. To address these issues, we are putting out a thorough automated approach that is intended to assess subjective responses in an all-encompassing manner.

This paper is organized as follows: (i) Section 2 summarizes the previous research conducted in this area. (ii) Section 3 Part A gives a detailed description of the tools and technologies integrated into this system along with the mathematical equation used for marks calculation. (iii) Section 3 Part B is about the complete algorithm followed while developing this system (iv) Section 3 Part C focuses on the methodology and stepwise processing during answer sheet grading. (iv) Section 4 presents the experimental results obtained during model training and implementation. (v) Section 5 demonstrates a detailed comparison of our proposed system with earlier research. (vi) Section 6 contains the conclusion.

2. LITERATURE REVIEW

(Nikhil, 2024), have proposed an automatic answer sheet checking system that evaluates the an-swers written in an exam sheet by matching original or reference answers, based on some key-words and grammatical error checking algorithms. This system works by first scanning the imag-es of handwritten exam papers and then uploading them to the system so that it can evaluate and assign the score of each answer. It integrates various Mathematical Model Heuristic Rule-based algorithms, Comparison algorithms, and a Text-to-Text generative AI transformer. This approach addresses many challenges like establishing contextual understanding, reducing human bias, and improving model accuracy rates which are approximately around 90-95%.

(Nayudu & Rani, n.d.), have introduced an Efficient Exam Grading system with AI-Powered Answer Verification or in other words an Online Subjective Answer Grading application that will help improve the skills of students and enhance their personalities. This application moti-vates students to engage with the subject matter deeply, get a stronger grasp of concepts, and write well-explained subjective answers which further helps to improve their writing skills. The proposed system aims to assist the subjective answer evaluation using Python, Flask, and inte-grating different machine learning and artificial intelligence algorithms. Many approaches have been proposed for the initial step of keyword extraction including TF-IDF, Count Vectorizer, and Yet Another Keyword Extractor (YAKE) tool. Then the BM25L algorithm and cosine simi-larity are applied for text summarization and document ranking. Finally, the BERT model is in-tegrated to measure the semantic similarity between the handwritten and expected answers. The future scope for this system involves working with various issues like evaluating answers con-taining multiple languages, improving the accuracy rates of the trained model, and decreasing the error rates further as compared to the current accuracy and error rates. The current accuracy of this model is around 88% without using MNB and the error rate is said to decrease by 31% after using the Multinomial Naive Bayes (MNB).

(Kulkarni et al., 2024), proposed a Digital Handwritten Answer Sheet Evaluation application to evaluate handwritten answer sheets fairly using a single set of rules /rubrics, assign marks, and provide student feedback. This application uses the PDF2Image python library to separate and extract individual images from the uploaded PDFs. After this Google Cloud Vision, a cloud-based OCR service has been integrated for text recognition, character segmentation, and feature extraction. This is followed by employing a BERT tokenizer which creates word embeddings of text to capture their semantic meaning. Marks are then calculated for each answer based on co-sine similarity and four other parameters. These parameters are The Total number of questions (Q), Maximum possible score (M), Content analysis (A), and Handwriting recognition (H) and they are assigned a score between 0 and 1 by thorough processing of extracted answers. Then the overall score (O) is calculated as the weighted sum of H and A with weights assigned to each parameter. Finally, the final score (F) is computed using the formula F=(O/M)*Q and dis-played as output. This paper introduced only text-

based evaluation using Google Cloud Vision OCR and cosine similarity but failed to assess non-textual elements like diagrams.

(Azubogu et al., 2024), proposed a solution for developing a computer-based test platform us-ing Natural Language Processing (NLP) to assess the descriptive answer examinations. This plat-form represents several issues inherent in manual Grading methodologies including slow turna-round times, potential for bias, and limited scalability. This system not only provides timely and accurate feedback but also helps assess students' critical thinking, problem-solving, and creativi-ty. It reduces the stress and workload on lecturers, thus improving the general learning and teaching experience. The manual method of descriptive answers is fraught with many challenges, such as the high level of stress involved and inherent subjectivity in grading and the like. delays in producing results. The use of NLP by the system for evaluation provides timely and accurate feedback on issues such as Grading bias, slow evaluation, and scalability limitations.

(Tan et al., 2025), proposed an approach to grade the theoretical content using AI Techniques. Dr. B. Vanathi et al. (Vanathi, 2023), proposed an Automated Exam Paper Evaluation System that implements various deep learning and NLP-based approaches to evaluate the handwritten answers of students, assign a score, and provide feedback. First, the preprocessing of handwrit-ten answer images is done to improve readability by techniques such as binarization, noise reduc-tion, smoothing, and component analysis. For the next task of handwriting recognition and con-version to processable text, two approaches have been defined. The first is the use of a Rapid API OCR service and the other is a model formed by the combination of two deep learning ar-chitectures - vgg19 and GRU neural networks. Next text preprocessing is done and several met-rics like Levenshtein distance and cosine similarity are applied for answer evaluation in compari-son to answer key. Keywords that appear frequently are assigned to a smaller weighted value and those with less frequency are assigned to a larger weighted value. Fuzzy string matching is also used to locate matching strings in case they are written with a different structure in student answers. Marks are then calculated based on weighted values and metrics calculated. The IAM dataset has been used for model training and compared to manual evaluation the proposed mod-el was found to be 75-85 % accurate.

(S et al., n.d.2023) , proposed an evaluation model that uses pre-trained Cloud Vision and Textract OCR models for handwriting recognition due to higher accuracy rates and short prediction times. The BERT model then has been used for generating contextual embeddings from text sequences and keyword extraction due to its superior encoder architecture and GPT-3 model is incorporated to generate summaries for long descriptive answers due to its superior decoder architecture. The keywords from the expected answer and the extracted answer will be fed to a similarity checking algorithm and marks will be assigned accordingly. The system also allows a user to view the marks allotment for checked answers including output from the handwritten recognition model and GPT-3 summaries in case he/she finds the evaluation unsatisfactory. The datasets used for model training include the IAM dataset, crowd-sourced handwritten samples, and pseudo-academy ESA manuscript samples. The OCR accuracy achieved by the handwriting recognition model was approximately 98.6% for a best-case (clean) document and 66.66% for a worst-case (sloppy) document.

(Sanuvala & Fatima, 2021), proposed a basic version of an Automated Rating System that generates scores for long handwritten answer scripts. This method effectively improves accuracy rates for answer evaluation as compared to manual grading by word error detection, comparing sentence similarity values and rating calculation using NLP integrated with STS (OSTS) approaches. Comparison and grading are conducted using various ML/ Clustering models like Logistic Regression, Naïve Bayes, Gradient Boost Decision Tree, and Support Vector Machine for high document grade and low document grade. This method resolves several issues like finding the (dis)similarity degree among the documents, selecting the right features from the document, and enhancing accuracy, precision, and recall rates by 98.80%, 92.26%, and 94.08% respectively. Reduced computation time for the same has also been achieved by introducing hashing technique.

(Bashir et al., 2021), proposed a subjective answer evaluation using machine learning and Natural Language Processing. This approach leverages machine learning models to give confidence and suggestions to Similarity-induced scores. Word2vec for word embedding that keeps the semantic meaning intact. A hybrid of cosine similarity and WMD combined with the MNB model is used. A threshold of values is selected: WDM_LOWER 0.7, WDM_UPPER 1.6; COS_LOWER 0.2 COS_UPPER 0.7 for acceptable matching. Addition or deduction of half the number of values in that range is made based on whether the model-suggested score is greater or lesser than the Similarity equivalent score. The finalized score is obtained if Similarity scores match with the Model predicted class. The score predicted by the module achieved 88% accuracy. With the model suggestion, the error rate decreases from 15.6% to 13.94%.

(Chandrapati & Rao, 2024), propose a novel ensemble model, the Descriptive Answer Evaluation System (DAES), which integrates Topic Modelling (TM) and Question Answering (QA) models to automatically evaluate descriptive answers. The proposed hybrid model using LDA and T5 incorporating the Cosine similarity method achieved an accuracy of 91%. It employs the use of SBERT for sentence embedding. A training set of 7200 instances is taken with dataset partition 8:2 for training and testing. Through rigorous experimentation, the model obtained- accuracy 91%, precision 91%, recall 92%, f1-score 91%.

(Säuberli & Clematide, 2024), proposed using large language models (LLMs) like Llama2 and GPT-4 to generate and evaluate MCRC items in a zero-shot setting. They introduce a new evaluation protocol, which includes a metric called text informativity, that combines answerability (how well the item can be answered based on the text) and guess ability (how likely an item can be guessed without the text). This approach enables the automatic evaluation of test items by eliciting responses from LLMs and comparing them to human annotators' performance. The challenge is twofold:

- Automating the generation of MCRC items.
- Developing a robust method to evaluate the quality of these items efficiently, especially in languages like German, where there is limited data.

(Meenakshi et al., 2022), developed a web-based application that evaluates the exam scores quickly by employing Natural Language Processing (NLP) to analyses the descriptive answers. This system's methodology comprises cosine similarity, information extraction, keyword matching, Rapid Automatic Keyword extraction-based processing module, and NLTK. An added feature of generated scores represented in a pictorial graph has also been discussed. Prospects of the system include improved accuracy rates by integrating a synonyms model in addition to the TF-IDF technique. This system focuses only on the English language.

(Salim et al., 2022), proposed Indonesian automatic short answer (ASAG) grading system to autotune and improve the process of evaluation. This system uses bidirectional encoder representations from the transformer (BERT) based models with fine-tuning and advanced feature extraction. It employs use of two different ASAG datasets. It achieved a score of 0.9508 in Pearson's correlation and 0.4138 in root-mean-square error (RMSE) by the BERT-based model. This demonstrates the effectiveness of the proposed method in grading short answers accurately.

(G & G, 2020), have developed an online subjective answer verifying system integrating numerous automated techniques such as Artificial intelligence-based answer verifier, score computation, grammar API, similarity algorithm, Fuzzy logic for QST (Question Specific Things), text Gear grammar API and Decision Based scoring algorithm. This approach effectively sorts out the challenge of variation in grammar checking based on standard requirements. The developed model is based on a keyword-matching approach and the current accuracy rate achieved is 80-90%. Future scope involves enhancing efficiency approximately by 95-98% which may lead to improved model performance.

(Al Mahmud et al., 2020), have offered a keyword-based approach to evaluate broad subjective answer scripts and several other types of test questions such as viva-voce and objective questions to evaluate student performance. The research study has suggested Latent Semantic Analysis (LSA), Intelligent Essay Answers, and Syntactically Enhanced LSA (SELSA) integrated system. Several limitations like evaluating answers with diagrams, tables, and mathematical expressions have also been addressed. After resolving those functions, it would be useful in all types of evaluation.

(Balat et al., 2020), proposed a solution for automatic exam evaluation based on Brain Computer Interface (BCI) that records the brain signals of the students and then pre-processes these signals by applying both low and high-pass filters, followed by subsampling and segmentation. The extracted features are fed into a Linear discriminant analysis (LDA) model, achieving an accuracy of 91%. This highlights the challenges students encounter during the exams such as anxiety, lack of focus, tension, and lack of attention, and this issue negatively impacts the assessment of learning outcomes and leads to the ineffective evaluation of students' performance.

(Rowtula et al., 2019), have proposed an automated evaluation system for handwritten answers that improved existing solutions for the same by introducing two new approaches - 1) treating it as a self-supervised, feature-based classification problem and 2) using Natural Language Processing (NLP) and Information Retrieval and Extraction (IRE) for semantic analysis. This solution involves using a self-supervised scoring model that calculates a score based on the keyword comparison between Textual Reference Answers (TRA) and handwritten answers. A MALLET framework is used for connecting words with similar meanings and distinguishing the uses of words with multiple meanings. Parts of Speech (POS) tagging and Named Entity Recognition (NER) to get another fresh set of features for answer evaluation. The neural network used is trained on various feature sets including base features (unique terms, keyword recall & word

count), lexical features (tokens & unique terms - token ratio), syntactic features (word length, term strength &token strength), and NLP features (nouns phrase ratio, verb phrase ratio & named entities match count). The datasets used for training are the Classroom Dataset (CRD), Controlled Dataset (CD), and SciEntsBank Dataset (SE).

(Rahman & Akter, n.d.), have proposed a NLP-based solution for automatic answer script evaluation. The images of handwritten answer scripts are uploaded and cleaned and then the Python pytesseract library is implemented for text extraction and digitization. Next text summaries are generated for each answer to get the gist of it, focus on the most important parts, and speed up the evaluation process. Various NLP-based approaches have been discussed for summary creation including the Bag-of-Words (BoW) technique. Several metrics like Precision, Recall, and F-score have also been calculated to check how fact-relevant a given generated summary is. Then various text pre-processing techniques like tokenization, stemming, lemmatization, etc. have been applied. After this, several similarity measurement techniques like cosine similarity, Jaccard similarity, bigram similarity, and synonym similarity are used to analyses structural similarity and keyword similarity of generated answer summary and reference text (answer key). Grammatical errors and spelling correctness have also been considered by utilizing a Python package language check. The previously mentioned similarity measures and grammar & spelling checks are finally used as parameters in calculating marks given for each answer where each parameter is assigned a weighted value based on its importance.

(Bhonsle et al., 2019), proposed an adaptive approach for subjective answer evaluation for the current academic environment which has already gained greater attraction in this field using online education resources. This research study suggests the k- k-nearest neighbor algorithm-based kernel method and attribute reduction for characteristics simplicity and efficiency using four UCI datasets. The issues addressed are a decrease in the rate of error, quick response of feedback, and exhibiting an accuracy comparable to human decision.

(Brown & Program, 2017) , proposed a novel approach for an automated grading system for numerical answers on tests, using Computer Vision. The project uses Contour detection and filtering for identifying the digit boxes and CNN is used for identifying 12 supported pixel classes [0-9, ., -] with size description. The grading is done with some tolerance level using harmonic means as the aggregate function. The harmonic mean of recall and precision is taken as the threshold function. Most of the samples used for the training process came from the MNIST handwritten digit dataset consisting of approximately 6500 handwritten samples of each digit between 0 and 9. 95.6% of answers were successfully graded by the model.

(Raina et al., 2023) proposed a solution utilizing the BERT model for grading answers. The proposed BERT model outperforms other models by achieving an accuracy of 90% when it comes to calculating text and language similarity. (Sridevi et al., 2019), proposed an android-based application into which all the handwritten exam sheets of students along with the answer key can be uploaded for computerized marks allotment. This application also tackles the issue of transparency by generating a report with a detailed assessment of a given student. The proposed model starts by performing adaptive binarization of all the uploaded images followed by component analysis. Then the Tesseract Python library is employed for text extraction and digitization. Next, the NLTK library is used for text preprocessing (tokenization, stop word removal, etc.). Then these pre-processed and cleaned answers are compared against stored answer keys and marks are awarded based on keyword comparison and count. At its backend, this application uses Firebase for data storage. Future scope includes evaluation of scientific diagrams or pictorial representations by comparison to some previously stored reference images.

Table 1 provides a summary of the research conducted in this field, offering a comparative analysis that highlights the limitations, algorithms used and future scope presented in the most recent papers.

Table 1

Table 1 Comparative Analysis of Previous Research

Author & Year	Proposed	Cons	Algorithms Used	Future Scope
(Nikhil, 2024)	Question Paper Checking Using Generative AI	The model may struggle with comprehensive answer assessment, have a weak grasp of contextual understanding and demonstrate biases present in the training data.	• Heuristic Rule Based Algorithm • Comparison Algorithm • Mathematical Model • Text to Text generative ai transformer.	Focus on establishing contextual understanding, reducing human bias, improving model and accuracy rates.

(Nayudu & Rani, n.d.)	Efficient Exam Grading with AI Powered Answer Verification	The model might encounter difficulties implementing machine learning using an efficient and well-structured method	• Keyword identification using the technique of Term- frequency inverse document frequency (TF-IDF) • Count vectorizer • Yet another keyword extractor • Summarization method • Cosine similarity • BM25L • Similarity Check • BERT Model	The model can be updated to evaluate the answers of students which contain multiple languages and enhancing more accuracy and error to obtain better result.
(Kulkarni et al., 2024)	Digital Handwritten Answer Sheet Evaluation	Illegible or unrecognizable handwriting could negatively impact the marks even if the answer is correct.	• Google Cloud vision • PDF2Image Python Library • OCR • BERT Tokenizer • Cosine Similarity • A rubric of four well defined evaluation Parameter • Word Embedding Character Segmentation	• Evaluation of exam sheets written in multiple languages • Evaluating tables and diagrams.
(Azubogu et al., 2024)	Development of Natural language Processing Based Descriptive Answer Evaluation Platform (Grade- scriptive)	Limited metrics for marks calculation	• NLP (Natural Language Processing) • MERN Stack • Xenova/all-MiniLM-L12-v2 (for language understanding). • WordNet	• Incorporation of Advanced NLP Techniques. • Broaden Evaluation Criteria. • Subject: Specific Customization.
(Vanathi, 2023)	Automated Exam Paper Evaluation System	Complex deep learning model for OCR task	• Image processing – Binarization, noise reduction, smoothing, etc. • RapidAPI OCR service • and combination of vgg19 and GRU neural network for OCR task • Fuzzy string matching • Weights assigned to keywords based on relevance and frequency • Levenshtein distance and cosine similarity. • Lemmatization	• Decreasing the time spent assessing the answers • Automatic formula evaluation Recognizing various forms and styles of handwritten text.
(S et al., n.d.) (2023)	Eval - Automatic Evaluation of Answer Scripts using Deep Learning and Natural Language Processing	The proposed model puts little emphasis on grammatical structure of answers while assigning marks	• POS tagging • Cloud Vision and textract OCR Models • BERT model to extract keywords • GPT-3 model to generate long answer summaries	• Taking spelling and grammatical errors into account

			• Similarity checking algorithm	
(Sanuvala & Fatima, 2021)	A Study of Automated Student's Examination Paper Using Machine Learning Techniques	Choosing the appropriate features of documents for assessing similarity. Determining suitable ML algorithms for (dis)similarity index computation.	• OCR tool • Cosine Similarity • Examination • Natural Language Processing Machine Learning Algorithms.	Minimizing the execution time by integrating hashing techniques into this system.
(Bashir et al., 2021)	Subjective Answer evaluation using ML and NLP	Current models may struggle with synonyms, sentence structure variations, and the overall context of answers, which can affect scoring accuracy	• Word2vec • Cosine similarity • WMD MNB	Particular domains training of word2vec

Automated grading systems are becoming more common and helpful, but they still have some big issues to solve. (Nikhil, 2024) came up with a generative AI model for grading, but it struggles with understanding context and has biases. (Nayudu & Rani, n.d.) used keyword-based techniques like TF-IDF and cosine similarity, but their system doesn't handle multiple languages well. (Kulkarni et al., 2024) used OCR and BERT but found that handwriting problems can mess up grading. (Azubogu et al., 2024) built an NLP-based grading platform but did not include enough ways to calculate scores fairly. (Saharan et al., n.d.) worked with neural networks and random forests but ran into issues with not having enough good datasets.

3. PROPOSED METHODOLOGY

3.1. TOOLS AND TECHNIQUES

Several pre-trained handwriting recognition models were thoroughly investigated to find the most effective solution with a good balance between accuracy, cost-effectiveness, and ability to handle diverse handwriting styles. Considering the results achieved, this paper suggests the application of an Optical Character recognition API from the RapidAPI platform.

For the task of actual answer evaluation, this paper uses the Llama 3.2 3B text-only model, part of the versatile Llama 3.2 family which includes both lightweight and high-capacity variants as well as vision LLMs and text-only models. With its 3 billion parameters and 128k support context length, Llama 3.2 3B offers substantial language understanding and generation capabilities without requiring extensive hardware resources. It is also designed to support fine-tuning, allowing for adaptation to specialized tasks while preserving its streamlined architecture for fast deployment. Fine-tuning is a transfer learning technique where a pre-trained deep learning model is further trained on a customized and/or task-specific dataset. This process updates the model's weight, allowing it to adapt to new tasks while retaining knowledge from its original training.

After fine tuning, Llama 3.2 3B will rephrase the answers of input into a list of points and create a new list of topic-wise short points and categorize them by priority in order of relevance to the question; applied only to reference answers. The model will Then identify overlapping points between reference and student answers by awarding marks according to the priority of common points. For each point, the analysis would be done based on semantic relevance to make sure diverse phrasing is used. Taken together, language structures and synonyms do not mask the actual similarity of contents.

The paper defines four weighted priority levels P1, P2, P3 and P4. P1 includes non-negotiable points essential to answering the question. P2 covers points that are not strictly required but add critical elements for a complete response. P3 consists of non-essential points that provide secondary context and enhance the answer's overall depth. P4 includes optional, extra and unnecessary points. While P1 priority level will always be assigned to one or more points, other levels may not be assigned to any reference answer points.

This domain-specific finetuning is achieved by training the model on our customized dataset using prompt-chaining technique, enhancing the model's ability to better understand the complex task of answer evaluation and provide tailored and accurate outputs. This dataset has been prepared by combining question-answer pairs from hundreds of

student exam sheets. It consists of these question-answer pairs along with meticulous preparation including cleaned, summarized points as well as priority assigned points for each answer. This will be used to effectively train the model about the input format as well as the intended output format. The model was fine-tuned using Low-Rank Adaptation (LoRA) with a rank of 16, scaling factor (α) of 16, and no dropout, applied to key transformer layers. The training process used the AdamW optimizer with 8-bit precision, a learning rate of 2e-4 and a batch size of 16. A linear learning rate scheduler was applied with 5 warmup steps, while mixed precision (fp16 or bf16) and a weight decay of 0.01 were used to improve training efficiency.

Instead of using a single very detailed prompt, a sequence of multiple simpler prompts is used to train the model, breaking down the entire evaluation process into smaller, manageable steps. Four different prompts are defined where the output of one prompt becomes the input of the next. The first prompt is defined to rephrase the input answers into a more structured and consistent format. The second prompt contains instructions to create a new list of topic-based condensed points from the list of points obtained from the first prompt. The third prompt has instructions to assign priorities to the topic-based points from the second prompt and the fourth prompt defines the steps to identify common points between the reference and student answer.

For the process of diagram evaluation, the paper proposes the use of the state-of-art LightGlue (Lindenberger et al., 2023) deep learning framework. Its core transformer architecture integrates attention mechanism with traditional computer vision techniques to create a powerful tool for feature extraction, image matching and object detection. It can build on top of local feature detectors and descriptors (such as SIFT, SuperPoint, or DISK) to establish correspondences between key points in two diagrams or images. It also achieves high repeatability due to its iterative refinement process. The paper proposes leveraging this framework to extract and match key points between reference and hand-drawn diagrams, enabling precise calculation of a similarity score (S). The accuracy of this score is further improved through a custom image pre-processing pipeline applied to the images. Next, OCR is applied to extract labels from the diagram, and a similarity score (L) is calculated by comparing them with the expected diagram labels. On the basis of all the above defined parameters, marks for a particular answer can be calculated using the following equations 1 and 2:

$$M = P_1N_1 + P_2N_2 + P_3N_3 + P_4N_4 + W_SS + W_LL \quad (1)$$

$$M = \sum_{i=1}^{4} P_iN_i + W_SS + W_LL \quad (2)$$

Where,

M = Marks for a given answer

P_1 = Weighted parameter of first priority level

P_2 = Weighted parameter of second priority level

P_3 = Weighted parameter of third priority level

P_4 = Weighted parameter of fourth priority level

N_1 = Number of overlapping points in first priority level

N_2 = Number of overlapping points in second priority level

N_3 = Number of overlapping points in third priority level

N_4 = Number of overlapping points in fourth priority level

W_S = Weighted parameter of diagram feature similarity

W_S = 0, if no diagram is present in the question

W_L = Weighted parameter of diagram labels similarity

W_L = 0, if no diagram is present in the question

S = Similarity score of features in student diagram and expected diagram

L = Similarity score of labels in student diagram and expected diagram.

3.2. AUTOMATED GRADING SYSTEM ALGORITHM

1) Data collection

$D \leftarrow \{S_1, S_2, ..., S_n\}$

(Collect student scripts S_i)

2) Data preparation

For each $S_i \in D_i$:

- Extract: QA_i = Extract (S_i)
- Clean: QA_i' = Clean (QA_i)
- Summarize: P_i = Summarize (QA_i')
- Prioritize: W_i = Assign Weights (P_i)

3) Model training

LLM = FineTune ($Llama_{3.2}$, $\{W_i\}$)

4) Integration of diagram evaluation functionality

For each diagram Img_i:

Img_i' = Preprocess(Img_i)

F_i = LightGlue(Img_i')

L_i = OCR(Img_i')

5) OCR for text

T_i = OCR(S_i^{text})

6) Answer evaluation

Score:

$$Score_i = \sum(\alpha_i . (T_i)) + \beta . DiagramFeatures(F_i) + \gamma . LabelAccuracy(L_i)$$

7) Maintenance and Enhancement

Update Dataset $D' \leftarrow D + \{new\ types\}$

Re-train: $LLM' \leftarrow FineTune(LLM, D')$

Where S_i : The i-th student script (exam sheet or assignment).

QA_i : Extracted Question-Answer pairs from student script S_i.

QA_i' : Cleaned version of QA_i – unnecessary characters removed, formatting corrected.

P_i : Summarized points (heading-based key ideas) from the cleaned Q&A.

W_i : Weighted or prioritized points – each point is given a priority score for model training.

LLM: The finetuned Llama 3.2 model.

Img_i : The i-th student diagram(image) submitted as part of an answer.

Img_i' : Preprocessed version of the image – resized, denoised, normalized, etc.

F_i : Extracted features from the diagram using LightGlue (e.g., edges, structures).

L_i : Labels or annotations extracted from the diagram using OCR.

T_i : Text content of the handwritten answer extracted using OCR.

$Score_i$:Final evaluation score for student$_i$'s answer.

α, β, γ: Weights for each scoring component:

where α_i: weighted components for content relevance

β: weight for diagram features

γ: weight for label accuracy

3.3. METHODOLOGY

Step 1: Scanning and digitization of answer sheet

The student answer sheets are scanned to transform them into digitized documents i.e. PDFs. The pages of these answer sheet PDFs are then parsed and converted into images.

Step 2: Optical Character Recognition

The next step involves conversion of handwritten text from student answer sheets into processable text as shown in Figure 1. The newly generated images are opened in binary format and sent as a POST request to RapidAPI's OCR API for text extraction.

Figure 1

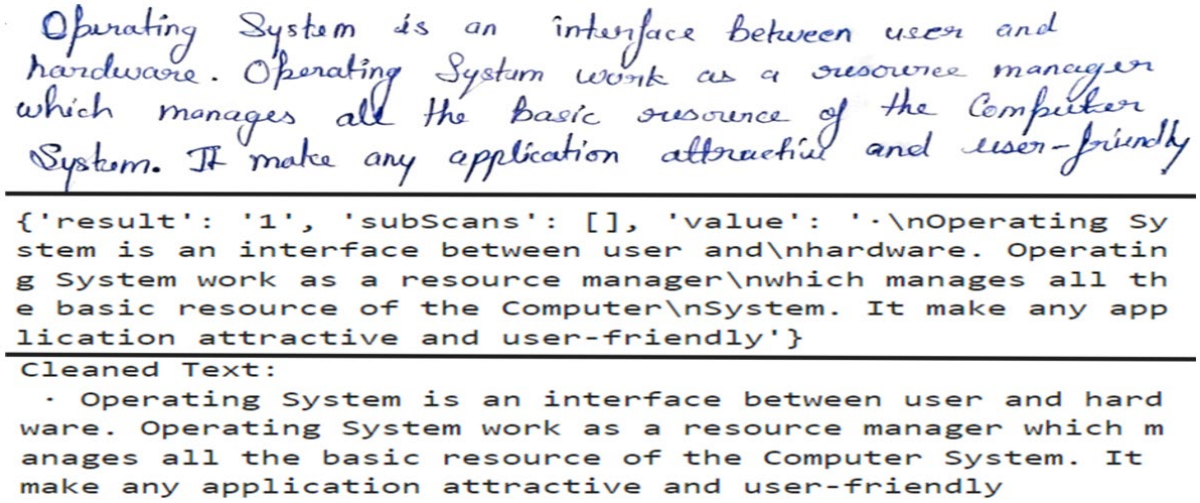

Figure 1 Optical Character Recognition

Step 3: Text cleaning and pre-processing

The extracted answers are cleaned to enhance content readability by replacing multiple spaces, newline characters and handling punctuation. These answers then undergo further pre-processing by finetuned Llama 3.2 to ensure consistent format where each answer is broken down into a uniform list of points.

Step 4: Creation of topic-based concise points

For a given student answer, its list of points is analyzed by Llama 3.2 to generate a new structured list of concise, topic-based points that correspond to the question's requirements. The objective is to identify key information and present it in a precise question-aligned format to facilitate a clear and reliable evaluation.

Step 5: Student answer points vs expected answer points

This step involves fine-tuned Llama 3.2 to identify and analyze which points overlap between the student answer and its corresponding reference answer. A precise calculation of marks would be provided based on the priority assigned for common points.

Step 6: Diagram Detection

If a diagram is expected in the answer, the relevant answer images are loaded and checked. If no diagram is found, marks are deducted; otherwise, the detected diagram is cropped and saved for further analysis. Figure 2 shows an example of diagram detection functionality.

Figure 2

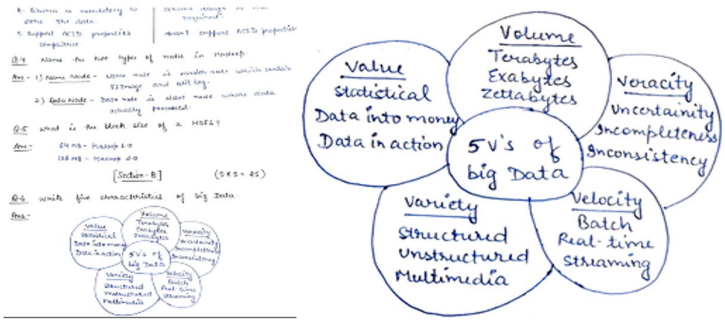

Figure 2 Diagram Detection

Step 7: Pre-processing of diagram image

Next, the diagram image undergoes a customized pre-processing pipeline involving resizing, noise reduction, edge enhancement and binarization. Contours are identified and overlaid on the original image to aid shape analysis and feature extraction. Figure 3 illustrates the sequential output of this pipeline.

Figure 3

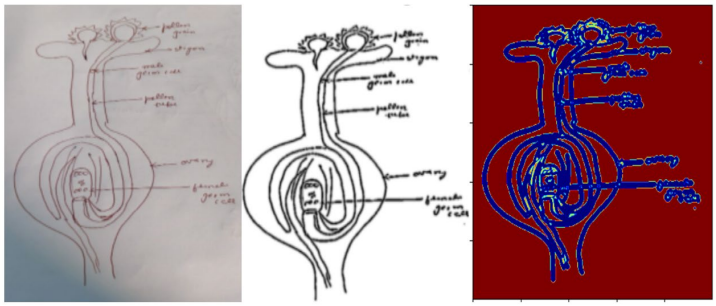

Figure 3 Image pre-processing

Step 8: Feature extraction and matching with expected diagram

The processed student diagram is compared with the reference image, also pre-processed using the same pipeline with the Light Glue framework. Light Glue detects and matches key points in both images, aligning corresponding regions as shown in Figure4. The similarity score (S) is calculated based on the number of matched key points and their spatial relationship, quantifying their resemblance.

Figure 4

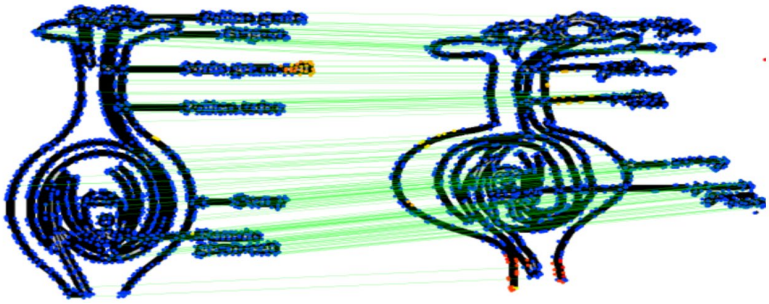

Figure 4 Comparison of student drawn and hand-drawn diagram

Step 9: Extraction and evaluation of diagram labels

Next, label matching verifies the accuracy of words and phrases in the diagram. The student diagram's labels, extracted via OCR as a single string, are compared to the reference diagram's labels, also formatted as a single string. Similarity analysis is performed using the BLEU score metric, focusing on word matches and phrase accuracy.

Step 10: Marks calculation for a single answer

The marks for a single answer will be calculated on the basis of all the above defined parameters, i.e. the priority of points discussed in both the student answer and reference answer, the similarity score of diagram matching if drawn along with the similarity score of its extracted set of labels.

Step 11: Repetition of above steps for each answer in exam sheet and output final score

This process will be repeated for each and every question in the answer sheet. At the end, the marks will be totaled and the final score for the entire answer sheet will be displayed. Figure 5 visualizes the complete methodology in the form of a flowchart and figure 6 is a snapshot of the final system formed in process.

Figure 5

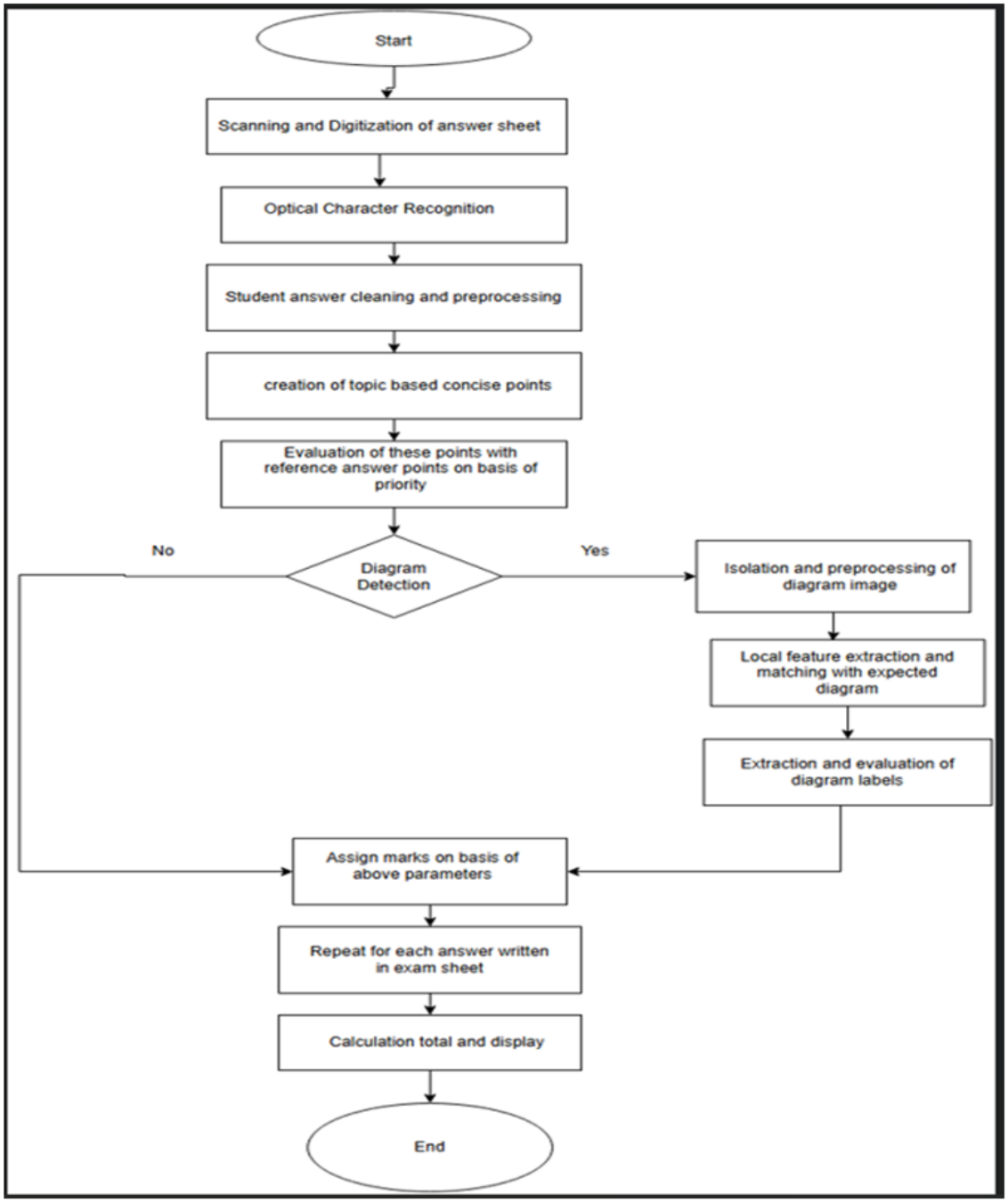

Figure 5 Flowchart for methodology

Figure 6

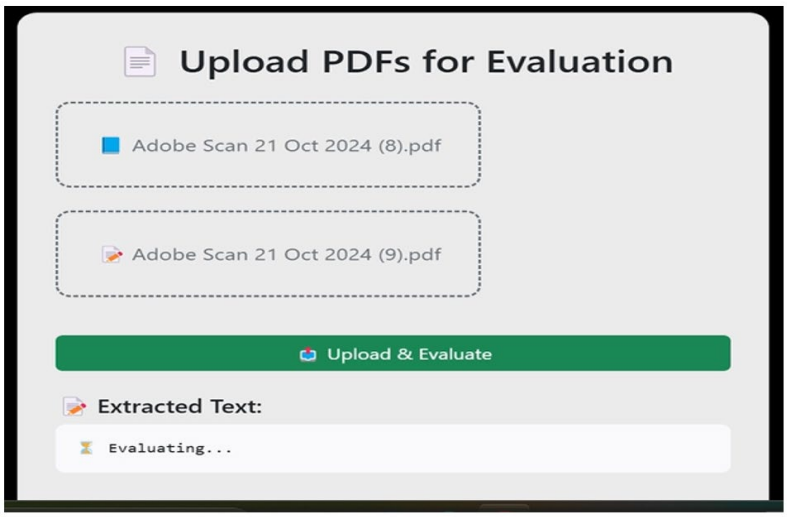

Figure 6 Snapshot of the final system in process

4. IMPLEMENTATION AND RESULTS

The accuracy of Optical Character Recognition (OCR) API was checked on various images and handwriting styles. It was found that the performance of handwritten OCR is significantly influenced by the quality of input image and the clarity of handwriting. Clean images with good or average handwriting provide clear and well-formed output with an accuracy of 80-85%. However, the same is not true for noisy, low-quality images and images with unclear handwriting including irregular character shapes, smudges or distortion with an accuracy of 50-55%. While the blurry and low-quality image output can be improved with certain pre-processing techniques, images with unclear handwriting pose a significant challenge and can affect the overall accuracy of the proposed answer evaluation system.

While training the Llama 3.2 model for the task of answer evaluation, we experimented with various prompt-tuning strategies to assess model performance and achieve highest degree of accuracy possible. The model was initially trained to use a single, very complex prompt to generate the required output which did not perform well. However, after incorporating prompt chaining where the overall process of answer evaluation was broken down into simple steps using multiple, sequential and smaller prompts, the model's understanding of the task improved significantly with its outputs now more aligned with the expected output.

Next the model training was conducted with task-based chain of prompts and instruction-based chain of prompts. Experimental results showcased how instruction-based prompt approach consistently outperformed the task-based prompt approach. Considering the results obtained, the paper employs the prompt chaining approach with instruction-based prompts due to its superior performance and ability to better guide the model's behavior and understanding of context through process.

While fine tuning the model with different prompt tuning approaches and on various types of answers, several other inferences were drawn. The answers written by students in an exam can be in varied formats such as paragraph form, simple list of points, heading-wise point formats, table-based formats and comparison-based questions. The best results have been achieved for heading-based point format followed closely by list of points formats and paragraph-based answers. Next are the comparison-based questions which are typically presented in a multi-column format. With such a layout, the Optical Character Recognition API may fail to maintain the intended structure and struggle with accurate text extraction. These questions construct a hit or miss situation where the answer may or may not be extracted correctly, thus affecting such questions' overall evaluation accuracy. The OCR output for a comparison-based question is shown in figure 7.

Figure 7

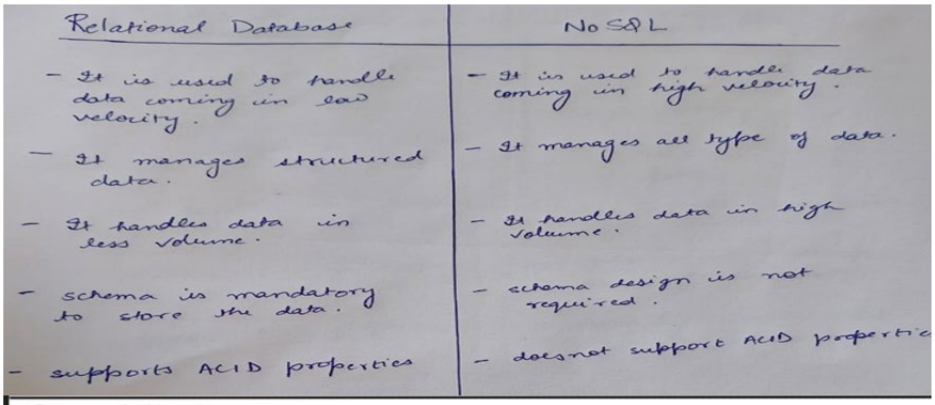

Figure 7 OCR output obtained from a comparison-based question

The worst results were achieved with table-based questions as shown in figure 8 where the OCR API was completely unsuccessful in extracting the answer due to the complex structure of rows and columns. Without clear dividers or consistent alignment, the content extracted from various cells got mixed up making such questions' evaluation not possible with the current system.

Figure 8

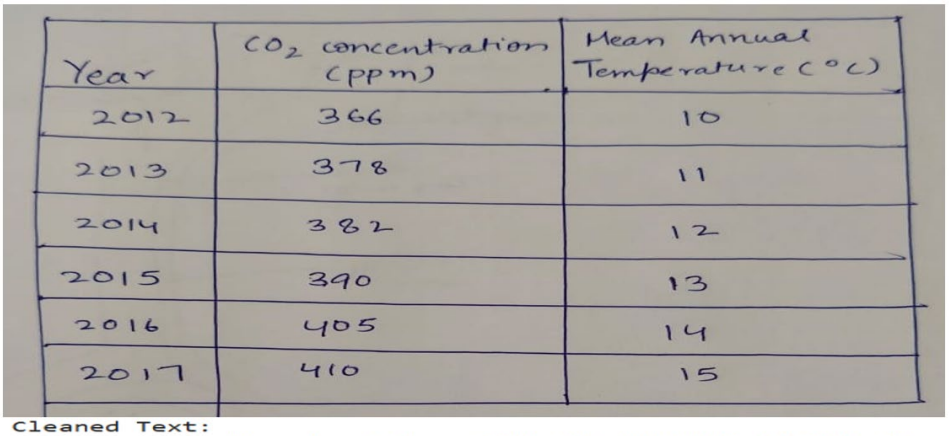

Figure 8 Inaccurate OCR output obtained from a table-based question

Other than that, to assess the effectiveness of our approach, we compared the model's grading decisions with those of human evaluators. This comparison highlighted the model's alignment with human judgment, ensuring reliability in assessing answer quality, addressing variations in expression & structure and diagram checking with LightGlue achieving around 95% correct matches and maintaining 70-80% accuracy even in challenging scenarios. Our model was able to significantly reduce bias and complete the grading process in a fraction of the time taken by human evaluators, proving to be much more scalable and efficient. Figure 9 shows the results of this comparison between human evaluators and our proposed system on basis of various factors through separate graphs and Figure 10 summarizes them into a single graph for overall analysis.

Figure 9

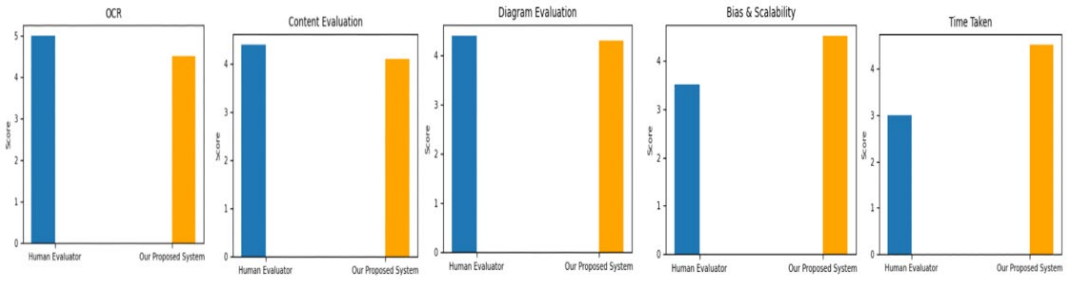

Figure 9 Comparison of human evaluator and Our Proposed System

Figure 10

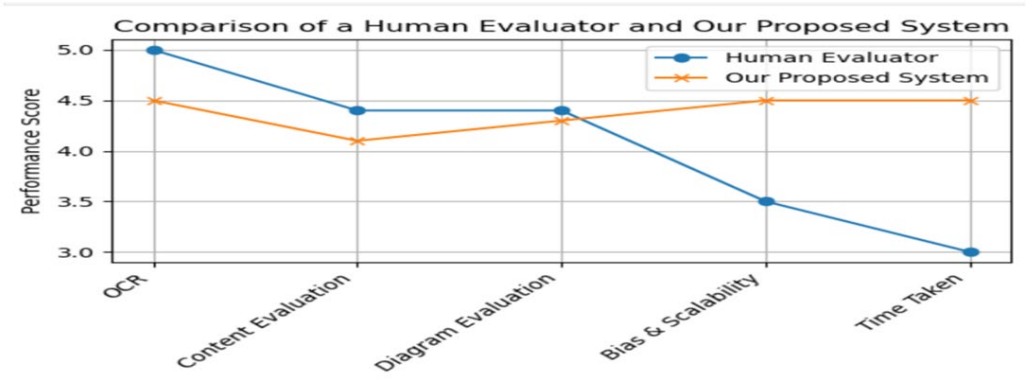

Figure 10 Comparison of human evaluator and Our Proposed System

5. COMPARATIVE ANALYSIS WITH EARLIER RESEARCH

Our system builds on a foundation of prior studies and addresses key limitations while enhancing overall functionality and accuracy. A detailed comparison with earlier research in this field highlights our advancements in key areas such as diagram analysis, which makes our system multimodal, as well as contextual evaluation and semantic understanding of text, particularly in handling varied grammatical structures and vocabulary. Table 2 summarizes the proposed system compared to existing approaches, showcasing its broader applicability and effectiveness.

Table 2

Table 2 Table of Comparative analysis with earlier research

Author & Year	Technologies Used	Features	Results Obtained	Future Scope
(Nikhil, 2024)	Text to Text generative ai transformer, Heuristic Rule Based Algorithm, Comparison Algorithm	User-friendly interface, Immediate feedback	Reduced time and effort required for grading subjective exams	Reduction of bias from training data
(Kulkarni et al., 2024)	Google Cloud vision, Tokenization and Word Embedding, BERT Tokenizer, Cosine Similarity	Consistent evaluation, Feedback generation for students	Improves grading efficiency and fairness, Provides timely feedback to students	Enhancement of the system to evaluate diagrams
(Azubogu et al., 2024)	NLP, MERN Stack, Xenova/all-MiniLM-L12-v2, WordNet	Good approach for semantic and linguistic understanding of answers	Reduced bias, faster turnaround time as compared to human evaluation	Incorporation of detailed evaluation criteria
Our Proposed Solution	OCR API from RapidAPI, Llama 3.2 (3B) LLM, LightGlue	Multimodal approach with diagram evaluation, Increased parameters for evaluation,	Effectively accounts for variations in language structure, tone and vocabulary enabling a deeper understanding of the content's semantics, Ensures	Multilingual Support

	Consistent grading criteria, Minimal bias while checking	consistency in evaluation of different types of answers	
			Evaluation of mathematical problems and table-based answers

6. CONCLUSION

The paper presents a new approach to implement an automated evaluation system for the grading of answer sheets. The proposed system builds on the research previously conducted within this field, introducing new techniques to overcome prior shortcomings. This includes evaluating answers based on context, underlying concepts, and nuanced text, rather than highly relying on keyword extraction techniques. It caters to evaluation of both long, subjective questions and short answer questions. The overall system integrates a new diagram analyzer functionality that enables the assessment of both textual and visual elements of an answer, thus further improving the accuracy and reliability of the system. This system combines OCR, LLM finetuning, and image feature extraction and matching to implement a system that will boost the current scenario of automated grading systems. However, the proposed system also has certain limitations: reduced evaluation accuracy in the case of comparison-based answers, table-based answers, and answer sheets with unclear handwriting.

INFORMED CONSENT

All interviews and image data collection were conducted with the prior informed consent of the participants.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Al Mahmud, T., Hussain, M. G., Kabir, S., Ahmad, H., & Sobhan, M. (2020). A Keyword Based Technique to Evaluate Broad Question Answer Script. Proceedings of the 2020 9th International Conference on Software and Computer Applications, 167–171. https://doi.org/10.1145/3384544.3384604

Azubogu, K., Asogwa, E. C., Ezeugbor, I. C., Okwuchukwu Ejike, C., & Onyeizu, M. N. (2024). Development of Natural Language Processing-Based Descriptive Answer Evaluation Platform (Gradescriptive). Engineering and Technology Journal, 09(08). https://doi.org/10.47191/etj/v9i08.47

Balat, H. F., El-dosuky, M. A., El-Razek, E.-S. M. A., & Rashed, M. Z. (2020). Automatic Exam Evaluation based on Brain Computer Interface. International Journal of Computer Applications, 175(25), 15–21. https://doi.org/10.5120/ijca2020920792

Bashir, M. F., Arshad, H., Javed, A. R., Kryvinska, N., & Band, S. S. (2021). Subjective Answers Evaluation Using Machine Learning and Natural Language Processing. IEEE Access, 9, 158972–158983. https://doi.org/10.1109/ACCESS.2021.3130902

Bhonsle, V., Sapkal, P., Mukadam, D., & Raut, V. (2019). An Adaptive Approach for Subjective Answer Evaluation. International Journal for Research and Innovation, 1(2). www.viva-technology.org/New/IJRI

Brown, M., & Program, E. (2017). Automated Grading of Handwritten Numerical Answers.

Chandrapati, L. M., & Rao, Ch. K. (2024). Descriptive Answers Evaluation Using Natural Language Processing Approaches. IEEE Access, 12, 87333–87347. https://doi.org/10.1109/ACCESS.2024.3417706

G, J., & G, C. S. (2020). Online Subjective answer verifying system Using Artificial Intelligence. 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 1023–1027. https://doi.org/10.1109/I-SMAC49090.2020.9243601

Kulkarni, M., Adhav, G., Wadile, K., Chavan, R., & Deshmukh, V. (2024). Digital Handwritten Answer Sheet Evaluation System. https://doi.org/10.21203/rs.3.rs-3978232/v1

Lindenberger, P., Sarlin, P.-E., & Pollefeys, M. (2023). LightGlue: Local Feature Matching at Light Speed. http://arxiv.org/abs/2306.13643

Meenakshi, A. T., Pradeep, B. M., & Vishaka, M. (2022). Web app for quick evaluation of subjective answers using natural language processing. Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 22(3), 594–599. https://doi.org/10.17586/2226-1494-2022-22-3-594-599

Nayudu, P. P., & Rani, I. R. (n.d.). Efficient Online Exam Grading with AI Powered Answer Verification. Retrieved www.joics.org

Nikhil, D. (2024). Question Paper Checking Using Generative Ai. International Journal for Research in Applied Science and Engineering Technology, 12(5), 4362–4368. https://doi.org/10.22214/ijraset.2024.62488

Rahman, M. M., & Akter, F. (n.d.). An Automated Approach for Answer Script Evaluation Using Natural Language Processing. Retrieved www.ijcset.net

Raina, S., Amin, H., Sanghvi, S., Bharti, S. K., & Gupta, R. K. (2023). Automatic Subjective Answer Evaluator Using BERT Model (pp. 531–538). https://doi.org/10.1007/978-981-99-3315-0_40

Rowtula, V., Oota, S. R., & C.V, J. (2019). Towards Automated Evaluation of Handwritten Assessments. 2019 International Conference on Document Analysis and Recognition (ICDAR), 426–433. https://doi.org/10.1109/ICDAR.2019.00075

S, P. M., Chavan, S. M., Bathula, R., Saikumar, S., & Dayalan, G. (n.d.). International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Eval-Automatic Evaluation of Answer Scripts using Deep Learning and Natural Language Processing. In Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE (Vol. 2023, Number 1). Retrieved www.ijisae.org

Saharan, R., Chauhan, R. K., Singh, S., & Sharma, P. (n.d.). AUTOMATED CONTENT GRADING USING MACHINE LEARNING Theoretical Content Grading from Exam Papers.

Salim, H. R., De, C., Pratamaputra, N. D., & Suhartono, D. (2022). Indonesian automatic short answer grading system. Bulletin of Electrical Engineering and Informatics, 11(3), 1586–1603. https://doi.org/10.11591/eei.v11i3.3531

Sanuvala, G., & Fatima, S. S. (2021). A Study of Automated Evaluation of Student's Examination Paper using Machine Learning Techniques. 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 1049–1054. https://doi.org/10.1109/ICCCIS51004.2021.9397227

Sridevi, V., Kumar S., S., Supraja, B., & Udhayakumar, S. (2019). Knowledge Representation and Answer Evaluation System using Language Processing Algorithm. 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), 1–4. https://doi.org/10.1109/ViTECoN.2019.8899525

Säuberli, A., & Clematide, S. (2024). Automatic Generation and Evaluation of Reading Comprehension Test Items with Large Language Models. http://arxiv.org/abs/2404.07720

Tan, L. Y., Hu, S., Yeo, D. J., & Cheong, K. H. (2025). A Comprehensive Review on Automated Grading Systems in STEM Using AI Techniques. Mathematics, 13(17), 2828. https://doi.org/10.3390/math13172828

Vanathi, B. (2023). Automated Exam Paper Evaluation System. In International Journal of Current Science (Vol. 13, Number 2). www.ijcspub.org