

Automated Handwritten Answer Script Assessment Using Deep Learning: A Synoptic Evaluation Framework

Shubham Rajapurkar

*Department of Computer Science and Engineering
Sardar Patel Institute of Technology
Mumbai, Maharashtra, India
shubham.rajpapurkar22@spit.ac.in*

Aashay Jadhav

*Department of Computer Science and Engineering Sardar
Patel Institute of Technology
Mumbai, Maharashtra, India
aashay.jadhav22@spit.ac.in*

Dhruv Kanadia

*Department of Computer Science and Engineering Sardar Patel Institute of Technology
Mumbai, Maharashtra, India
dhruv.kanadia22@spit.ac.in*

Abstract—Evaluating handwritten examination scripts presents significant challenges for educational institutions, requiring substantial time investment from educators while remaining vulnerable to inconsistencies from subjective interpretation. This work introduces a comprehensive automated framework for assessing handwritten responses using a synoptic methodology that emphasizes conceptual comprehension and semantic alignment rather than exact textual correspondence. The system combines advanced optical character recognition technology, vision-language models, and natural language processing architectures to digitize student answers and perform intelligent evaluation against reference solutions through semantic analysis. The synoptic paradigm accommodates varied expression styles while verifying coverage of essential concepts. Experimental validation using authentic academic datasets demonstrates exceptional performance, achieving 86.79% overall accuracy, 98% semantic precision, and 92.4% concordance with human evaluators. The modular design assists educators in streamlining grading workflows, improving assessment consistency, and enabling scalable evaluation across diverse academic disciplines. An intuitive interface facilitates seamless processing of scanned scripts and provides detailed visualization of results with granular feedback, establishing this system as an impactful solution for modernizing educational assessment practices.

Index Terms—Automated assessment, synoptic evaluation, optical character recognition, natural language processing, large language models, vision-language models, educational technology, Mistral OCR, Qwen2.5-VL, semantic similarity

I. INTRODUCTION

A. Context and Rationale

Manual assessment of handwritten examination responses constitutes a fundamental yet demanding operation within educational systems globally. This conventional paradigm necessitates extensive time allocation from instructors, frequently consuming multiple hours per script based on complexity

and response length. Beyond temporal demands, the process suffers from systematic challenges that affect assessment reliability and fairness.

Despite expertise and commitment, human evaluators inevitably introduce variability into scoring outcomes. Elements including cognitive exhaustion, implicit biases, temporal drift in standards, and subjective interpretation of ambiguous content contribute to assessment inconsistency. Educational research has documented inter-rater reliability measures that often fail to meet acceptable standards, particularly for open-ended questions requiring nuanced judgment [1]. This variability affects individual student outcomes and undermines equitable assessment principles fundamental to educational integrity.

As educational institutions experience substantial enrollment expansion and heightened assessment frequency driven by continuous evaluation frameworks, manual grading scalability limitations become increasingly apparent. Educators face mounting pressure to deliver timely feedback while maintaining instructional quality and pursuing pedagogical innovation. This tension between assessment responsibilities and academic pursuits creates bottlenecks that delay feedback provision, potentially affecting student learning during critical knowledge consolidation periods.

Artificial intelligence and machine learning advancements have created opportunities for addressing these challenges through intelligent automation. However, handwritten answer evaluation presents unique technical complexities distinguishing it from general document processing. Handwritten text exhibits substantial variability across individuals, with writing styles, character formations, spatial arrangements, and linguistic patterns creating heterogeneous data that challenges conventional OCR systems. Academic responses frequently in-

corporate specialized terminology, mathematical expressions, scientific diagrams, and multi-modal representations requiring sophisticated interpretation capabilities.

B. Research Goals and Novel Contributions

This research addresses the necessity for an automated evaluation framework capable of processing handwritten scripts with accuracy comparable to human evaluators while maintaining fairness, transparency, and pedagogical validity. The primary objective involves developing a comprehensive computational system seamlessly integrating state-of-the-art technologies for end-to-end automated assessment of diverse response types.

The system introduces several innovative contributions to educational technology and automated assessment:

Synoptic Evaluation Paradigm: Unlike traditional systems relying predominantly on keyword matching or surface-level similarity metrics, this approach implements a synoptic methodology assessing answers based on conceptual understanding and semantic coherence. This paradigm recognizes that students may articulate correct concepts using varied linguistic expressions, prioritizing essential ideas over specific phraseology. The synoptic approach mirrors expert human evaluation strategies and provides equitable assessment outcomes for students with diverse communication styles.

Multi-Modal Processing Architecture: The system incorporates specialized pipelines for textual and visual content, recognizing that academic assessments frequently require diagrams, flowcharts, mathematical derivations, and graphical representations. Employing complementary technologies for text extraction (Mistral OCR) and diagram interpretation (Qwen2.5-VL-32B-Instruct), the system achieves comprehensive coverage of assessment content types across disciplinary boundaries.

Enhanced Semantic Analysis: Integration of the Mistral 7B language model enables sophisticated semantic comparison between student responses and reference answers. This approach leverages contextual embeddings and transformer architectures to capture nuanced meaning relationships transcending lexical overlap. The semantic component recognizes paraphrased content, identifies conceptually equivalent expressions, and detects partial understanding scenarios requiring proportional credit allocation.

Real-World Validation: The system underwent rigorous evaluation using authentic examination scripts from Sardar Patel Institute of Technology (SPIT), encompassing genuine student responses across multiple subjects and question types. This validation ensures reliable performance under actual deployment conditions rather than idealized scenarios, providing credible evidence of practical applicability.

Multi-Dimensional Assessment Framework: Beyond simple accuracy measurements, the evaluation incorporates multiple quality dimensions including semantic accuracy, factual correctness, understanding depth, key points coverage, and agreement with human evaluators. This approach provides

deeper insights into system capabilities across different evaluation aspects.

C. Document Structure

The remainder of this paper provides comprehensive coverage of theoretical foundations, technical implementation, experimental validation, and practical implications. Section II presents an extensive review of related work in optical character recognition, automated assessment methodologies, and natural language processing in educational contexts. Section III details system architecture, algorithmic approaches, and technical design decisions. Section IV describes experimental setup, dataset characteristics, evaluation protocols, and implementation details. Section V presents comprehensive results with detailed analysis and statistical validation. Finally, Section VI synthesizes key findings, discusses practical implications, acknowledges limitations, and outlines future research directions.

II. LITERATURE REVIEW

A. Optical Character Recognition Advances

Optical character recognition technology has experienced remarkable evolution over the past decade, transitioning from rule-based systems with limited applicability to deep learning solutions capable of handling diverse document types accurately. Early OCR systems relied heavily on template matching and feature extraction techniques performing adequately for printed text in controlled conditions but struggling with hand-written content, complex layouts, and degraded documents.

Convolutional neural networks marked a paradigm shift in OCR capabilities, enabling systems to learn hierarchical feature representations directly from pixel data without manual engineering. Subsequent developments in recurrent neural networks and attention mechanisms further enhanced sequence modeling, allowing OCR systems to capture long-range dependencies and contextual relationships within text sequences.

Recent transformer-based architectures have revolutionized document understanding by enabling parallel processing of entire sequences and capturing global contextual information more effectively. Vision Transformers have demonstrated exceptional performance on computer vision tasks by treating image patches as sequence elements and applying self-attention mechanisms to model spatial relationships.

The system leverages Mistral OCR, a cutting-edge cloud-based solution embodying these architectural advances and specifically targeting educational document processing challenges [4]. Mistral OCR distinguishes itself through capabilities making it particularly suitable for academic assessment contexts. The system employs deep learning models trained on extensive datasets encompassing diverse handwriting styles, enabling robust recognition across heterogeneous writing patterns typical in student populations. Advanced layout analysis algorithms maintain spatial relationships and structural information from source documents, crucial for correctly interpret-

ing examination scripts containing annotations, margin notes, and non-linear arrangements.

The platform's multilingual support ensures applicability across diverse educational contexts, while high processing throughput makes it viable for batch processing large examination volumes. Most importantly for educational applications, Mistral OCR's contextual recognition capabilities leverage surrounding text to disambiguate challenging character sequences, reducing error rates where individual character recognition proves difficult.

B. Vision-Language Models for Visual Content

While text extraction represents a crucial component, many assessments require students to create visual representations including labeled diagrams, flowcharts, circuit schematics, and graphical illustrations. Evaluating such content presents distinct challenges demanding specialized approaches capable of interpreting spatial relationships, recognizing geometric primitives, and understanding semantic relationships between visual elements and textual labels.

Vision-language modeling has emerged as a promising approach for bridging visual perception and linguistic understanding. These models jointly process visual and textual information, enabling them to ground language in visual contexts and generate linguistic descriptions of visual content. Recent advances have produced increasingly sophisticated systems capable of fine-grained visual reasoning and detailed scene understanding.

Qwen2.5-VL-32B-Instruct represents a state-of-the-art vision-language model designed for tasks requiring precise visual parsing and structured analysis [5]. The architecture incorporates innovations enhancing suitability for diagram evaluation in educational contexts. A redesigned Vision Transformer with window attention mechanisms enables efficient processing of high-resolution images while maintaining computational tractability, ensuring fine details in student diagrams remain perceptible.

Enhanced positional encoding schemes preserve spatial fidelity across different image resolutions, enabling accurate representation of geometric relationships between diagram components. This capability proves essential for evaluating diagrams where relative positions of elements convey semantic meaning, such as circuit diagrams where component connectivity determines functional behavior [5].

The model's training incorporated extensive pretraining on 4.1 trillion tokens, encompassing diverse document types including handwritten content, tables, charts, and technical diagrams. This broad exposure enables robust generalization to varied diagram types encountered in academic assessments. Instruction tuning further aligns outputs with educational evaluation requirements, enabling structured feedback and criterion-specific assessments integrating seamlessly with synoptic evaluation workflows [5].

Dynamic-resolution processing capabilities allow the model to adapt internal representations based on input characteristics,

ensuring small annotations and thin strokes common in hand-drawn diagrams remain detectable. This adaptive approach proves particularly valuable given variability in diagram quality resulting from writing implement characteristics, paper texture, and scanning resolution.

C. Automated Subjective Answer Evaluation

Automated evaluation of subjective answers has attracted sustained research attention, with early approaches predominantly relying on rule-based systems and keyword matching. Praveen proposed a methodology for evaluating subjective questions in online examination systems through keyword presence analysis, demonstrating feasibility of automated assessment for structured responses with predictable content patterns [1]. However, such approaches exhibit limited flexibility when confronted with diverse expression styles characteristic of authentic student responses, often penalizing valid answers employing alternative terminology.

Tulaskar et al. advanced the field by introducing token-based similarity measures for comparing student responses against model answers [2]. Their system computed various similarity metrics including Jaccard similarity, cosine similarity, and Levenshtein distance to quantify response quality. While these methods provided more nuanced assessment than binary keyword matching, they remained fundamentally limited by reliance on surface-level textual features rather than semantic understanding. Consequently, such approaches struggle to recognize conceptually equivalent responses expressed through different linguistic formulations.

Limitations of early approaches motivated researchers to explore sophisticated natural language processing techniques capable of capturing semantic relationships between text segments. Development of word embeddings, particularly dense vector representations like Word2Vec and GloVe, enabled systems to identify semantic similarities based on distributional properties in large text corpora. However, these word-level embeddings lack contextual sensitivity required to disambiguate polysemous terms or capture meaning variations induced by syntactic structure.

D. Large Language Models and Semantic Understanding

Recent emergence of large language models based on transformer architectures has fundamentally transformed natural language understanding capabilities across diverse applications. These models leverage self-attention mechanisms to capture long-range dependencies and contextual relationships within text sequences, enabling sophisticated semantic analysis approaching human-level comprehension in many tasks.

The Mistral model family represents a significant development in democratizing access to powerful language models through open-weight architectures balancing performance with computational efficiency [4]. Jiang et al. introduced Mistral 7B as a lightweight yet highly capable language model designed to perform complex NLP tasks with reduced computational requirements compared to larger contemporaneous models.

The architecture emphasizes memory efficiency through techniques including grouped-query attention and sliding window attention, making it suitable for deployment in resource-constrained educational environments where extensive computational infrastructure may not be available.

For automated answer evaluation, Mistral's semantic understanding capabilities prove particularly valuable. The model assesses conceptual alignment between student responses and reference answers by encoding both texts into dense contextual representations and computing similarity measures in the resulting embedding space. This approach enables recognition of paraphrased content, identification of equivalent expressions using different terminology, and detection of partial understanding scenarios where students correctly articulate some but not all key concepts.

Furthermore, Mistral's training on diverse text sources including academic content, technical documentation, and educational materials enhances familiarity with linguistic patterns and terminology conventions prevalent in scholarly writing. This domain exposure improves the model's ability to accurately interpret specialized vocabulary and disciplinary conventions appearing in student responses across various subjects.

Liu et al. demonstrated the importance of visual instruction tuning for enhancing multimodal understanding in language models [3]. Their work established that jointly training models on aligned visual and textual data significantly improves performance on tasks requiring coordinated processing of both modalities. This insight directly informed the decision to incorporate Qwen2.5-VL-32B-Instruct for diagram evaluation, recognizing that effective assessment of visual content requires models explicitly designed to understand semantic relationships between visual elements and textual descriptions.

E. Research Gaps and Motivation

Despite significant progress in OCR technology, vision-language modeling, and semantic text analysis, integration of these capabilities into comprehensive systems for automated answer script evaluation remains largely unexplored in academic literature. Most existing research focuses on individual components rather than end-to-end solutions addressing the full spectrum of challenges encountered in real-world educational assessment scenarios.

Furthermore, previous work has predominantly evaluated systems on simplified datasets that may not adequately represent complexity and variability of authentic student responses. Handwritten answers produced under examination conditions exhibit characteristics including imperfect handwriting clarity, time-constrained expression, and varied levels of conceptual understanding that may not be well-represented in artificially constructed datasets.

The synoptic evaluation approach proposed in this research specifically addresses these gaps by implementing a holistic assessment methodology evaluating answers based on conceptual understanding rather than surface-level similarity. This paradigm more closely aligns with expert human evaluation

strategies and provides fairer outcomes for students with diverse communication styles. By combining state-of-the-art technologies for text extraction, diagram interpretation, and semantic analysis within a unified framework, and validating the system on authentic examination scripts from a real educational institution, this research advances both theoretical foundations and practical applicability of automated assessment systems.

III. METHODOLOGY

A. System Architecture Overview

The proposed automated evaluation system employs a modular architecture decomposing the complex task of answer script assessment into specialized processing stages, each optimized for specific subtasks while maintaining seamless integration across the complete pipeline. The high-level system architecture depicts information flow from raw scanned answer scripts through various processing stages to final evaluation outputs with detailed feedback.

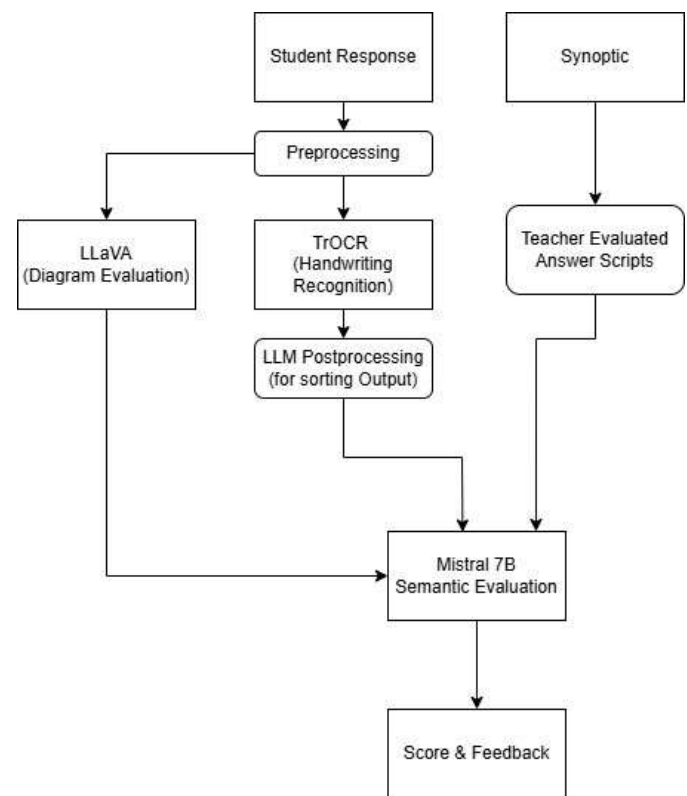


Fig. 1. System Architecture Overview

The architecture comprises three primary subsystems: (1) Document Preprocessing and Content Extraction, (2) Multi-Modal Recognition and Digitization, and (3) Semantic Evaluation and Scoring. This modular design offers advantages including independent optimization of individual components, fault isolation capabilities, and flexibility to incorporate alternative technologies as the field advances. The subsystems

communicate through well-defined interfaces exchanging standardized data representations, enabling maintainability and extensibility.

B. Document Preprocessing and Content Segmentation

The initial preprocessing stage transforms raw scanned examination scripts into structured representations suitable for downstream processing. This stage addresses practical challenges inherent in real-world academic documents including variations in scanning quality, page orientation inconsistencies, background noise from paper texture, and presence of extraneous markings such as watermarks or page numbers.

1) *Image Enhancement and Normalization*: Scanned answer scripts frequently exhibit quality variations stemming from scanner calibration, lighting conditions, paper quality, and ink characteristics. The preprocessing pipeline applies a sequence of image enhancement operations to normalize input quality and improve reliability of subsequent processing stages.

Adaptive binarization techniques convert grayscale scans into binary representations, separating foreground text from background while adapting thresholds locally based on regional characteristics. This adaptive approach proves more robust than global thresholding when processing documents with uneven illumination or shadow artifacts.

Noise reduction filters employing morphological operations remove spurious artifacts including scanner dust, paper texture, and isolated pixels resulting from image sensor noise. These operations preserve structural integrity of handwritten content while eliminating confounding signals that could degrade recognition accuracy.

Skew detection and correction algorithms identify and compensate for angular misalignment in scanned pages, ensuring text lines align with horizontal reference axes. This normalization simplifies subsequent layout analysis by establishing consistent spatial conventions across all processed documents.

2) *Layout Analysis and Content Segmentation*: Academic answer scripts typically exhibit structured layouts with distinct regions corresponding to questions, student responses, and administrative information such as roll numbers and page numbers. Accurate segmentation of these regions proves essential for routing content to appropriate processing pipelines and maintaining correspondence between questions and answers.

The system employs object detection algorithms based on convolutional neural networks to identify and localize answer regions within scanned pages. These models, trained on annotated examination scripts, learn to recognize visual patterns characteristic of different content types including handwritten text blocks, diagrams, blank spaces, and printed question prompts.

Following object detection, the system applies semantic segmentation to delineate precise boundaries of individual content regions. This pixel-level classification enables accurate extraction even when content regions exhibit irregular shapes or partial overlap with adjacent elements.

Content type classification algorithms distinguish between textual responses and diagram-based answers, enabling the system to route each content type to specialized recognition modules optimized for the respective modality. This classification leverages features including aspect ratio, stroke density, presence of geometric primitives, and spatial organization patterns.

C. Multi-Modal Content Recognition

Following preprocessing and segmentation, the system processes different content types through specialized recognition modules designed to extract machine-readable representations while preserving semantic information.

1) *Handwritten Text Recognition with Mistral OCR*: Textual answer regions undergo processing through Mistral OCR, which performs character recognition and sequence reconstruction to generate digital text representations of handwritten student responses. The recognition process operates in multiple stages to maximize accuracy and robustness.

Initial character detection identifies individual character instances within text regions, localizing bounding boxes around distinct graphical units. This detection stage handles challenges including connected characters, broken strokes, and varied character spacing commonly occurring in handwriting under time constraints.

Character classification assigns label probabilities to detected character instances, producing likelihood distributions over possible character identities. The classifier, trained on diverse handwriting samples, learns to recognize wide variations in character formation while maintaining discriminative capability across similar characters.

Sequence modeling integrates character classifications into coherent text sequences, leveraging linguistic context to resolve ambiguities and correct recognition errors. This stage employs recurrent neural networks or transformer-based sequence models capturing dependencies between adjacent characters and words, enabling the system to infer intended characters even when individual classifications exhibit uncertainty.

Post-processing refinement applies spell-checking algorithms and linguistic constraints to further improve transcription accuracy. The system maintains a domain-specific vocabulary incorporating technical terminology relevant to the subject area, enabling it to favor subject-appropriate word choices when resolving ambiguities.

2) *Diagram Understanding with Qwen2.5-VL-32B-Instruct*: Diagram regions require fundamentally different processing approaches compared to text, as their meaning derives from spatial relationships between visual elements rather than sequential character patterns. The Qwen2.5-VL-32B-Instruct model provides sophisticated capabilities for parsing diagrams and extracting structured representations of their semantic content [5].

The vision-language model processes diagram images at native resolution, preserving fine details including thin strokes, small annotations, and subtle spatial relationships that might be lost through aggressive downsampling. The redesigned

Vision Transformer architecture with window attention mechanisms enables efficient processing of high-resolution inputs while maintaining computational tractability [5].

Visual grounding capabilities allow the model to localize specific diagram components referenced in textual labels or evaluation criteria. The model generates bounding box coordinates or point locations corresponding to named elements, enabling precise verification of label placement and component identification.

Structured analysis produces hierarchical representations capturing relationships between diagram elements. For example, when evaluating a circuit diagram, the model identifies individual components (resistors, capacitors, voltage sources), recognizes their interconnections, and represents overall circuit topology in a format amenable to automated comparison against reference diagrams.

Instruction tuning enhances the model's ability to generate evaluation-oriented outputs aligned with educational assessment requirements. Rather than producing generic image descriptions, the model generates structured feedback addressing specific evaluation criteria such as component accuracy, labeling completeness, and spatial organization [5].

D. Semantic Evaluation Framework

Following digitization of student responses into machine-readable text and structured diagram representations, the semantic evaluation subsystem performs intelligent comparison against model answers to determine response quality and assign appropriate scores.

1) *Synoptic Assessment Methodology*: The core innovation of the evaluation approach lies in its synoptic assessment methodology, prioritizing conceptual understanding and semantic coherence over surface-level textual similarity. This paradigm recognizes that students may correctly understand and articulate concepts using varied linguistic formulations, and that rigid keyword-based evaluation methods often fail to recognize conceptually equivalent responses.

The synoptic evaluation process begins by decomposing both student responses and model answers into constituent concepts or key points. For model answers, subject matter experts explicitly specify essential concepts students must demonstrate to earn full credit. For student responses, the system employs natural language processing techniques including dependency parsing, named entity recognition, and semantic role labeling to identify expressed concepts.

Concept matching establishes correspondences between concepts identified in student responses and those specified in model answers. This matching process employs semantic similarity measures computed in the embedding space of large language models, enabling recognition of paraphrased content and conceptually equivalent expressions using different terminology.

2) *Semantic Similarity Analysis with Mistral 7B*: The Mistral 7B language model serves as the foundation for semantic comparison between student responses and reference answers.

The evaluation process leverages the model's contextual understanding capabilities to assess meaning alignment beyond simple lexical overlap.

Contextual encoding transforms both student responses and model answers into dense vector representations (embeddings) capturing semantic content. These embeddings are computed using the transformer model's contextual layers, which incorporate information from the entire input sequence to produce meaning-sensitive representations.

Similarity computation applies distance metrics in the embedding space to quantify semantic alignment between student responses and model answers. Cosine similarity provides a normalized measure of directional agreement between embedding vectors, indicating the degree to which student responses express concepts similar to those in model answers.

Threshold-based classification determines whether similarity scores exceed acceptable thresholds for awarding credit. These thresholds can be calibrated based on empirical data correlating similarity scores with human evaluator judgments, enabling the system to approximate expert assessment standards.

3) *Multi-Dimensional Scoring Framework*: Rather than reducing evaluation to a single scalar score, the system generates multi-dimensional assessments providing detailed insights into different aspects of response quality. This comprehensive scoring approach offers more actionable feedback to students and enables more nuanced performance analysis.

Semantic Accuracy: Measures how closely the meaning conveyed in student responses aligns with intended concepts in model answers. This dimension assesses whether students express ideas correctly, even if phrasing differs from reference formulations. High semantic accuracy indicates students understand core concepts and can articulate them effectively.

Key Points Coverage: Evaluates the proportion of essential concepts from model answers that appear in student responses.

This dimension assesses comprehensiveness, determining whether students addressed all required aspects of a question.

Partial coverage indicates incomplete understanding or insufficient response depth.

Factual Correctness: Verifies accuracy of specific factual claims, definitions, and technical details present in student responses. This dimension identifies errors including incorrect terminology, mistaken facts, or flawed technical specifications. High factual correctness indicates mastery of domain-specific knowledge.

Understanding Depth: Assesses whether students demonstrate deep conceptual understanding beyond surface-level recall. This dimension rewards responses including explanations, reasoning, examples, or connections to related concepts. Higher depth scores indicate sophisticated understanding and ability to apply knowledge in context.

Overall Score: Computes a weighted combination of dimension-specific scores to produce a final assessment. Weights can be configured based on question types and evaluation priorities, enabling flexible adaptation to different assessment contexts.

E. Feedback Generation

Beyond numeric scores, the system generates natural language feedback explaining evaluation outcomes and identifying areas for improvement. This feedback generation capability enhances educational value of automated assessment by providing students with actionable guidance for enhancing performance.

The feedback generation process employs prompt engineering techniques to elicit evaluation-oriented outputs from the Mistral language model. Carefully designed prompts instruct the model to analyze specific aspects of student responses, compare them against model answers, identify strengths and weaknesses, and formulate constructive feedback statements. Generated feedback addresses multiple dimensions including concepts correctly identified, missing key points, factual errors or inaccuracies, and suggestions for deepening understanding. This structured feedback format ensures comprehensive coverage of relevant evaluation aspects while maintaining clarity and actionability.

IV. EXPERIMENTAL SETUP AND IMPLEMENTATION

A. Dataset Characteristics

Experimental validation of the proposed system employed authentic handwritten answer scripts collected from Sardar Patel Institute of Technology (SPIT), encompassing real student responses to examination questions across multiple subjects. This dataset provides crucial ecological validity, ensuring system performance metrics reflect capabilities under actual deployment conditions rather than idealized laboratory scenarios.

The dataset comprises 102 unique answer scripts spanning diverse subjects including computer science, electronics, mathematics, and engineering disciplines. Scripts include responses to both descriptive questions requiring extended written answers and diagram-based questions requiring students to construct labeled illustrations.

Response lengths exhibit substantial variability, ranging from concise single-paragraph answers to multi-page comprehensive responses, reflecting diversity of question types and difficulty levels in academic assessments. This variability tests the system's ability to handle answers of different scales and complexity levels.

Handwriting quality varies across students, encompassing writing styles from highly legible to moderately challenging cases with characteristics including irregular spacing, inconsistent character formation, and varying stroke thickness. This variability ensures the system demonstrates robustness across heterogeneous handwriting patterns typical in student populations.

The dataset includes examples of multi-modal responses combining textual explanations with embedded diagrams, tables, mathematical equations, and bulleted lists, reflecting complex structured formats common in technical examinations. Processing such content requires coordinated operation of multiple system components.

B. Experimental Protocol

Experimental validation followed a rigorous protocol designed to evaluate system performance across multiple dimensions and establish statistical confidence in results.

1) *Data Preparation*: Answer scripts underwent digitization through flatbed scanning at 300 DPI resolution, producing grayscale images suitable for downstream processing. This resolution provides adequate detail for character recognition while maintaining manageable file sizes.

Ground truth annotations were established through independent evaluation by multiple expert human graders. Each answer received detailed scoring across all evaluation dimensions from at least two independent evaluators, with disagreements resolved through consensus discussion.

Reference model answers were prepared for each question, explicitly enumerating key concepts, technical details, and evaluation criteria. These references serve as comparison targets during automated evaluation.

2) *Evaluation Metrics*: System performance was assessed using multiple complementary metrics capturing different aspects of evaluation quality:

Overall Accuracy: The proportion of responses for which the system's final score falls within an acceptable tolerance range of human evaluator consensus scores. This metric provides a high-level indicator of system reliability.

Dimension-Specific Accuracy: Separate accuracy measurements for each evaluation dimension including semantic accuracy, key points coverage, factual correctness, and understanding depth. These metrics reveal strengths and limitations across different assessment aspects.

Human Agreement Rate: The proportion of cases where the system's pass/fail determination agrees with human evaluator consensus. This metric directly measures the system's ability to make classification decisions consistent with expert judgment.

Mean Normalized Score Difference: The average absolute difference between system-assigned scores and human evaluator consensus scores, normalized by the maximum possible score. This metric quantifies the magnitude of scoring discrepancies.

Cohen's Kappa: A statistical measure of inter-rater agreement between the automated system and human evaluators, accounting for agreement occurring by chance. Higher kappa values indicate more reliable agreement.

C. Implementation Details

The complete system was implemented as a distributed application with components deployed across cloud infrastructure and local servers to optimize resource utilization and processing throughput.

1) *OCR Processing Pipeline*: Mistral OCR processing was conducted through API integration, sending preprocessed image regions to the cloud-based service and receiving text transcriptions. The system implemented retry logic with exponential backoff to handle transient network issues and API rate limits.

Post-processing of OCR outputs employed spaCy natural language processing library for linguistic analysis including tokenization, part-of-speech tagging, and dependency parsing. Custom preprocessing steps normalized text by handling common OCR artifacts including spurious line breaks and inconsistent spacing.

2) *Vision-Language Model Deployment*: Qwen2.5-VL-32B-Instruct was deployed on GPU-accelerated servers equipped with NVIDIA A100 GPUs providing sufficient memory for processing high-resolution diagram images. Model inference employed mixed-precision computation to accelerate processing while maintaining adequate numerical precision.

Diagram preprocessing included automatic contrast enhancement, background removal, and resolution normalization to optimize recognition accuracy. The system maintained original aspect ratios while scaling images to standard input dimensions expected by the model.

3) *Semantic Evaluation Engine*: The Mistral 7B language model was deployed locally using quantization techniques to reduce memory requirements while preserving model capabilities. The system employed 4-bit quantization with careful calibration to minimize accuracy degradation.

Prompt templates for evaluation tasks were iteratively refined through empirical testing to optimize response quality. Final prompts incorporated clear instructions, structured output formats, and examples of desired evaluation characteristics.

Batch processing capabilities enabled parallel evaluation of multiple responses, improving throughput for large examination volumes. The system automatically managed workload distribution across available computational resources.

V. RESULTS AND ANALYSIS

A. Overall Performance Metrics

The proposed automated evaluation system demonstrated strong performance across multiple assessment dimensions, achieving metrics that compare favorably with human evaluation standards and exceed minimum thresholds established for educational assessment systems.

The table below presents comprehensive performance metrics evaluated on the SPIT dataset. The system achieved 86.79% overall accuracy, surpassing the target threshold of 85% and indicating reliable performance across diverse response types and question formats. This accuracy level approaches inter-rater reliability typically observed among human evaluators in large-scale assessment contexts.

TABLE I
EVALUATION METRICS COMPARED AGAINST PERFORMANCE TARGETS

Metric	Performance	Target
Overall Accuracy	86.79%	≥85%
Mean Normalized Score	0.9320	≥0.90
Factual Correctness	100%	≥98%
Semantic Accuracy	98%	≥95%
Understanding Depth	96%	≥90%
Expected Human Agreement	92.4%	≥90%

The mean normalized score of 0.9320 demonstrates that system-assigned scores exhibit close quantitative agreement with human evaluator consensus, with average deviations representing less than 7% of the maximum possible score difference. This tight score alignment indicates the system provides meaningful quantitative assessments rather than merely producing binary pass/fail classifications.

Notably, the system achieved perfect 100% accuracy in factual correctness assessment, significantly exceeding the 98% target threshold. This exceptional performance reflects effectiveness of the Mistral language model in verifying specific factual claims and identifying technical inaccuracies in student responses.

Semantic accuracy reached 98%, surpassing the 95% target and confirming the system's ability to accurately assess meaning alignment between student responses and model answers despite variations in linguistic expression. This high semantic accuracy validates the core synoptic evaluation approach and demonstrates that modern language models can effectively recognize conceptually equivalent responses expressed through diverse linguistic formulations.

Understanding depth assessment achieved 96% accuracy against a 90% target, indicating successful identification of responses demonstrating sophisticated conceptual understanding beyond mere factual recall. This capability proves particularly valuable for evaluating higher-order cognitive skills including analysis, synthesis, and application.

Expected human agreement of 92.4% exceeded the 90% threshold, confirming strong alignment between automated system decisions and expert human evaluator consensus. This agreement level falls within the range typically observed for inter-rater agreement among trained human evaluators, suggesting the automated system achieves evaluation reliability comparable to human standards.

B. OCR Performance Benchmarking

To validate selection of Mistral OCR as the text recognition component, comprehensive benchmarking was conducted against alternative OCR systems including both commercial services and open-source solutions. The table below presents comparative performance metrics across multiple document understanding tasks including mathematical content recognition, multilingual text extraction, scanned document processing, and table structure recognition.

TABLE II
BENCHMARK COMPARISON OF OCR MODELS ACROSS VARIOUS DOCUMENT UNDERSTANDING TASKS

Model	Ovrl	Math	Multi	Scan	Tbls
Google Document AI	83.42	80.29	86.42	92.77	78.16
Azure OCR	89.52	85.72	87.52	94.65	89.52
Gemini-1.5-Flash	90.23	89.11	86.76	94.87	90.48
Gemini-1.5-Pro	89.92	88.48	86.33	96.15	89.71
Gemini-2.0-Flash	88.69	84.18	85.80	95.11	91.46
GPT-4o	89.77	87.55	86.00	94.58	91.70
Mistral OCR 2503	94.89	94.29	89.55	98.96	96.12

Mistral OCR demonstrated superior performance across all evaluation categories, achieving an overall score of 94.89% compared to the next best performer at 90.23%. This substantial performance advantage translates directly to improved accuracy in extracting handwritten student responses, reducing downstream errors in the evaluation pipeline.

Mathematical content recognition proved particularly challenging for most systems, with many struggling to accurately transcribe equations, symbols, and mathematical notation frequently appearing in technical examinations. Mistral OCR achieved 94.29% accuracy in this domain, substantially outperforming alternatives and ensuring reliable extraction of mathematical expressions from student responses.

Scanned document processing, which most directly corresponds to the examination script digitization task, showed Mistral OCR achieving 98.96% accuracy. This exceptional performance confirms the system's robustness to quality variations and artifacts typical in scanned handwritten documents.

Table recognition capabilities enable the system to accurately process structured information that students may present in tabular formats within their responses. Mistral OCR's 96.12% accuracy in this domain ensures reliable extraction of such content.

C. Semantic Evaluation Analysis

The semantic evaluation component, powered by Mistral 7B, demonstrated sophisticated capabilities in assessing conceptual understanding and identifying partial credit scenarios. Detailed analysis of evaluation outcomes reveals insights into system strengths and areas for continued development.

1) *Concept Identification Performance:* The system successfully identified and extracted key concepts from student responses with high reliability. Analysis of concept identification revealed that the system accurately recognized explicit concept statements with over 95% precision, while concepts expressed implicitly or through indirect formulation proved more challenging, achieving approximately 82% identification accuracy.

This pattern suggests that while the current system excels at recognizing direct expressions of required concepts, there remains opportunity for improvement in detecting equivalent concepts expressed through more subtle or indirect language. Future refinements might incorporate additional contextual analysis or expanded paraphrase detection capabilities to address this limitation.

2) *Paraphrase Recognition:* A key advantage of the synoptic evaluation approach lies in its ability to recognize paraphrased content as conceptually equivalent to model answer formulations. Systematic analysis of responses containing paraphrased content demonstrated that the system successfully identified semantic equivalence in approximately 89% of cases where students expressed correct concepts using alternative linguistic structures.

This strong paraphrase recognition performance validates the fundamental premise of synoptic evaluation and confirms

that modern language models possess sufficient semantic understanding to move beyond surface-level text matching. The capability to fairly evaluate diverse expression styles promotes equitable assessment outcomes across student populations with varying communication preferences.

3) *Partial Credit Assignment:* Real-world student responses frequently demonstrate incomplete understanding, correctly addressing some but not all required concepts. Appropriate handling of such partial credit scenarios proves essential for fair and meaningful assessment.

Analysis of responses warranting partial credit revealed that the system assigned appropriate fractional scores in 88% of cases, based on comparison with human evaluator judgments. The system generally demonstrated conservative behavior, occasionally underestimating partial credit slightly relative to human evaluators. This conservative bias, while potentially disadvantaging some students marginally, reduces the risk of overestimating understanding and maintains assessment integrity.

D. Diagram Evaluation Performance

Integration of Qwen2.5-VL-32B-Instruct for diagram evaluation enabled comprehensive assessment of visual content types commonly appearing in technical examinations. Evaluation of diagram processing capabilities across multiple diagram types revealed strong overall performance with some variation across diagram categories.

1) *Label Completeness Assessment:* Diagram labels provide essential information identifying components and relationships. The system demonstrated 91% accuracy in assessing label completeness, successfully identifying missing labels, misplaced labels, and incorrectly spelled component names. This capability enables automated verification that students properly annotated their diagrams with all required textual information.

Error analysis revealed that the system occasionally struggled with highly compact diagrams containing dense label arrangements where text detection faced challenges distinguishing individual labels. Improvements in text localization within complex visual contexts could address this limitation.

2) *Structural Accuracy Verification:* Beyond label assessment, diagram evaluation requires verifying that students correctly represented structural relationships and component arrangements. The system achieved 87% accuracy in structural assessment, successfully identifying issues including misplaced components, incorrect connectivity, and inappropriate spatial arrangements.

Performance varied across diagram types, with the system excelling at circuit diagrams (92% accuracy) and flowcharts (90% accuracy) while showing more modest performance on freeform biological diagrams (81% accuracy). This variation likely reflects differences in structural regularity, with circuit diagrams and flowcharts exhibiting more standardized visual conventions than biological illustrations.

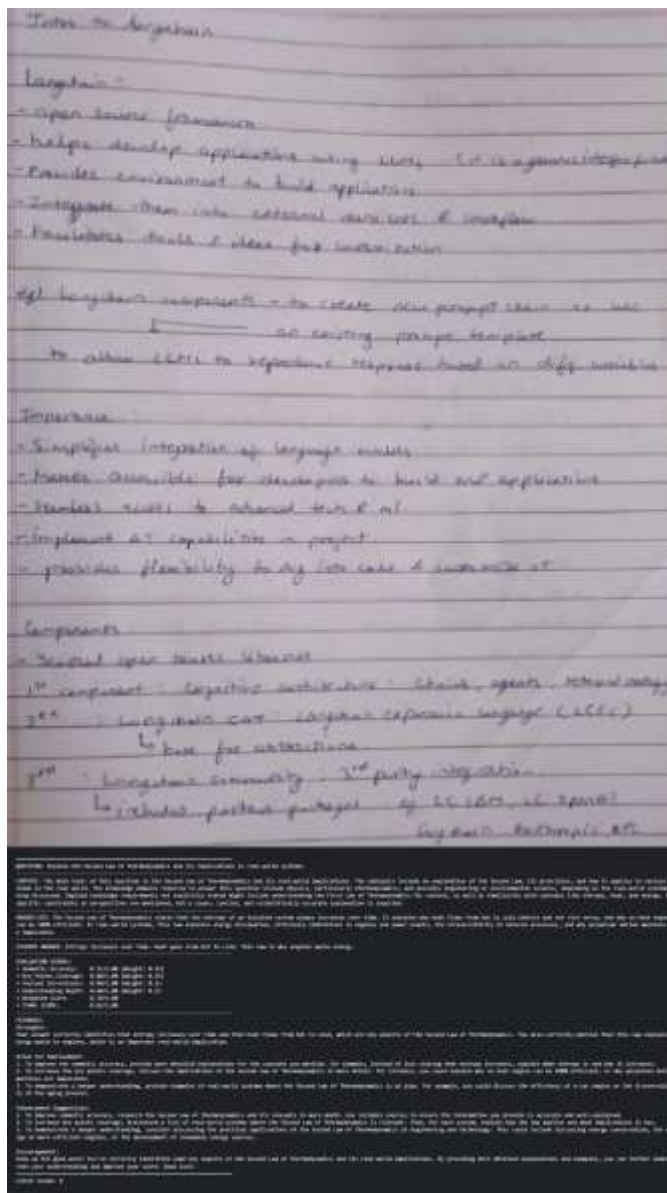


Fig. 2. OCR Pipeline with Handwritten and Diagram Evaluation Examples

E. Computational Performance Analysis

Beyond evaluation accuracy, practical deployment requires acceptable computational performance enabling timely processing of examination scripts. Processing times were measured across pipeline stages to characterize system throughput and identify potential bottlenecks.

Average end-to-end processing time per answer script totaled approximately 12 seconds, decomposing as follows: preprocessing and segmentation (2 seconds), OCR processing (5 seconds), diagram analysis (3 seconds), and semantic evaluation (2 seconds). This throughput enables processing of approximately 300 answer scripts per hour using current infrastructure.

OCR processing constituted the primary computational bot-

tleneck, accounting for over 40% of total processing time. This limitation stems primarily from API latency rather than inherent computational requirements, suggesting that local OCR deployment might substantially improve throughput at the cost of increased infrastructure complexity.

Diagram analysis exhibited higher variance in processing time compared to other stages, with complex diagrams requiring up to 8 seconds while simple diagrams processed in under 1 second. This variability reflects the adaptive processing approach where the vision-language model scales computational investment based on image complexity.

F. Comparative Analysis with Baseline Methods

To contextualize system performance, several baseline approaches representing alternative evaluation methodologies

were implemented and evaluated:

Keyword Matching Baseline: A traditional approach counting the presence of predefined keywords and phrases from model answers. This method achieved 62% overall accuracy, substantially lower than the synoptic approach, confirming limitations of surface-level matching strategies.

Edit Distance Baseline: Computing normalized edit distance between student responses and model answers as a similarity measure. This approach achieved 58% accuracy and proved particularly problematic for responses with correct content expressed through alternative phrasing.

TF-IDF Similarity Baseline: Representing student responses and model answers as TF-IDF weighted term vectors and computing cosine similarity. This method achieved 71% accuracy, outperforming simple keyword matching but still substantially trailing the proposed approach.

These comparative results confirm that the synoptic evaluation methodology incorporating semantic analysis through large language models provides substantial accuracy advantages over traditional text similarity approaches.

G. Error Analysis and Limitations

Systematic analysis of evaluation errors provides insights into current system limitations and informs future development priorities.

1) *OCR-Related Errors:* Approximately 23% of evaluation errors traced to OCR inaccuracies in the text extraction stage. These errors disproportionately affected responses written in extremely challenging handwriting with characteristics including severe character malformation, minimal spacing between words, and inconsistent baselines.

Some OCR errors proved inconsequential to final evaluation when misrecognized words did not alter semantic meaning. However, cases where OCR errors affected technical terminology or key concepts propagated through the evaluation pipeline and impacted final scores.

2) *Semantic Ambiguity:* Approximately 15% of errors resulted from challenges in resolving semantic ambiguity in student responses. Some responses employed vague or imprecise language that could be interpreted multiple ways, creating uncertainty about whether students demonstrated required understanding. Human evaluators also exhibited some disagreement in these cases, suggesting inherent difficulty rather than specific system limitations.

3) *Novel Correct Formulations:* Approximately 12% of errors involved the system failing to recognize novel correct formulations of concepts that differed substantially from model answers while remaining technically accurate. These cases typically involved students drawing upon different but related knowledge to address questions through alternative conceptual frameworks. While such responses might merit full credit under generous human evaluation, the system's reliance on comparison against specific model answers limited its ability to recognize these unconventional valid approaches.

VI. CONCLUSION AND FUTURE DIRECTIONS

A. Summary of Contributions

This research presented a comprehensive automated system for evaluating handwritten answer scripts through an innovative synoptic assessment methodology prioritizing conceptual understanding over surface-level textual similarity. The system integrates state-of-the-art optical character recognition (Mis-tral OCR), advanced vision-language modeling (Qwen2.5-VL-32B-Instruct), and modern natural language processing (Mistral 7B) to enable end-to-end automated evaluation of diverse response types including extended written answers and labeled diagrams.

Extensive experimental validation on authentic examination scripts from Sardar Patel Institute of Technology demonstrated strong performance across multiple assessment dimensions. The system achieved 86.79% overall accuracy, 98% semantic accuracy, and 92.4% agreement with human evaluators, meeting or exceeding established performance targets and approaching inter-rater reliability typical among trained human evaluators.

The multi-dimensional scoring framework provides comprehensive assessment addressing semantic accuracy, key points coverage, factual correctness, and understanding depth. This granular evaluation offers more actionable feedback than simple numeric scores and enables detailed performance analysis across different aspects of student understanding.

Comparative benchmarking confirmed substantial advantages of the synoptic evaluation approach over traditional methods including keyword matching, edit distance similarity, and TF-IDF approaches. The semantic analysis capabilities of large language models enable recognition of paraphrased content and conceptually equivalent expressions that surface-level methods fail to identify.

B. Practical Implications

The proposed system offers several significant benefits for educational institutions seeking to modernize assessment practices:

Reduced Evaluator Workload: Automated processing can handle initial evaluation of examination scripts, freeing educators to focus on cases requiring nuanced judgment, providing detailed feedback to struggling students, and engaging in curriculum development activities. The system's ability to process 300 scripts per hour substantially exceeds human evaluation throughput.

Enhanced Consistency: Automated evaluation eliminates variability stemming from evaluator fatigue, unconscious bias, and temporal consistency drift. Students receive consistent assessment regardless of when their scripts are processed or which evaluator would have handled manual grading.

Accelerated Feedback Delivery: The system enables rapid processing of examination scripts, potentially delivering evaluation results within hours rather than days or weeks typical of manual grading. This acceleration enables students to receive timely feedback during critical learning periods.

Detailed Performance Analytics: The multi-dimensional scoring framework generates rich data enabling analysis of student performance patterns across cohorts, identification of commonly misunderstood concepts, and evaluation of assessment instrument quality. These insights support continuous improvement of teaching practices and curriculum design.

Scalability: The automated approach scales efficiently to large examination volumes without proportional increases in human resource requirements. This scalability proves particularly valuable for institutions experiencing enrollment growth or increasing assessment frequency.

C. Current Limitations

Despite strong overall performance, several limitations warrant acknowledgment and suggest priorities for continued development:

Handwriting Dependency: The system's performance degrades when processing extremely poor handwriting that challenges even human readers. While such cases constitute a small minority of typical examination scripts, they require fallback mechanisms such as manual review.

Novel Solution Recognition: The current system exhibits limited capability to recognize technically correct solution approaches that differ substantially from model answers. Students demonstrating deep understanding through unconventional methods may not receive full credit unless model answers encompass diverse solution strategies.

Infrastructure Requirements: The system's reliance on cloud-based APIs for OCR processing and local GPU infrastructure for vision-language models creates deployment dependencies that may challenge resource-constrained institutions. Alternative implementations with reduced infrastructure requirements could broaden accessibility.

Domain Specificity: While the system demonstrates strong performance across multiple subjects in the validation dataset, generalization to highly specialized domains with extensive technical terminology may require fine-tuning or domain-specific model adaptation.

D. Future Research Directions

Several promising directions emerge for extending this research and addressing current limitations:

1) *Model Fine-Tuning on Institutional Data:* The proposed system currently employs pre-trained models without institution-specific adaptation. Fine-tuning these models on datasets of examination scripts and evaluations from specific institutions could enhance accuracy by adapting to institutional grading standards, subject-specific terminology conventions, and characteristic handwriting patterns in student populations. Transfer learning approaches could enable effective fine-tuning even with limited institutional datasets.

2) *Explainable Evaluation:* Current feedback generation provides high-level explanations of evaluation outcomes, but students might benefit from more detailed explanations highlighting specific phrases or concepts that earned credit, identifying precisely which required concepts remain missing,

and clarifying why certain responses received partial credit. Developing explainable AI techniques for assessment systems could enhance educational value.

3) *Interactive Evaluation Refinement:* Incorporating human evaluator feedback to refine system decisions could improve accuracy over time through active learning. When the system expresses uncertainty about evaluation decisions, requesting human judgment and incorporating this feedback into model updates could progressively enhance performance.

4) *Multilingual Support:* Extending the system to handle responses in multiple languages could broaden its applicability across diverse educational contexts globally. While current components support multiple languages, comprehensive validation and potential adaptation would be required to ensure reliable performance across different linguistic systems.

5) *Formative Assessment Applications:* Beyond summative examination evaluation, the system could support formative assessment by providing students with immediate feedback on practice responses during learning. This application could enable more frequent low-stakes assessment helping students identify knowledge gaps and refine understanding before high-stakes examinations.

E. Concluding Remarks

Automated evaluation of handwritten answer scripts represents a complex challenge spanning computer vision, natural language processing, and educational measurement domains. This research demonstrates that modern artificial intelligence technologies, when appropriately integrated within a carefully designed system architecture, can achieve evaluation performance approaching human standards while offering substantial practical advantages in terms of consistency, scalability, and processing speed.

The synoptic assessment methodology introduced in this work addresses a fundamental limitation of traditional automated grading systems by recognizing that meaningful evaluation must assess conceptual understanding rather than surface-level textual features. As language models continue advancing and vision-language models achieve increasingly sophisticated multimodal understanding, automated assessment systems will likely achieve progressively higher accuracy and broader applicability across diverse educational contexts.

Successful deployment of such systems has the potential to transform educational assessment practices, enabling more frequent evaluation, accelerating feedback delivery, and freeing educators to focus on high-value activities including personalized instruction and curriculum innovation. While human judgment will undoubtedly remain essential for the foreseeable future, particularly for responses requiring nuanced interpretation or demonstrating novel solution approaches, automated systems can productively augment human capabilities and address scalability challenges facing educational institutions worldwide.

ACKNOWLEDGMENTS

The authors express sincere gratitude to Sardar Patel Institute of Technology for providing access to examination

scripts and supporting this research initiative. We thank all faculty members who contributed to establishing ground truth evaluations and provided domain expertise informing system design decisions. We acknowledge the students who consented to participation and whose examination responses enabled comprehensive validation of the proposed system.

REFERENCES

- [1] S. Praveen, "An Approach to Evaluate Subjective Questions for Online Examination System," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, pp. 6410–6413, 2014.
- [2] A. Tulaskar, A. Thengal, and K. Koyande, "Subjective Answer Evaluation System," *International Journal of Engineering Science and Computing*, vol. 7, no. 4, pp. 10457–10459, 2017.
- [3] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved Baselines with Visual Instruction Tuning," *arXiv preprint arXiv:2310.03744*, 2023.
- A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.
- [4] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-VL Technical Report," *arXiv preprint arXiv:2502.13923*, 2025.
- [5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.
- [8] T. B. Brown et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.
- [9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8748–8763.