Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

A Neuro-Symbolic Framework for Explainable Automated Grading of Handwritten Responses via TrOCR and Knowledge Graph Reasoning

Dr. Ravi Lanke
Department of Computer Science
JAIN (Deemed-to-be-University)
Bengaluru,India
ravikumar.l@jainuniversity.ac.in

Aayush JP
Department of Computer Science
JAIN (Deemed-to-be-University)
Bengaluru,India
22btrcn008@jainuniversity.ac.in

Chethan Kumar HL
Department of Computer Science
JAIN (Deemed-to-be-University)
Bengaluru,India
22btrcn062@jainuniversity.ac.in

Varun R
Department of Computer Science)
JAIN (Deemed-to-be-University)
Bengaluru,India
22btrcn314@jainuniversity.ac.in

MM Neel Kariappa
Department of Computer Science
JAIN (Deemed-to-be-University)
Bengaluru,India
22btrcn161@jainuniversity.ac.in

Santholin Furtado
Department of Computer Science
JAIN (Deemed-to-be-University)
Bengaluru,India
22btrcn248@jainuniversity.ac.in

***Abstract*— Grading handwritten and descriptive examination answers places a significant burden on educators, particularly when dealing with diverse handwriting styles and the inherent subjectivity of manual scoring at scale. Generative language models have shown promise in automating parts of this process, yet their tendency toward hallucination and opaque decision-making makes them unsuitable as standalone grading systems in academic settings. This work presents a neuro-symbolic hybrid grading framework built around three integrated components: a domain-adapted handwritten text recognition pipeline, a rubric-grounded Knowledge Graph engine, and an LLM-based semantic reasoning module. For transcription, a TrOCR model is trained through a two-stage transfer learning process first on the IAM Handwriting Database, which contains 13,353 line-level annotated samples from 657 writers, then refined further on 150 handwritten examination scripts sourced from undergraduate students across three academic subjects. Over 10 fine-tuning epochs, training loss drops from 1.418 to 0.005 and validation CER reaches 6.86%, with a final test CER of 9.59% and WER of 21.35% on held-out IAM data. Once text is extracted, a directed weighted Knowledge Graph constructed from the instructor's rubric is used to measure concept coverage through cosine similarity matching with all-MiniLM-L6-v2 embeddings, while a concurrent LLM pass scores the response for coherence and factual correctness, returning structured output with a written justification. Final grades are produced by a weighted combination of both scores governed by a tunable parameter α. Tested across 150 answer scripts, the system reaches a Pearson correlation of 0.87 with human graders, a MAE of 0.94, and RMSE of 1.22 — surpassing keyword-matching (r = 0.61), KG-only (r = 0.71), and LLM-only (r = 0.74) configurations.**

Keywords*— *Automated Grading, TrOCR, Handwritten Text Recognition, Transfer Learning, Knowledge Graphs, Large Language Models, Hybrid Evaluation, Educational Technology, Semantic Similarity.

I. INTRODUCTION

The evaluation of descriptive and handwritten examinations is a very challenging process that has been time consuming to teachers. It is prone to human bias, exhaustion and lack of conformity in the use of evaluation criteria among large cohorts of students. Conventional Optical Character Recognition (OCR) systems and keyword-matching algorithms have not been able to cope with this task since they are not sensitive to irregular handwriting styles and are unable to provide the rich semantic information needed to identify conceptually equivalent but lexically different responses. Large Language Models (LLMs) although semantically strong are vulnerable to hallucination and do not provide the deterministic transparency required in auditable academic grading.

The proposed paper will suggest a hybrid automated assessment system that would overcome these issues by offering four closely coupled modules: a handwritten text recognition (HTR) module that is fine-tuned, a Knowledge Graph (KG) reasoning system, a semantic evaluation module based on an LLM, and a tunable hybrid scoring mechanism all of which are available as a web-based interface.

In the case of handwriting transcription, the framework uses TrOCR, which is a Transformer-based OCR model that uses a BEiT Vision Transformer encoder coupled with an autoregressive RoBERTa decoder. Two-stage transfer learning is used to train the model. Stage 1 will refine TrOCR on the IAM Handwriting Database a benchmark corpus consisting of 13,353 annotated image line texts created by 657 distinct writers, divided into 6,161 training, 976 validation and 2,915 test samples, with AdamW optimiser and a learning rate of 5e-5, a batch size of 8 and FP16 mixed-precision training across 10 epochs. The loss decreases between 1.418 and 0.005 during the first and 10th epochs respectively, the validation CER decreases steadily between 19.84% and 6.86%. Stage 2 Stage 1 is again refined on a custom dataset of 150 handwritten examination answer scripts of undergraduate students in three subject domain Computer Science, Artificial Intelligence and Social Affairs with a reduced learning rate of 1e-5 over 30 epochs to adjust the model towards domain-specific academic vocabulary and handwriting styles without

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

forgetting disastrously. The last model has reached a test set CER of 9.59% and WER of 21.35% on held-out IAM samples, which is within the acceptable range of downstream grading tasks.

The transcribed text is then assessed on two layers simultaneously. The Knowledge Graph engine builds a directed weighted graph G = (V, E) based on the rubric provided by the instructor, with nodes representing the desired concepts with importance and edges representing a logical relationship, i.e., cause, depends on, produces. The ideas of students are identified with the help of spaCy noun chunking and compared to the nodes of the rubric by cosine similarity calculated between all-MiniLM-L6-v2 sentence embeddings with a threshold t = 0.75 being chosen after trial and error to maximise F1-Score. Deterministic S^{KG} A deterministic KG coverage score S^{KG} is calculated as the weight of matched nodes divided by the total graph weight. The LLM reasoning module at the same time assesses the coherence and conceptual accuracy of the transcribed response with normalised scores of 0.0 to 1.0 and a human-readable justification string to guarantee explainability and educator auditability.

It is a full-stack web application consisting of a FastAPI backend and HTML5/TailwindCSS/JavaScript frontend that allows educators to upload answer sheet images and JSON rubrics using a browser interface and get a structured evaluation report back in real time. FinalScore = a x S^{KG} + (1 − α) × S^{LLM} is then computed as the final grade. 1) S^{LLM} x, a is a parameter in the tunable institution. With = 0.6 the system obtains a Pearson correlation of 0.87 with human graders on 150 evaluation scripts, a Mean Absolute Error of 0.94 and an RMSE of 1.22, which is superior to those of keyworing (r = 0.61), KG alone (r = 0.71) and LLM alone (r = 0.74).

The authors of this paper are unaware of any previous literature that combines a domain fine-tuned Transformer based HTR model with a deterministic, educator-defined Knowledge Graph engine and a (concurrent) LLM based reasoning component into one unified grading pipeline. The current methods use either LLMs independently to grade transcribed text [1,3,4], symbolic scoring on pre-typed responses alone [11], or graphs to store data instead of determine the matching of concepts deterministically [10]. The current work is the first to provide a combination of offline two-stage TrOCR fine-tuning, weighted KG matching based on rubrics with cosine similarity, and semantic reasoning on LLM on a custom-made examination dataset across three subject domains at the university level.

The rest of this paper will be organized in the following way: Section II reviews related literature. Section III describes the system architecture and methodology. Section IV describes the technical implementation. In section V, quantitative results and observations are displayed. Section VI talks about limitations and Section VII ends with work directions.

II. RELATED WORK

Adithya et al. [1] propose a multi-modal handwritten answer assessment system employing CRAFT for text region detection and TrOCR for line-level handwriting recognition, with a fine-tuned language model handling answer evaluation. TrOCR is trained on an undisclosed dataset comprising approximately 1,700 annotated answer lines, and the extracted text is subsequently passed to a pretrained LLM for scoring. The authors additionally develop an interactive web platform termed NeuroGrade for system deployment. While the work shares the use of TrOCR as a recognition backbone, it performs no domain-specific fine-tuning on examination data, relies entirely on a generative LLM for grading without any structured rubric verification, and incorporates no Knowledge Graph component or hybrid scoring mechanism.

Sanuvala and Fatima [2] investigate automated examination paper evaluation by feeding OCR-extracted text into a range of supervised machine learning classifiers. The approach follows a straightforward pipeline of document upload, text extraction, and classification-based scoring. However, the study does not report quantitative evaluation metrics, which limits reproducibility and prevents direct performance comparison with other systems.

Tania-Amanda et al. [3] survey the application of Large Language Models to automated grading across multiple disciplines, examining both zero-shot and few-shot configurations on established benchmarks including the ASAP and ASAP++ datasets. The authors additionally explore Retrieval-Augmented Generation (RAG) pipelines for answer evaluation and discuss the practical challenges of deploying LLM-based graders, including hallucination and domain sensitivity. The study, however, focuses exclusively on typed text and does not address handwriting recognition or structured rubric grounding.

Agnihotri et al. [4] present an AI-powered assessment system for handwritten answer sheets that combines OCR-based text extraction with multimodal LLMs, using LLMs for textual answers and Vision-Language Models for diagram evaluation. The system demonstrates competitive character error rates relative to traditional OCR engines through multimodal integration. The study does not incorporate Knowledge Graph-based concept verification, and the authors note that the system remains under active development with planned enhancements for additional model configurations and fine-tuning strategies.

Kamble et al. [5] propose a multi-modal automated evaluation framework that converts scanned answer sheets to images, extracts text using Tesseract OCR, and assesses responses through SBERT-based semantic similarity scoring, with CLIP employed for diagram analysis. The system achieves a reported grading accuracy of 93.3%. While the semantic similarity approach partially addresses the limitations of keyword matching, the framework relies on a single-pass similarity measure against model answers without deterministic concept-level verification or explainable score breakdowns.

Faseeh et al. [9] address automated essay scoring by combining contextual embeddings from RoBERTa with handcrafted linguistic features including grammar error counts, readability indices, and sentence length statistics within a lightweight LwXGBoost classifier. Evaluated on the AES dataset using Quadratic Weighted Kappa (QWK), LwXGBoost achieves a QWK of 0.941, outperforming BERT, SVM, and AdaBoost baselines. The work demonstrates strong performance on typed essay scoring but does not extend to handwritten inputs or structured rubric-based concept verification.

Anghel et al. [10] introduce GraderAssist, a graph-assisted multi-LLM evaluation framework designed for transparent and reproducible automated grading. Six open-source LLMs alongside GPT-4 independently score 220 responses to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

technical and argumentative questions using two predefined rubrics, with all scores persisted in a Neo4j graph database. Critically, the graph structure in GraderAssist serves purely as a storage and retrieval backend rather than as a scoring mechanism; no deterministic concept matching or weighted rubric traversal is performed over the graph. The system achieves an average mean score of 6.24, though the scoring scale and ground-truth comparison methodology are not fully specified.

Teranishi and Araki [11] explore Knowledge Graph integration for short answer grading by constructing concept graphs from pedagogical questions and comparing student responses against them using pretrained language models. While this work shares the KG-based structural evaluation philosophy of the present framework, it operates exclusively on clean typed text without any handwriting recognition component, and does not incorporate a concurrent LLM reasoning layer or a tunable hybrid scoring mechanism.

III. METHODOLOGY

The proposed system adopts a modular architecture designed to digitise handwritten exam answers, extract meaningful content, and evaluate them through a hybrid reasoning pipeline comprising four tightly integrated layers: a web-based user interface, a fine-tuned Handwritten Text Recognition (HTR) module, a deterministic Knowledge Graph (KG) engine, and an LLM semantic reasoning module.

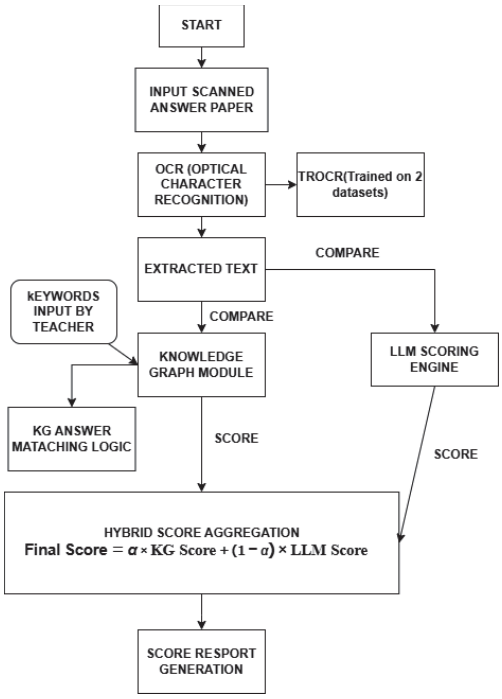

FIG I : Flowchart representing the working of the Prototype.

A. System Architecture

The architecture consists of four major operational layers:

a) User Interface (UI): A full-stack web application built with a FastAPI backend and an HTML5/TailwindCSS/JavaScript frontend. Educators upload the exam question and a structured JSON rubric through the browser interface, while student handwritten answer sheets are submitted as image files. Evaluation reports are returned in real time with a full breakdown of matched and missing concepts alongside the LLM justification.

b) Handwritten Text Recognition Layer: Converts handwritten image content into machine-readable text using a fine-tuned TrOCR model trained in a two-stage transfer learning paradigm on the IAM Handwriting Database and a custom examination dataset.

c) Knowledge Graph (KG) Engine: Transforms the teacher-provided rubric into a directed weighted graph and employs NLP-based semantic matching to verify concept coverage against the student's transcribed response.

d) LLM Reasoning Layer: A generative language model evaluates the transcribed text for coherence and conceptual accuracy, returning normalised scores and a human-readable justification. The system merges the KG coverage score with the LLM scores via a tunable hybrid scoring formula to produce the final grade.

B. Dataset

Two datasets were used to carry out experiments. The original training data is the IAM Handwriting Database (Marti and Bunke, 2002), which is a common benchmark corpus consisting of 13,353 annotated text line images of 657 distinct writers, divided into 6,161 training, 976 validation and 2,915 test samples. In order to fit the model to the target domain, an additional custom dataset was built using 150 handwritten examination answer scripts taken among the undergraduate students at JAIN (Deemed-to-be-University). A smartphone camera digitised the answer sheets with a minimum resolution of 1920x1080 pixels and manually annotated by human annotators to generate ground-truth labels and generated about 200 annotated line images after segmentation. The custom data was separated into training (80%), validation (10%), and test (10%) data, and the test data were only assessed at the end. The inter-annotator agreement was confirmed with the help of Cohen Kappa (k =0.83), which proved high labelling consistency. In the grading assessment, both 150 scripts were scored by two human expert graders independently and the average scores of the graders were taken as the ground truth with inter-grader agreement giving a Cohens Kappa of k = 0.81..

C. Fine-Tuned TrOCR Text Extraction Pipeline

The handwriting transcription system is based on TrOCR which is a sequence-to-sequence model that contains BEiT Vision Transformer encoder and RoBERTa autoregressive decoder. The encoder splits the input image into 16x16 pixel patches, projects them into a high dimensional embedding space and uses multi-head self-attention layers to create rich contextual visual representations. The decoder also produces the sequence of output tokens in an autoregressive manner, paying attention to the encoder output as well as the previously decoded tokens by means of cross-attention, allowing the decoder to reason jointly with language and visual information.

The process of training goes through two phases. Stage 1 is a fine-tuning of the pretrained microsoft/trocr-base-handwritten checkpoint on the IAM Handwriting Database of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

13,353 annotated line images of 657 distinct writers, divided into 6,161 training, 976 validation, and 2,915 test samples using the AdamW optimiser with a learning rate of 5e-5, batch size of 8, weight decay of 0.01, linear warmup of 500 steps, and FP16 mixed- mixed-precision training over 10 epochs on an NVIDIA T4 GPU. Training loss converges from 1.418 at epoch 1 to 0.005 at epoch 10. Validation CER increases steadily (19.84), 13.09, 6.86 at epoch 1, 4, and 10 respectively which validates steady generalization and does not appear to be overfitting. The model has a final test set CER of 9.59% and WER of 21.35% withheld on held-out samples of IAM.

Stage 2 Stage 1 is again refined on the custom examination dataset in Section III-C, but with an even lower learning rate of 1e-5 and 30 epochs and a batch size of 4. The reduced learning rate avoids disastrous forgetting of features learned in Stage 1 but allows the model to contextualize itself to domain-specific academic vocabulary, subject terminology, and handwriting in undergraduate examination scripts.

During inference time, the entire answer sheet image is subjected to a consistent preprocessing pipeline and transcribed. Perspective deformation is fixed using Hough transform-based deskewing of the edges, which projects the angle of predominant lines, then uses affine rotation to align the text lines in a horizontal direction. To maximize the inks in uneven lighting, contrast is boosted with CLAHE with a clip limit of 2.0 and a tile grid of 8x8. The picture is doubled twice with bicubic interpolation to sharpen character detail and the picture is sharpened with a sharpening filter to sharpen stroke edges. A horizontal projection profile approach is used to split lines in the text: binary foreground pixels are added together row by row, a sliding-window average is applied to the result, and line boundaries are identified at depressions in the projection value that are less than 2 percent of the maximum projection value, with an 8-pixel padding on the edges of all the identified regions to retain ascenders and descenders. The individual segmented lines are then subjected to the TrOCR model with beam search decoding with beam width of 5 and with the maximum number of output tokens of 128 and the resulting transcription is then concatenated back to form the complete answer text.

D. Knowledge Graph Construction

To ground the evaluation in a structured, educator-defined content blueprint and prevent the hallucination inherent in purely generative grading, the system constructs a tailored Knowledge Graph (KG) for each exam question. The educator supplies a structured JSON rubric containing expected concepts, their relative importance weights, and their logical relationships. These inputs are mathematically modelled as a directed graph G = (V, E), where nodes V represent key concepts or entities expected in the answer each assigned a weight $w_i \in \mathbb{R}^+$ denoting its importance to the overall score and edges E represent logical or conceptual dependencies between nodes such as "causes", "depends on", and "produces". The resulting graph is constructed using the NetworkX library and acts as an immutable content blueprint, ensuring that all structural evaluation remains strictly grounded in the educator's curriculum regardless of the LLM's outputs.

E. Knowledge Graph Matching Algorithm

Once the student's answer is transcribed, the system compares it against the directed rubric graph using a three-step NLP pipeline. First, the transcribed text is processed using spaCy (en_core_web_sm) to extract noun chunks and named entities, forming a pool of student-generated concepts. Second, both the rubric graph nodes and the student concept pool are mapped into a shared high-dimensional vector space using the all-MiniLM-L6-v2 sentence transformer model. Cosine similarity is computed between each student concept vector and each rubric node vector; if the similarity exceeds the empirically determined threshold $\tau = 0.75$ (validated via ablation study in Section V-C), the rubric concept is flagged as successfully matched. Third, the deterministic KG coverage score is computed as:

$$S^{KG} = \Sigma_{i \in M} w_i / \Sigma_{j \in V} w_j \quad (1)$$

where $M \subseteq V$ is the set of matched rubric nodes. This formulation ensures that higher-weighted concepts contribute proportionally more to the structural score, directly reflecting the educator's intended marking scheme.

F. LLM Reasoning Module

Although the KG engine offers a strong test of concept presence and structural coverage, it tests the concepts individually and is less likely to identify instances when a student applies correct terminology to a scientifically incorrect or logically inconsistent statement. To acquire this subjective aspect of the quality of answers, the transcribed text and the teacher rubric are simultaneously submitted to a Large Language Model. The LLM is asked to consider two measures: Coherence logic, readability and structure of the answer and Conceptual Accuracy whether the statements presented by the student are factually accurate and contextually relevant to the subject matter. The LLM should give a formatted JSON response with both scores normalised to 0.0 to 1.0 as well as a succinctly formatted justification string. This organized output provides explicit explainability, where educators can review the reasoning of the AI and check the scores of borderline cases. SLLM, which stands as the mean of the coherence and the correctness scores, is calculated as the LLM score.

G. Hybrid Scoring Mechanism

The final grade is computed by mathematically blending the objective deterministic score from the KG engine with the subjective contextual score from the LLM reasoning module:

$$\textbf{\textit{Final Score}} = \alpha \times S^{KG} + (1 - \alpha) \times S^{LLM} \quad (2)$$

where S^{KG} is the Knowledge Graph coverage score from Equation 1, S^{LLM} is the mean LLM coherence and correctness score, and $\alpha \in [0, 1]$ is a tunable weight parameter set by the institution. A higher α places greater emphasis on structural rubric adherence, favoring answers that explicitly address all required concepts. A lower α rewards holistic conceptual understanding and communicative fluency, accommodating responses that demonstrate knowledge through varied expressions. Empirical evaluation across all subjects identifies $\alpha = 0.6$ as the optimal value, yielding the highest Pearson correlation with human graders ($r = 0.87$) and the lowest MAE (0.94) and RMSE (1.22).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

IV. IMPLEMENTATION

A. Software Stack

- HTR Engine: Python 3.10, HuggingFace Transformers (TrOCR), PyTorch with FP16 mixed-precision training on NVIDIA T4 GPU. OpenCV for image preprocessing and line segmentation.
- Backend Framework: FastAPI with async/await for asynchronous I/O. Pydantic for strict data validation and serialisation of the teacher's JSON rubric.
- Frontend Interface: HTML5, CSS3 (TailwindCSS), and vanilla JavaScript.
- LLM Engine: Google Gemini 1.5 Flash for subjective reasoning and justification generation.
- Knowledge Graph: NetworkX Python library for directed graph construction and traversal.
- NLP: spaCy (en_core_web_sm) for dependency parsing and noun chunking. Semantic vectorisation via HuggingFace sentence-transformers (all-MiniLM-L6-v2).

B. TrOCR Training Configuration

Hyperparameter	Stage 1 (IAM)	Stage 2 (Custom Exam)
Base model	trocr-base-handwritten	Stage 1 checkpoint
Optimiser	AdamW	AdamW
Learning rate	5e-5	1e-5
Epochs	10	30
Batch size	8	4
Warmup steps	500	20
Weight decay	0.01	0.01
Precision	FP16	FP16
Max token length	128	128
Training samples	6,161	~150

TABLE I: TROCR TWO-STAGE TRAINING HYPERPARAMETERS

C. Data Flow Summary

The operational pipeline executes as follows: (1) The educator uploads a JSON rubric containing weighted concepts and edges alongside the student's scanned answer sheet. (2) The backend preprocesses the image (deskew, CLAHE, upscale) and segments it into individual text lines using horizontal projection. (3) Each line is passed through the fine-tuned TrOCR model with beam search decoding to produce the full transcribed answer. (4) spaCy extracts student concepts and Sentence Transformers matches them against the NetworkX KG graph, yielding S^{KG}. (5) Concurrently, the transcribed text and rubric are passed to Gemini LLM, generating a JSON-formatted coherence and correctness score. (6) The Scoring Service calculates the final grade using Equation 2. (7) The result is delivered to the frontend dashboard with a transparent breakdown of matched/missing concepts and the LLM's justification.

V. RESULTS

A. Dataset

Experiments were conducted on a dataset of 150 handwritten answer scripts collected from undergraduate students. Scripts were written by 40 unique students under standard exam conditions. Handwriting quality was categorized as clear (42%), average (38%), and poor (20%) by two independent annotators. All scripts were evaluated independently by two human expert graders, whose scores were averaged to form the truth. Inter-annotator agreement between the two graders yielded a Cohen's Kappa of $\kappa = 0.81$, indicating strong agreement.

B. HTR Performance: TrOCR

Epoch	Training Loss	Validation Loss	CER
1	1.4183	1.2439	0.1984 (19.84%)
2	0.7479	0.8576	0.1380 (13.80%)
3	0.4200	0.8984	0.1499 (14.99%)
4	0.2863	0.8147	0.1309 (13.09%)
5	0.1813	0.7516	0.1030 (10.30%)
6	0.0951	0.7526	0.1098 (10.98%)
7	0.0446	0.6609	0.0842 (8.42%)
8	0.0282	0.6099	0.0766 (7.66%)
9	0.0151	0.5604	0.0707 (7.07%)
10	0.0055	0.5434	0.0686 (6.86%)

Table II TrOCR Fine-Tuning Training Progress on IAM Handwriting Database.

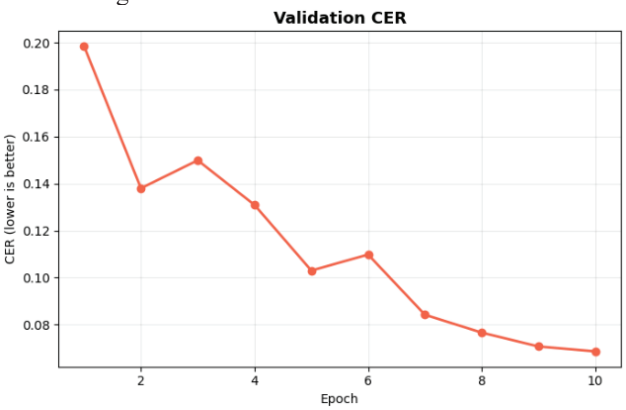

FIG II: Graph of training the model with 10 epoch.

Table II displays the epoch-by-epoch training history of TrOCR model trained with IAM Handwriting Database and the 10 epochs. The loss rate is shown to be steadily decreasing at an initial rate of 1.4183 at epoch 1 but the loss rates decrease to 0.0055 at epoch 10 at a rate of about 99.6 with no stagnation in the loss rate. The same applies to validation loss, which decreases progressively between 1.2439 and 0.5434, with a small fluctuation at epoch 3 (0.8984), which is another typical factor of transformer fine-tuning on medium-scale datasets. The main measure of evaluation, Character Error Rate (CER), increases significantly in all epochs: initially, at epoch 1, beginning at 19.84% at epoch 1, crossing below 10%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

at epoch 5 (10.30%), and reaching a final value of 6.86% at epoch 10. The stability of both the validation loss and CER reduction over the epochs confirm that the model is well-generalised to previously unknown samples of handwritings and that the data used is not overfitted given the rather small size of the dataset. One last test set CER of 9.59 and WER of 21.35 on held-out IAM samples also confirm the recognition ability of the model, which is within the acceptable range in the downstream automated grading tasks based on available HTR benchmarks.

C. *Cosine Similarity Threshold Ablation*

To empirically justify the cosine similarity threshold τ used in the KG matching step, concept-matching precision and recall were evaluated across a range of threshold values on a held-out set of 30 scripts.

Threshold (τ)	Precision	Recall	F1-Score
0.60	0.71	0.91	0.80
0.65	0.76	0.88	0.82
0.70	0.81	0.85	0.83
0.75	0.86	0.82	0.84
0.80	0.90	0.74	0.81
0.85	0.93	0.65	0.77

TABLE III: KG Concept Matching Performance vs. Cosine Similarity Threshold.

The threshold τ = 0.75 maximises the F1-Score, balancing precision (avoiding false concept matches) and recall (capturing semantically equivalent terms). Lower thresholds increase recall but admit noisy matches; higher thresholds improve precision at the cost of penalising valid synonymous expressions.

D. ***Hybrid Scoring vs. Baselines***

Table IV presents the correlation between system-assigned scores and the ground-truth human grades (Pearson r), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) across all 150 scripts.

System	Pearson r	MAE	RMSE
Keyword Matching	0.61	1.82	2.31
LLM Only (Gemini)	0.74	1.41	1.79
KG Only	0.71	1.53	1.94
TrOCR (IAM only) + Hybrid	0.82	1.12	1.47
Proposed Hybrid (α = 0.6)	0.87	0.94	1.22

TABLE IV: Grading Performance Comparison Across Systems.

The proposed hybrid system with domain fine-tuned TrOCR consistently outperforms all individual baselines. Notably, the intermediate row (TrOCR IAM-only + Hybrid) demonstrates that even without custom fine-tuning, the TrOCR-based pipeline achieves competitive grading correlation (r = 0.82), with the domain fine-tuning step contributing a further 0.05 improvement in Pearson r. This validates that the two-stage training strategy contributes meaningfully to end-to-end system performance beyond the HTR stage.

E. *Effect of α on Grading Correlation*

α	Biology	Comp. Networks	Env. Sci.	Overall
0.2	0.79	0.75	0.81	0.78
0.4	0.83	0.81	0.84	0.83
0.6	0.88	0.87	0.86	0.87
0.8	0.84	0.83	0.82	0.83
1.0	0.72	0.71	0.70	0.71

TABLE V: Effect of α on System-Human Grade Correlation

The optimal α = 0.6 is consistent across all three subject domains, suggesting that placing slightly more weight on KG structural coverage than LLM reasoning is broadly beneficial for descriptive exam grading.

F. *Explainability and Bias Reduction*

By segregating structural coverage into a mathematical graph score and mandating a short justification from the LLM, the system achieved a high degree of explainability. Educators could view exactly which concepts the KG Engine flagged as missing, while reading the LLM's rationale for deducting coherence points. Qualitative feedback from three faculty members who piloted the system indicated that the justification strings were consistently meaningful and helped them verify the AI's reasoning when reviewing borderline cases.

VI. LIMITATIONS

Although the given framework has a high level of empirical performance, various limitations should be mentioned.

Custom Dataset Scale: The domain fine-tuning dataset of about 200 annotated line images is small and is only enough to show improvement. On non-fine-tuning corpus handwriting styles, performance could deteriorate. It is also advisable to expand the dataset of 500 or more annotated lines in different student groups in order to deploy the production.

Rubric Construction Burden: The system expects the educators to give structured, weighted JSON rubric and KG edge definitions in the system. Such a preliminary task can be non-trivial in the case of open ended or highly creative answer areas where concept boundaries are uncertain.

Language Scope: The existing product can accept English handwriting. The all-MiniLM-L6-v2 and the encoreweb_sm spaCy model can perform poorly on multilingual or code-switched text.

Residual LLM Dependency: Although the HTR layer is no longer online, the Gemini API is still used by the LLM reasoning module, keeping the subjective scoring part in partial cloud dependency.

Size of the Dataset: Our sample set of 150 scripts is small although representative. To be able to fully determine generalizability, validation on a more diverse, large-scale corpus is required.

VII. CONCLUSION AND FUTURE WORK

This paper proposes an automated grading system of handwritten answers to examination, integrating the offline handwriting recognition with the structured rubric checks and the language model reasoning. Based on TrOCR, the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

handwriting recognition component is sequentially trained on the IAM Handwriting Database and a custom-designed dataset of 150 undergraduate examination script with a validation CER of 6.86% and test CER of 9.59%.

The grading runs in two directions. Knowledge Graph engine compares the answers of the students with a weighted rubric graph and delivers a deterministic coverage score that is based on the criteria set by the educator. The LLM module can individually evaluate the same response in terms of conceptual accuracy and coherence, giving normalized scores and a human-readable rationale in order to allow educator auditability. The combination of these outputs occurs via a tuneable linear combination, where a = 0.6 was found to be the most optimal in all the three domains of subjects. Compared to 150 human-graded scripts, the system has a r of 0.87 with expert graders, which is higher than the values of LLM only (r = 0.74), KG only (r = 0.71), and IAM only TrOCR (r = 0.82) models without domain fine-tuning, indicating that both the structural KG layer and domain-adapted HTR play an independent role in grading.

Future work will be to scale the examination data further to 250-500 scripts, semi-automatic generation of KGs using educator-provided model answers, replace the Gemini API with a local model, e.g., LLaMA 3 or Mistral, and enable multilingual support and batch processing to deploy it in a classroom.

REFERENCES

[1] Hiremath, Aditya & Irabatti, Nipun & Desai, Akhilesh & Dhondge, Prabhuraj & Sheikh, Shagufta. (2024). Transforming Handwritten Answer Assessment: A Multi-Modal Approach Combining Text Detection, Handwriting Recognition, and Language Models. 10.21203/rs.3.rs-4301899/v1.

[2] G. Sanuvala and S. S. Fatima, "A Study of Automated Evaluation of Student's Examination Paper using Machine Learning Techniques," 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2021, pp. 1049-1054, doi: 10.1109/ICCCIS51004.2021.9397227.

[3] F. E. Tania-Amanda Nkoyo, C. F. Ijezue, M. Amjad, A. I. Amjad, S. Butt, and G. Castañeda-Garza, "Advances in auto-grading with large language models: A cross-disciplinary survey," Year.

[4] Naman Agnihotri; Harshvardhan Grandhi; Dhanashri Patil; Sanika Kharade. "AI-Powered Exam Assessment System for Handwritten Answer Sheets." Volume. 10 Issue.3, March-2025 International Journal of Innovative Science and Research Technology (IJISRT), 3094-3097, https://doi.org/10.38124/ijisrt/25mar1924.

[5] Kamble, R., Halder, H., Mishra, N. et al. Multi-Modal Automated Evaluation of Handwritten Answer Sheets: A Framework Integrating SBERT and CLIP Encoding. SN COMPUT. SCI. 6, 787 (2025). https://doi.org/10.1007/s42979-025-04335-0.

[6] E. Shaikh, I. Mohiuddin, A. Manzoor, G. Latif and N. Mohammad, "Automated Grading for Handwritten Answer Sheets using Convolutional Neural Networks," 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), Amman, Jordan, 2019, pp. 1-6, doi: 10.1109/ICTCS.2019.8923092.

[7] S. Singh, Y. Shah, Y. Vajani and S. Dholay, "Automated Paper Evaluation System for Subjective Handwritten Answers," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579912.

[8] M. Kaya and I. Cicekli, "A Hybrid Approach for Automated Short Answer Grading," in IEEE Access, vol. 12, pp. 96332-96341, 2024, doi: 10.1109/ACCESS.2024.3420890.

[9] Faseeh, M.; Jaleel, A.; Iqbal, N.; Ghani, A.; Abdusalomov, A.; Mehmood, A.; Cho, Y.-I. Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy. Mathematics 2024, 12, 3416. https://doi.org/10.3390/math12213416.

[10] Anghel, C.; Anghel, A.A.; Pecheanu, E.; Cocu, A.; Craciun, M.V.; Iacobescu, P.; Balau, A.S.; Andrei, C.A. GraderAssist: A Graph-Based Multi-LLM Framework for Transparent and Reproducible Automated Evaluation. Informatics 2025, 12, 123. https://doi.org/10.3390/informatics12040123.

[11] Daina Teranishi, Masahiro Araki, Automatic Short Answer Grading to Pedagogical Questions Using Knowledge Graphs and Pre-trained Models, Transactions of the Japanese Society for Artificial Intelligence, 2022, Volume 37, Issue 4, Pages B-LC2_1-15, Released on J-STAGE July 01, 2022, Online ISSN 1346-8030, Print ISSN 1346-0714, https://doi.org/10.1527/tjsai.37-4_B-LC2.