LEVERAGING LLMs FOR AUTOMATED GRADING OF ANSWER SCRIPTS IN COMPUTER SCIENCE

Aneesh Ashwinikumar Sathe 21BCE1625
Dr. Janaki Meena
School of Computer Science and Engineering

MOTIVATION / INTRODUCTION

Grading handwritten answer scripts is a repetitive, time-intensive task that demands interpreting varied handwriting, diagrams, and pseudocode—often leading to fatigue, inconsistency, and human error. This research introduces an automated grading system that integrates Vision-Language Models (VLMs) and Large Language Models (LLMs) to evaluate answer scripts related to Computer Science subjects. By combining Fine-tuned Qwen2.5-VL-3B-Instruct's visual interpretation for recognizing and extracting text from handwriting, with Fine-tuned Google Gemma-3-4B-Instruct's domain expertise in the subject, the system ensures faster feedback, consistent scoring, reduced faculty workload, and scalability for large cohorts. This multimodal approach interprets handwritten content, diagrams, and code while providing detailed, context-aware feedback for efficient and intelligent evaluations.

OBJECTIVES

The core objective is to develop an LLM based evaluation system that can intelligently assess and grade scanned answer scripts.

1. **Improve Grading Speed**: Automate evaluation of handwritten scripts, diagrams, and pseudocode for faster feedback.

2. **Ensure Consistency**: Standardize grading to eliminate human bias and ensure fair scoring across all submissions.

3. **Reduce Faculty Effort:** Automate repetitive tasks to lessen the grading burden on educators.

4. **Enable Scalability**: Handle large-scale assessments like MOOCs and exams efficiently with accurate results.

SCOPE OF THE PROJECT

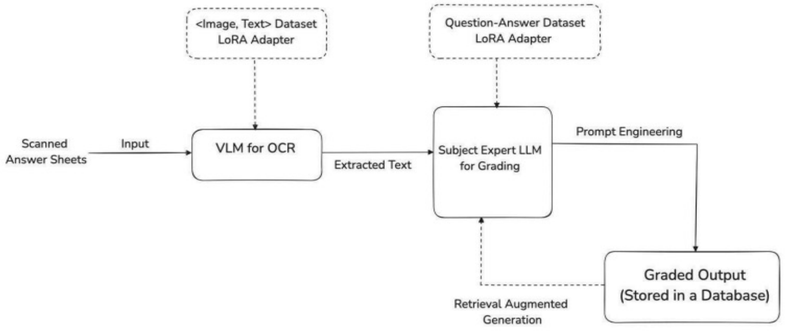
The scope focuses on automating the grading process to minimize manual effort, standardizing evaluation to ensure fairness, and enabling scalable deployment across various educational platforms such as universities, online courses, and competitive programming assessments.

METHODOLOGY

The methodology employs Low-Rank Adaptation (LoRA) to fine-tune Qwen2.5-VL-3B-Instruct on a custom <image, text> dataset for accurately extracting handwritten text and generating a descriptions of diagrams. Similarly, Gemma-3-4B-Instruct is fine-tuned on Question-Answer pairs related to a niche subject like Data Structures and Algorithms for introducing domain expertise. LoRA introduces low-rank matrices to reduce trainable parameters, preserving base model knowledge while ensuring computational efficiency. Qwen handles OCR and visual parsing, while Gemma evaluates answers against structured rubrics. Prompt engineering steers both models for precise, context-aware assessment—minimizing overfitting and enabling scalable deployment across MOOCs and institutional workflows.

ARCHITECTURE

The scanned answer script is input to the system, where the VLM performs OCR to extract handwritten text and generate descriptions of any diagrams. These outputs are then processed by the domain-specific LLM, which evaluates the responses against predefined rubrics and returns structured, rubric-aligned feedback via prompt engineering.



<Image, Text> Dataset LoRA Adapter
Question-Answer Dataset LoRA Adapter
Scanned Answer Sheets
Input
VLM for OCR
Extracted Text
Subject Expert LLM for Grading
Prompt Engineering
Graded Output (Stored in a Database)
Retrieval Augmented Generation

RESULTS

The following graphs depict the loss during fine-tuning of the VLM and LLM models. Additionally, The graded outputs from the proposed system are compared against GPT-4o, a state-of-the-art proprietary model.

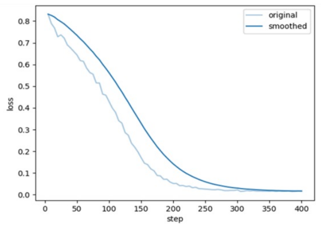


Fig 1. VLM Loss Graph

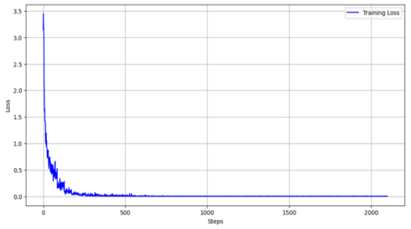
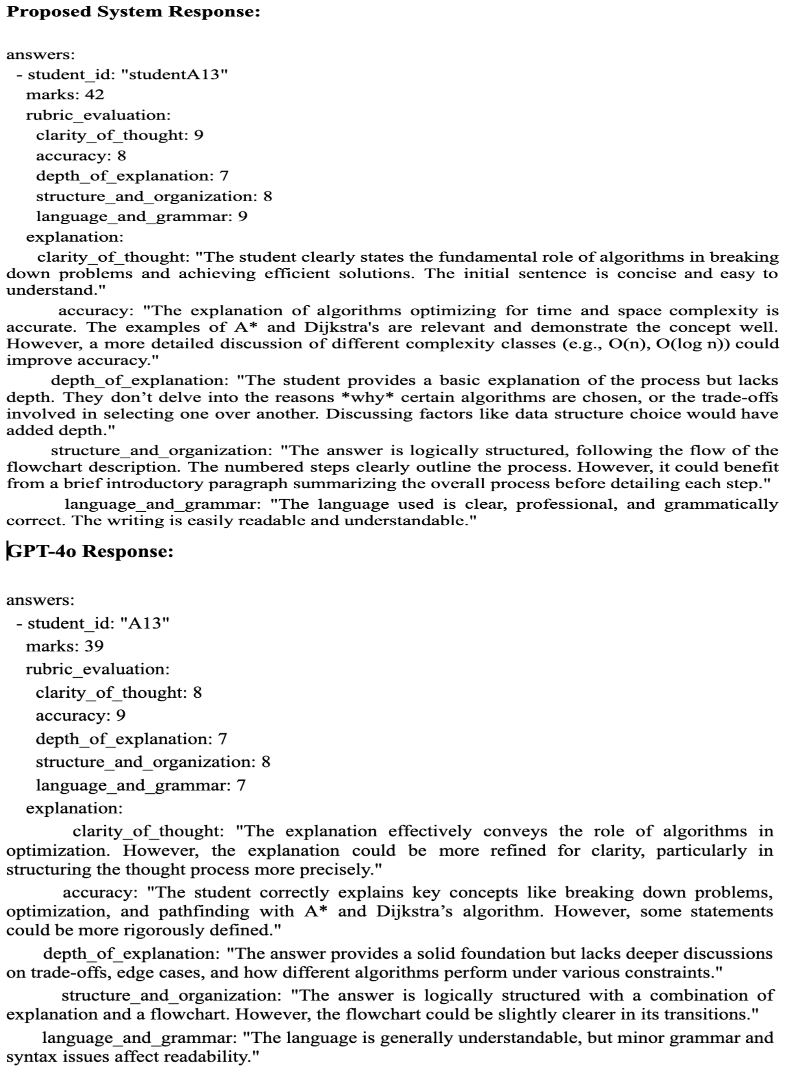


Fig 2. LLM Loss Graph



Proposed System Response:

```
answers:
  - student_id: "studentA13"
    marks: 42
    rubric_evaluation:
      clarity_of_thought: 9
      accuracy: 8
      depth_of_explanation: 7
      structure_and_organization: 8
      language_and_grammar: 9
    explanation:
      clarity_of_thought: "The student clearly states the fundamental role of algorithms in breaking down problems and achieving efficient solutions. The initial sentence is concise and easy to understand."
        accuracy: "The explanation of algorithms optimizing for time and space complexity is accurate. The examples of A* and Dijkstra's are relevant and demonstrate the concept well. However, a more detailed discussion of different complexity classes (e.g., O(n), O(log n)) could improve accuracy."
        depth_of_explanation: "The student provides a basic explanation of the process but lacks depth. They don't delve into the reasons *why* certain algorithms are chosen, or the trade-offs involved in selecting one over another. Discussing factors like data structure choice would have added depth."
        structure_and_organization: "The answer is logically structured, following the flow of the flowchart description. The numbered steps clearly outline the process. However, it could benefit from a brief introductory paragraph summarizing the overall process before detailing each step."
          language_and_grammar: "The language used is clear, professional, and grammatically correct. The writing is easily readable and understandable."
```

GPT-4o Response:

```
answers:
  - student_id: "A13"
    marks: 39
    rubric_evaluation:
      clarity_of_thought: 8
      accuracy: 9
      depth_of_explanation: 7
      structure_and_organization: 8
      language_and_grammar: 7
    explanation:
          clarity_of_thought: "The explanation effectively conveys the role of algorithms in optimization. However, the explanation could be more refined for clarity, particularly in structuring the thought process more precisely."
        accuracy: "The student correctly explains key concepts like breaking down problems, optimization, and pathfinding with A* and Dijkstra's algorithm. However, some statements could be more rigorously defined."
      depth_of_explanation: "The answer provides a solid foundation but lacks deeper discussions on trade-offs, edge cases, and how different algorithms perform under various constraints."
        structure_and_organization: "The answer is logically structured with a combination of explanation and a flowchart. However, the flowchart could be slightly clearer in its transitions."
      language_and_grammar: "The language is generally understandable, but minor grammar and syntax issues affect readability."
```

CONCLUSION

The proposed system, comprising a total of 7B parameters (Qwen2.5-VL-3B + Gemma-3-4B), achieves performance comparable to the state-of-the-art GPT-4o model, which has an estimated 175B parameters. This represents a ~96% reduction in model size, while maintaining grading accuracy within ±7% of GPT-4o across evaluated tasks.

CONTACT DETAILS

MAIL ID : aneeshashwinikumar.sathe2021@vitstudent.ac.in
MOBILE NO: 8779691940

REFERENCES

Hong, W., et al. (2024). CogVLM2: Visual Language Models for Image and Video Understanding. AI Research Papers.

Alhafni, B., Vajjala, S., Bannò, S., Maurya, K. K., & Kochmar, E. (2024). LLMs in Education: Novel Perspectives, Challenges, and Opportunities. Education Research Journal.