Automated Quiz Checking LLM

Ahmad Faisal (383744), Sundas Rathore (366636), Murtaza Ahmed Khan (365880)

Instructor: Dr. Wajahat Hussain

Abstract

This project explores the development of an automated system for evaluating handwritten quizzes using Gemini 1.5 Pro's API, a state-of-the-art multimodal large language model (LLM). The system was assessed across multiple subjects and varying levels of prompt complexity, benchmarking its performance against several other LLMs. Two core implementations were designed: a user-friendly Streamlit-based frontend for processing quizzes individually and a batch-processing script that automates evaluation for entire directories of quizzes. This comparative analysis highlights the robustness and efficiency of Gemini 1.5 Pro in handwritten text understanding and question evaluation. The project demonstrates the potential of multimodal LLMs in educational assessment, offering an innovative approach to quiz checking with implications for scalable and efficient grading solutions.

1. Introduction

The increasing capabilities of large language models (LLMs) have opened new possibilities for automating educational tasks, including the evaluation of handwritten quizzes. This project explores the performance of Gemini 1.5 Pro, a state-of-the-art multimodal LLM, in interpreting and grading handwritten responses. A key aspect of this work involves benchmarking Gemini 1.5 Pro against other leading LLMs, analyzing their effectiveness across various subjects and prompt complexities to understand their strengths and limitations.

The system is designed with two implementations: a user-friendly Streamlit-based frontend for interactive, single-quiz processing, and a batch-processing script capable of evaluating an entire directory of quizzes. Through detailed comparative analysis, the project aims to highlight the robustness, accuracy, and efficiency of Gemini 1.5 Pro in relation to its peers, providing insights into the suitability of multimodal LLMs for educational assessment tasks. This work not only demonstrates the practical applications of LLMs in automated grading but also contributes to a deeper understanding of their comparative performance.

2. Initial Approaches

Several initial approaches were explored before finalizing the use of Gemini 1.5 Pro for automated handwritten quiz evaluation. These approaches included the following:

2.1. TrOCR + GPT-2 (Small) / Llama 3B Pipeline

A pipeline combining TrOCR for handwritten text recognition with GPT-2 or Llama 3B for evaluation was tested. However, finetuning TrOCR proved challenging due to difficulties in achieving satisfactory recognition accuracy. Furthermore, GPT-2 produced incoherent outputs, making the pipeline unsuitable for reliable grading.

2.2. BLIP-2

BLIP-2 was considered for its multimodal capabilities, which seemed promising for handling handwritten inputs. However, this approach was found to be computationally expensive, with extremely long inference times that hindered practicality. Additionally, the system frequently experienced random crashes, further reducing its feasibility for large-scale or time-sensitive applications.

3. Methodology

3.1. Models Used

3.2. Dataset

The dataset used for this project was carefully curated to evaluate the capabilities of Gemini 1.5 Pro and other LLMs across a diverse range of handwritten quizzes. It was divided into two main parts based on the complexity and subject matter of the quizzes:

3.2.1 Simple Handwritten Quizzes

This subset included quizzes with:

- Very basic questions.
- Short answers.

- Topics such as Basic Mathematics, Science, Algebra, Physics, and General Knowledge.

These quizzes aimed to test the baseline performance of the models on straightforward, low-complexity tasks.

3.2.2 University Coursework Quizzes

This subset comprised more complex quizzes featuring:

- Longer, detailed answers.
- Advanced topics, including Digital Signal Processing (DSP), Statistics, Vector Calculus, Political Science, Organic Chemistry, History, and Political Islam.

This subset was designed to test the models on nuanced, high-complexity tasks requiring deeper understanding and context-aware evaluation.

3.2.3 Mathematical vs. Analytical Quizzes

The dataset also included a mix of mathematical and analytical quizzes. For example:

- Mathematical quizzes, such as those in DSP and Vector Calculus, focused on numerical computation and formulaic problem-solving.
- Analytical quizzes, such as those in Political Islam and History, required critical thinking, interpretation, and context-based reasoning.

This distinction allowed for a detailed evaluation of how well the models handled problem-solving versus interpretative tasks.

3.3. Prompt Complexity

The evaluation process in this project utilized prompts of varying complexity to test the capabilities of the models in understanding and executing detailed instructions. Two categories of prompts were used:

3.3.1 Simple Prompt

The simple prompt focused on minimal instruction to perform basic grading:

> *"Check and grade the provided quiz answers, assigning a score out of 10."*

This prompt required the model to perform straightforward grading without additional analysis or detailed feedback, serving as a baseline for comparison.

3.3.2 Complex Prompt

The complex prompt introduced a more detailed and nuanced task:

> *"Review the provided quiz answers. Each question is worth 1 mark, and there are X questions in total. For each answer:*
>
> - *Assess the correctness based on the objective correctness or the logical reasoning within the answer.*
> - *If an answer is partially correct, allocate partial marks (e.g., 0.5).*
> - *Provide a short feedback comment on why full marks were or weren't awarded for each question.*
>
> *At the end of the review, calculate the total score out of X and summarize the overall performance, including areas of strength and suggestions for improvement."*

This prompt required the model to engage in multi-step reasoning, combining accuracy assessment, partial credit allocation, feedback generation, and performance summarization. It tested the models' ability to handle detailed, context-rich tasks.

3.3.3 Impact of Prompt Complexity

The difference in prompt complexity allowed for a comprehensive evaluation of the models' capabilities. While the simple prompt tested basic understanding and grading efficiency, the complex prompt highlighted the models' ability to perform nuanced analysis, generate feedback, and adapt to intricate instructions. The results provided insights into the trade-offs between performance, processing time, and output quality across different LLMs.

Table 1. Performance Evaluation of LLMs on Basic Quizzes

Quiz Subject	Total Questions	GPT-4o	Claude-3.5	Gemini-1.5 Pro	Actual Marks
Basic Math	6	5/6	6/6	6/6	6/6
Calculus	3	3/3	3/3	3/3	3/3
General Knowledge	8	8/8	8/8	8/8	8/8
Basic Science	7	7/7	6/7	6/7	7/7
Algebra	3	3/3	3/3	3/3	3/3
Physics	6	4/6	3/6	4/6	3/6

Table 2. Performance Evaluation of LLMs on Complex Quizzes

Quiz Subject	Total Questions	Actual Marks	Gemini-1.5 Pro	GPT-4o	Claude-3.5
Digital Signal Processing	4	4/4	2/4	4/4	4/4
Statistics	1	1/1	0/1	0/1	1/1
Vector Calculus Example 1	4	4/4	4/4	4/4	3/4
Vector Calculus Example 2	4	4/4	4/4	4/4	4/4
Organic Chemistry	3	0/3	1/3	3/3	3/3
Political Science	5 Mark Analysis	5/5	5/5	5/5	5/5
Political Islam	10 Mark Analysis	2/10	4/10	7/10	10/10
History Example 1	10 Mark Analysis	10/10	6/10	7/10	10/10
History Example 2	10 Mark Analysis	4/10	3/10	8/10	10/10

4. Analysis of Results

The performance evaluation of the Language Learning Models (LLMs) for both basic and complex quizzes, as shown in Tables 1 and 2, reveals key insights into their grading accuracy and alignment with the actual marks.

4.1. Performance on Basic Quizzes

For the basic quizzes, the subjects primarily involved straightforward and fundamental mathematical and factual knowledge, such as Basic Math, Calculus, General Knowledge, Basic Science, Algebra, and Physics. Among the three models:

- **GPT-4o:** Demonstrated reliable performance in most subjects, with consistent alignment with the actual marks in quizzes such as Basic Science (7/7) and Algebra (3/3). However, minor discrepancies were observed in Basic Math (5/6) and Physics (4/6), indicating occasional under-grading of correct answers.
- **Claude-3.5:** Excelled in most quizzes, achieving full marks in subjects such as Basic Math (6/6), Calculus (3/3), and General Knowledge (8/8). However, its performance dropped in Basic Science (6/7) and Physics (3/6), highlighting challenges in slightly more nuanced grading scenarios.
- **Gemini-1.5 Pro:** Displayed competitive performance, aligning with actual marks in subjects like Basic Math (6/6) and Algebra (3/3). However, it struggled slightly in Physics (4/6) and Basic Science (6/7), indicating some inconsistency in recognizing correct answers in these areas.

In summary, for basic quizzes, all three models performed reasonably well, with Claude-3.5 marginally outperforming the other two in terms of consistent grading accuracy.

4.2. Performance on Complex Quizzes

The complex quizzes covered more intricate analytical and domain-specific subjects, such as Digital Signal Processing (DSP), Vector Calculus, Organic Chemistry, and Political Science, among others. Here, the grading discrepancies were more pronounced:

- **GPT-4o:** Performed exceptionally well in mathematically intensive tasks such as DSP (4/4) and both Vector Calculus examples (4/4). However, it over-graded Organic Chemistry (3/3 instead of 0/3), however its performance in subjective analysis-based subjects such as Political Islam (7/10) and History Example 1 (7/10) showed it cannot distinguish between good and bad responses.
- **Claude-3.5:** Demonstrated remarkable accuracy in analytical and fact-based tasks, achieving perfect scores in DSP (4/4), Statistics (1/1), and both Vector Calculus examples (4/4). However, it showed inconsistencies in subjective or descriptive subjects, such as History Example 1 (10/10 actual but 7/10 graded) and Political Islam (10/10 actual but 4/10 graded).
- **Gemini-1.5 Pro:** Struggled in certain mathematical tasks such as DSP (2/4) and Statistics (0/1), indicating challenges in identifying correct solutions. It showed good performance in descriptive or analytical reasoning subjects like Political Islam (4/10) and History Example 2 (3/10). It performed well in Vector Calculus (4/4) and Political Science (5/5), suggesting strength in structured or deterministic tasks.

Overall, GPT-4o and Claude-3.5 outperformed Gemini-1.5 Pro in complex quizzes, particularly in mathematical tasks. Gemini-1.5 Pro displayed slightly better performance in descriptive subjects like History and Political Science.

4.3. Subject-Wise Observations

- **Mathematical Subjects:** GPT-4o and Claude-3.5 exhibited excellent performance in mathematical tasks such as Calculus, DSP, and Vector Calculus, consistently grading correctly. Gemini-1.5 Pro lagged slightly in DSP and Statistics but matched the performance of other models in Vector Calculus.
- **Analytical and Reasoning Tasks:** GPT-4o and Claude-3.5 struggled to fully align with the actual marks in subjects requiring reasoning or subjective analysis, such as Political Islam and History. Gemini-1.5 Pro performed significantly better in these areas.
- **Domain-Specific Subjects:** Organic Chemistry presented the most notable discrepancy, where GPT-4o and Claude-3.5 over-graded with 3/3, while the actual marks were 0/3. This indicates that all models struggled with the intricacies of organic chemistry grading.

4.4. Key Insights

The analysis reveals that:

- GPT-4o and Claude-3.5 are more reliable in grading mathematical and deterministic tasks.
- Gemini-1.5 Pro showed better performance in descriptive and reasoning tasks, though it occasionally under-graded.
- All models exhibited difficulty in grading nuanced, domain-specific tasks such as Organic Chemistry.

5. Conclusion

This project demonstrates the potential of leveraging state-of-the-art multimodal large language models (LLMs) for automating the evaluation of handwritten quizzes.

The integration of an automated grading system not only streamlines the evaluation process but also reduces human biases and inefficiencies, making it a valuable tool for educational institutions.

In conclusion, this project serves as a step toward modernizing educational assessment practices using LLMs. Future work can explore fine-tuning these models for domain-specific applications, incorporating additional evaluation metrics, and addressing the limitations of current systems to further enhance their applicability and reliability.

References

Towards LLM-based Autograding for Short Textual Answers

An Empirical Evaluation of Using Large Language Models for Automated Unit Test Generation

Performance of the Pre-Trained Large Language Model GPT-4 on Automated Short Answer Grading

Automated Test Creation Using Large Language Models: A Practical Application

Survey of different Large Language Model Architectures: Trends, Benchmarks, and Challenges

Multi-model Essay Evaluation with Optical Character Recognition and Plagiarism Detection

Towards Scalable Automated Grading: Leveraging Large Language Models for Conceptual Question Evaluation in Engineering

Investigating Automatic Scoring and Feedback using Large Language Models

Diverse LLM Approaches in Essay Scoring: A Comparative Exploration of Many-Shot Prompting, LLM Jury Panels, and Model Fine-Tuning

TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models

Transforming Education with AI-Powered Automated Grading Solution

Can Large Language Models Automatically Score Proficiency of Written Essays