NATIONAL UNIVERSITY OF SCIENCES & TECHNOLOGY

Machine Learning (CS-471)

Initial Project Report

End-to-End Application for Handwritten Quiz Checking with LLM

Class: BEE 13C

Group Members

Name	CMS ID
Zuha Fatima	385647
Muhammad Anser Sohaib	367628
Zohair Shakeel	365721

Contents

1 Abstract

This project addresses the need for an automated grading system for handwritten quizzes, leveraging Large Language Models (LLMs) to provide consistent and rapid feedback on student responses. Our goal is to develop an end-to-end application that extracts text from handwritten answers, evaluates the content using LLMs, and delivers feedback in an organized, accessible format. This approach uses Optical Character Recognition (OCR) to transform handwritten responses into text and applies LLM-driven grading with subject-specific prompts to assess accuracy, completeness, and clarity. The web-based application is designed for ease of use, allowing educators to upload, grade, and retrieve results in a single workflow, ultimately reducing manual grading time and ensuring consistent feedback standards.

2 Methodology

2.1 Workflow:

- **Data Ingestion**: Handwritten quizzes are uploaded via the web interface and processed in batches to save time.
- **OCR and Text Extraction**: Files are processed through OCR to extract text from handwritten responses.
- **Grading Pipeline**: The extracted text is sent to the LLM grading model (Mistral-8B), which evaluates the responses based on criteria specified in a detailed grading prompt.
- **Result Display and Storage**: Graded responses are displayed in the app and stored in the configured format, with an optional feature to export results to Excel.

2.2 Current Progress:

- **OCR Integration**: We've tested multiple OCR models from Hugging Face, selecting one that provides acceptable results, though accuracy varies with handwriting quality. We are exploring ways to optimize text extraction for lower-quality inputs.
- **LLM Grading with Mistral-8B**: Mistral-8B, an advanced LLM, was integrated as the grading model. By using a tailored prompt, we achieve consistent grading aligned with subject matter criteria. This prompt helps standardize feedback on topics such as accuracy and relevance of student responses.
- **Flask-Based Web Application**: A preliminary Flask app provides a user-friendly interface, allowing file uploads and displaying both OCR and grading outputs. The app is designed for ease of use, with feedback mechanisms like progress indicators to improve user interaction.

- **Backend Setup and NVIDIA API**: The backend connects with Hugging Face for OCR and uses an API (NVIDIA) to process extracted text with the grading LLM, ensuring a robust and scalable infrastructure.

2.3 Future Integrations and Improvements:

1. Advanced OCR and Image Preprocessing:

Integrate preprocessing techniques, such as image normalization and noise reduction, to improve OCR reliability. We may explore combining multiple OCR models or using ensemble techniques.

2. Enhanced Prompt Engineering with LangChain:

Use LangChain to chain multiple LLMs for a more nuanced grading approach, particularly for quizzes with varied question types.

3. Improved User Interface and Accessibility

Add more interactive elements, such as progress bars and detailed error messages, to make the user interface more intuitive and accessible.

2.4 Limitations Of Our Application:

Our application, while functional, faces several limitations:

1. **OCR Accuracy with Poor Handwriting**: The OCR model struggles with inconsistent or poor-quality handwriting, leading to recognition errors that impact grading accuracy.
2. **Reliance on Prompt Precision**: Grading quality heavily depends on carefully crafted prompts, which can limit flexibility and consistency across varied subjects and question types.
3. **Limited Support for Diverse Quiz Formats**: The application currently performs best with simple question-answer formats, struggling with more complex structures like multi-part questions or multiple-choice.
4. **Performance and Scalability**: Processing multiple quizzes simultaneously can lead to latency due to high resource demand from OCR and LLM inference, impacting response times for larger submissions.
5. **Incomplete LangChain Integration**: Lacking LangChain, the application does not yet support model chaining, which would improve consistency and adaptability in complex grading workflows.

3 Prompt Generation:

Prompt generation isa major part of our project. We generated the following prompt and gave it to our llm Mistral -8B to use for checking and marking:

prompt_qa = f"""

You are provided with a set of questions and answers. Your task is to grade the answers with a focus on the following:

*1. *Subject Relevance*: For each subject (e.g., grammar, physics, etc.), the grading criteria will differ slightly:*

*- *For English or Grammar quizzes: Focus on **spelling, **grammar, and **sentence structure*. Make sure the answer is clear and well-formed.*

*- *For Science or Physics quizzes: Focus on **concept accuracy, **relevance of explanation, and whether the **fundamental principles* of the subject are properly addressed.*

*2. *Grading Criteria*:*

*- *Accuracy*: Does the answer correctly address the question, with factual or conceptual correctness?*

*- *Completeness*: Is the answer fully developed? Does it include all necessary components (e.g., calculations for physics, complete sentences for English)?*

*- *Clarity*: Is the answer clear, concise, and logically structured? Is the spelling and grammar correct for English quizzes? Are the scientific explanations easily understandable for physics quizzes?*

*3. *Scoring*:*

*- Provide a score out of *10* based on the quality of the response, considering the subject-specific guidelines above.*

*- Keep the feedback brief and focused. *Point out any mistakes* without overwhelming the response with text.*

- Provide the total score at the end.

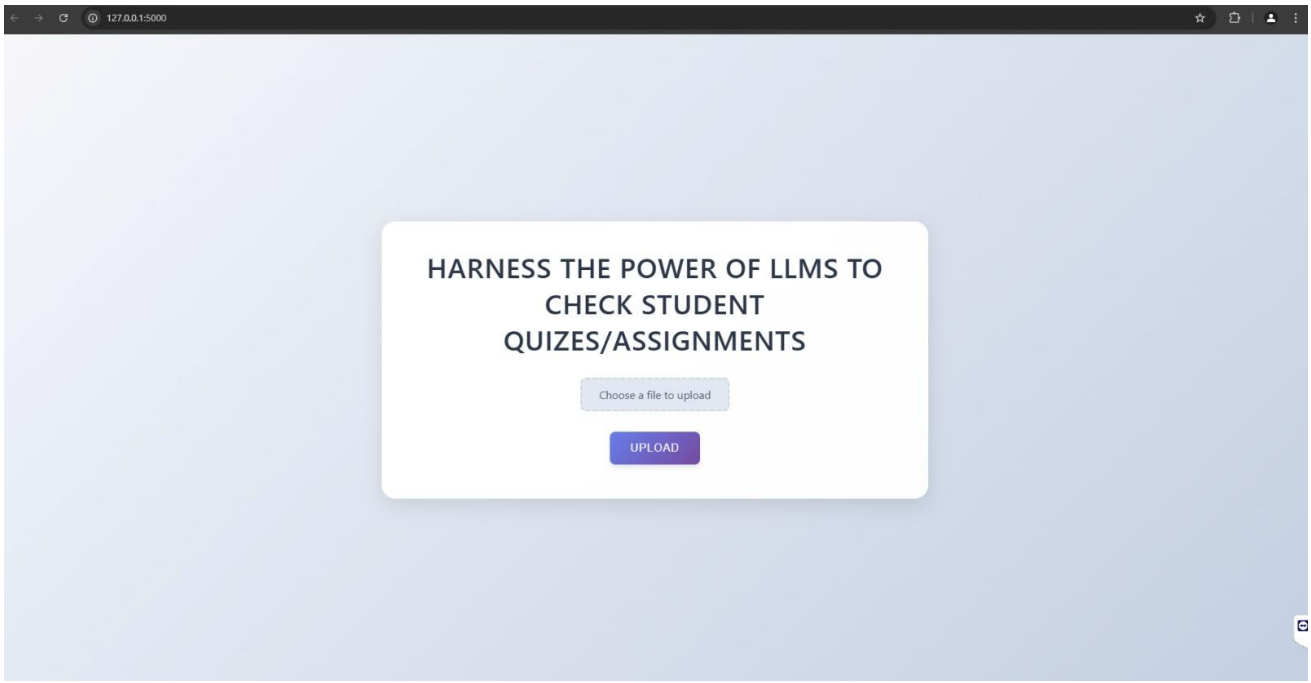
**Here is the text to be graded*:*

{res_text}

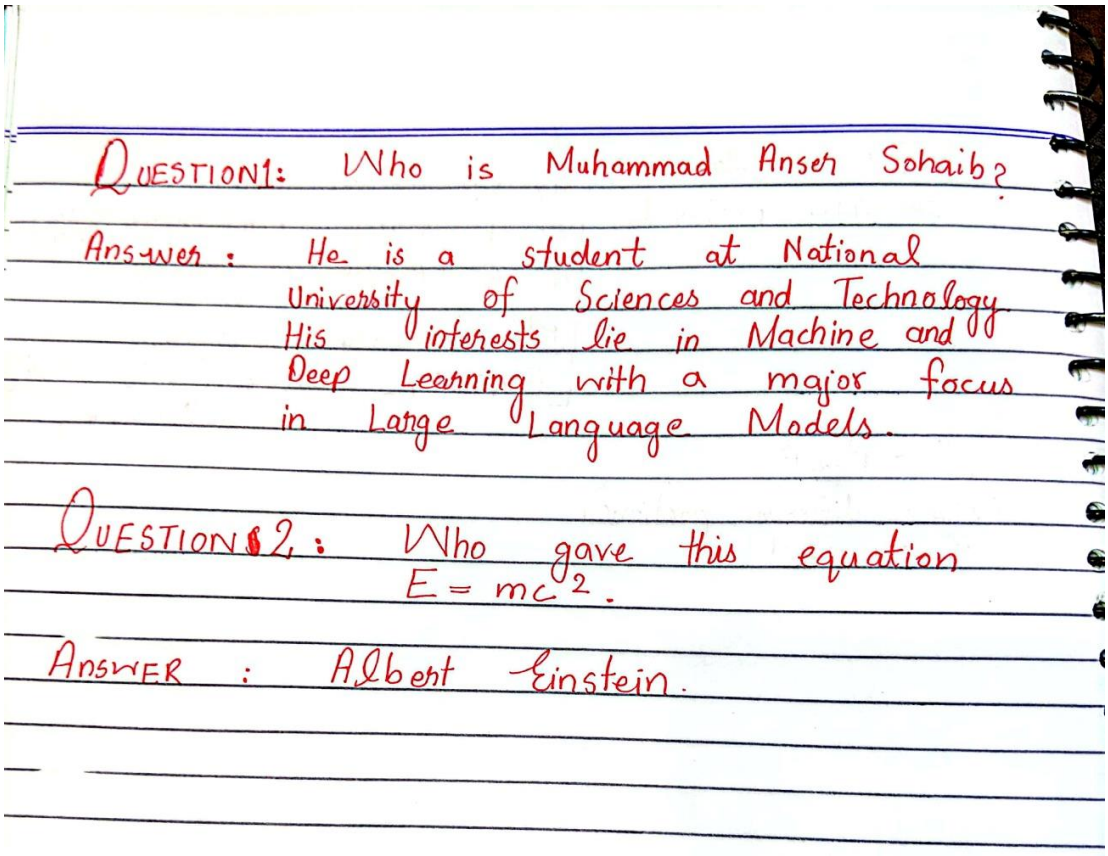
"""

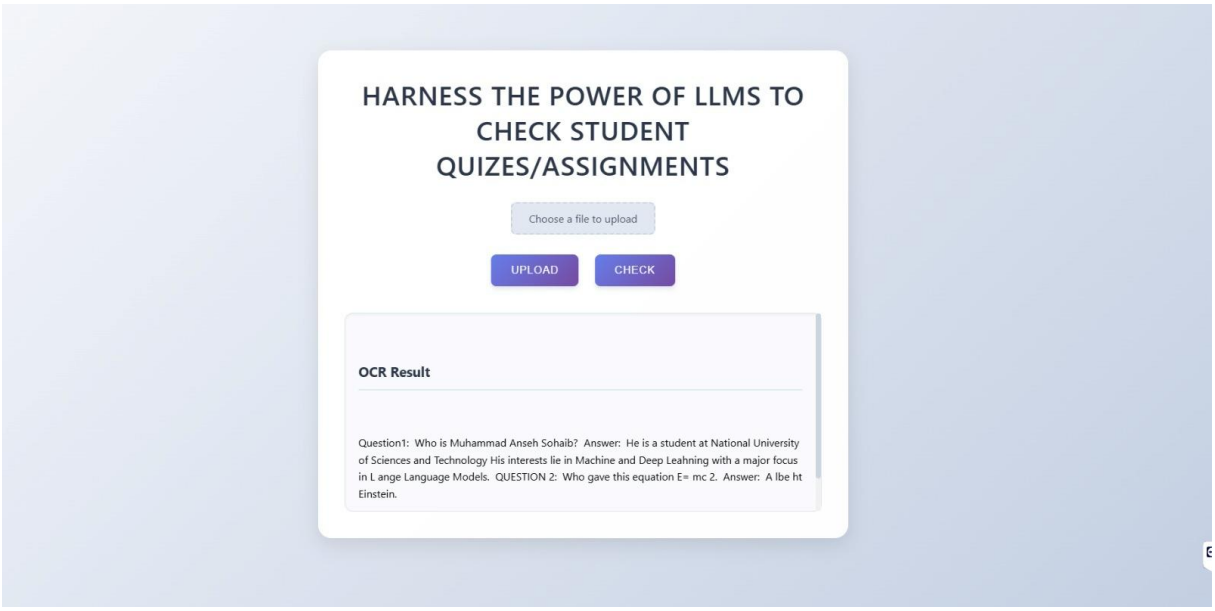
4 Output:

Web Application GUI:



Input:





Grading and Assessment:

Graded Text

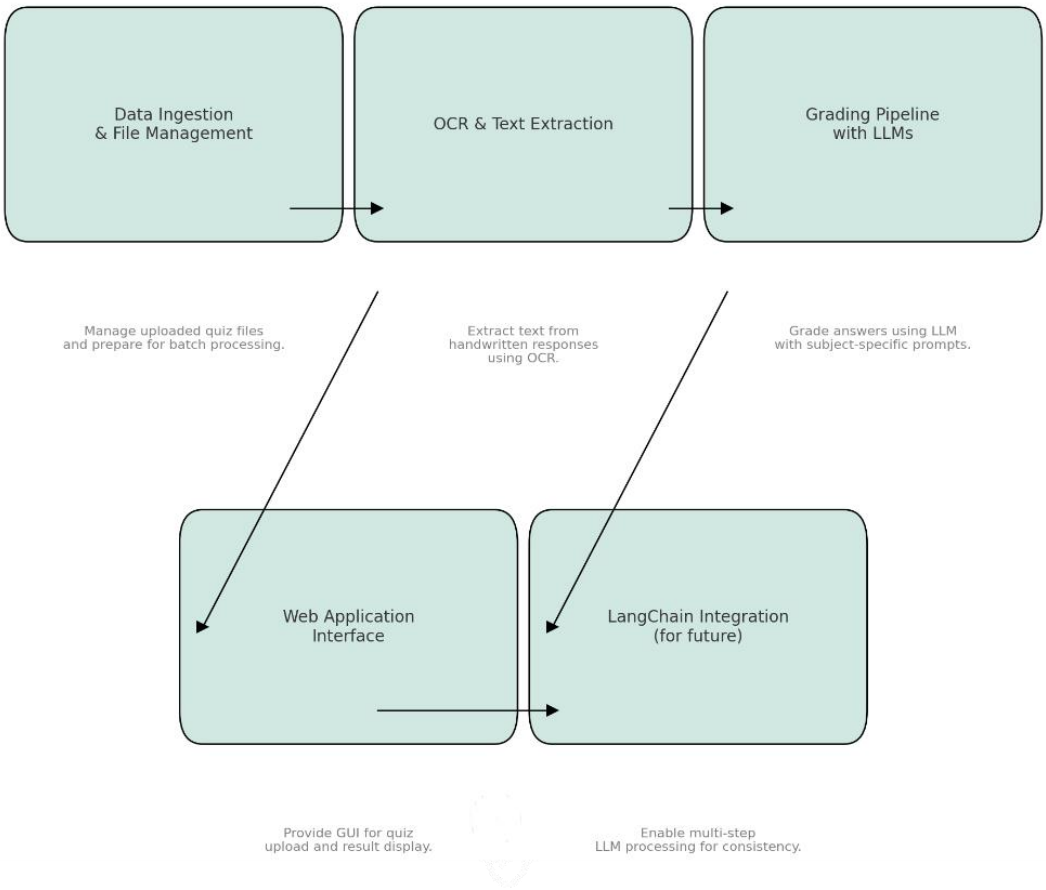
Question 1: The answer is relevant to the subject of personal information. The accuracy is high as it provides correct information about the person's educational background and interests. The answer is complete and clear. However, there are some spelling mistakes (Sohaib instead of Sohail, L ange Language Models instead of Language Models). Score: 8/10

Question 2: The answer is relevant to the subject of physics and the specific question. The accuracy is high as it correctly identifies Albert Einstein as the person who gave the equation E=mc^2. The answer is complete and clear. Score: 10/10

Total score: 18/20

(Note: The feedback was kept brief and focused, pointing out the spelling mistakes without overwhelming the response with text.)

5 Block Diagram



6 Conclusion

This project demonstrates significant potential to streamline handwritten quiz grading by leveraging state-of-the-art OCR and LLM technologies. The initial setup is functional, and further development will focus on enhancing robustness, improving grading accuracy, and refining the user experience. With these planned improvements, this application could become a valuable tool for educators, simplifying the grading process and freeing time for other instructional activities