Personalized Learning Assistant Experimentation Project Report

Gowreesh Gunupati, Katyaini Raj, Priyanka Gujar

Introduction

The primary objective of this project is to create an assisted learning system powered by large language models (LLMs) that simplifies the process of generating study materials. The platform allows users to upload text or PDF files, which are then processed to generate contextually relevant questions and summaries. This tool aims to enhance the learning experience by providing tailored educational content, fostering better comprehension and engagement.

The methodology utilizes two specialized models: GEMMA, a sequence-to-sequence model optimized for summarization tasks, and LLAMA 3.2, a decoder model designed for generating high-quality questions. The workflow integrates pre-processing steps, task-specific operations, and user input via an intuitive Streamlit-based interface. The attached flowchart illustrates the system architecture, highlighting the seamless interaction between components such as input processing, task selection, and model execution.

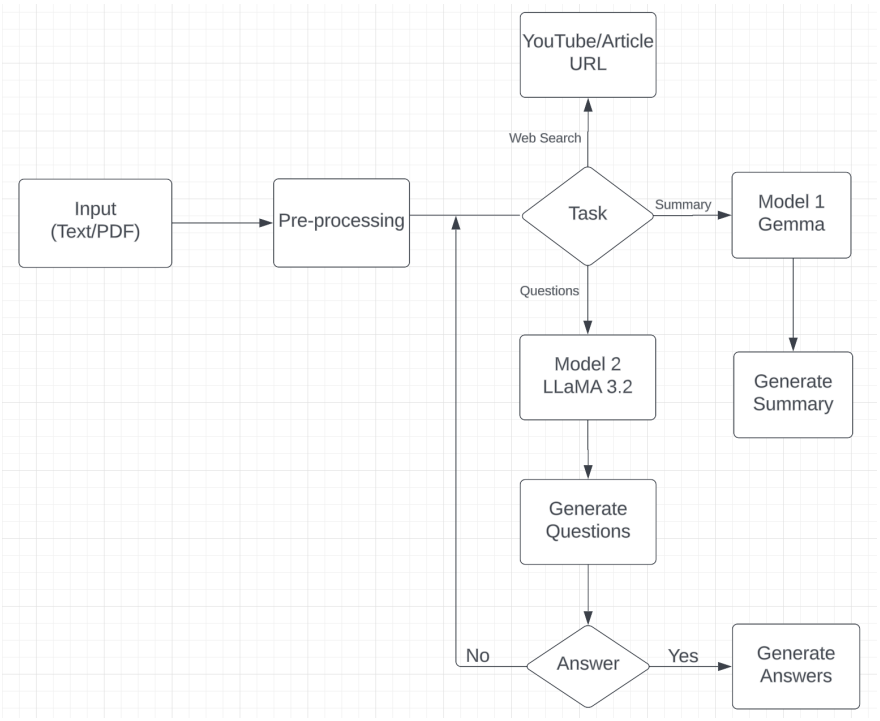

Figure 1: Application workflow

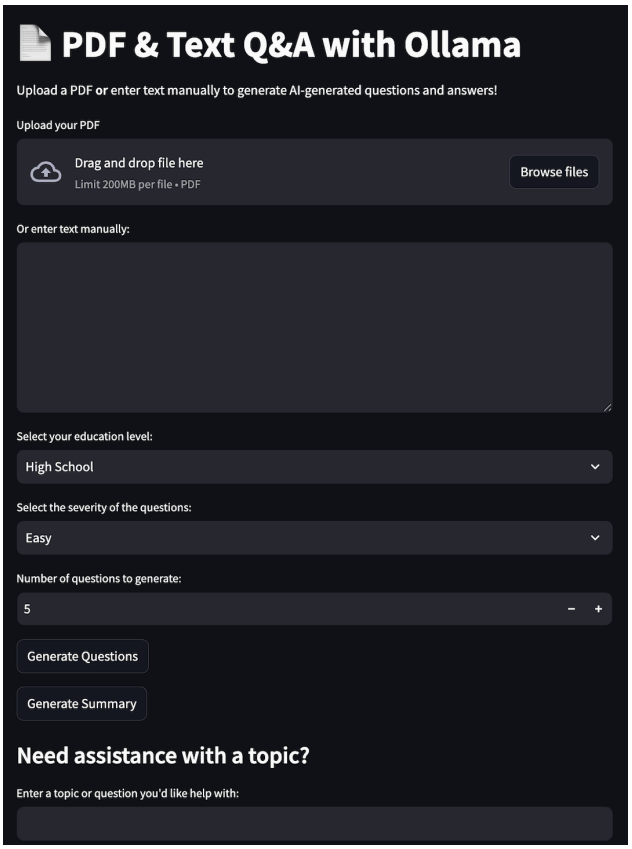

Figure 2: Application Interface

By leveraging cutting-edge LLMs and a user-friendly design, this project addresses the growing need for personalized educational tools in academia and beyond.

Objective

This project focuses on LLM-based tasks such as:

- Text summarization
- Question-answer generation
- Assisted learning applications

Significance

- Reduces the time and effort required to create study materials.
- Enhances accessibility to tailored educational content.

- Supports learners with diverse needs by providing customizable inputs.

The relevance of this project extends to academic research, industry applications (e.g., corporate training), and real-world impact in improving education systems globally.

Hypothesis

We aim to observe:

1. **Model Variability:** The suitability of each model for specific educational applications.
2. **Processing Efficiency:** On-device models will show comparable latency to cloud models while being more cost-effective and open-source
3. **Question Difficulty:** LLMs can generate questions of varying difficulty, with prompting techniques influencing model responses and interaction.

Experimental Setup and GitHub

Our application is available at github.

1. **Python Libraries:** pypdf, Pandas, NumPy, Matplotlib
2. **Software:** Ollama
3. **Frontend:** Streamlit
4. **Tools:** VSCode, Jupyter Notebook, Github

Methodology

The chosen methodology is designed to address the challenges of creating personalized study materials efficiently. By employing GEMMA for summarization and LLAMA 3.2 for question generation, the system ensures high-quality outputs tailored to the user's needs.

1. **Data Preprocessing**
    - Load and preprocess the given input using NLP techniques.
    - Extract text from uploaded PDFs using pypdf if necessary.
    - Tokenize the dataset to prepare inputs for LLMs.
2. **Context and QA Pipeline**
    - **Model 1 GEMMA:** Summarisation
        - **Input:** Pre-processed Text or a PDF file.
        - **Output:** Summary of the input
    - **Model 2 LLaMA 3.2:** Question-Answering Generation
        - **Input:** Pre-processed Text or a PDF file.

– **Output:** Precise answers and questions generated based on context.

3. **Evaluation**
   - Metrics: Use ROGUE score to determine how the model is performing for generating summaries.
   - Using human as a judge to see the quality of questions generated.
4. **Deployment**
   - Build a user-friendly interface using Streamlit for interactive input, knowledge retrieval, question and summary generation.

This methodology ensures that both summarization and question generation are optimized for diverse educational contexts.

Results

The figures below present a sample of summary and medium-level questions generated for a bachelor's student using a PDF on linear regression as input.

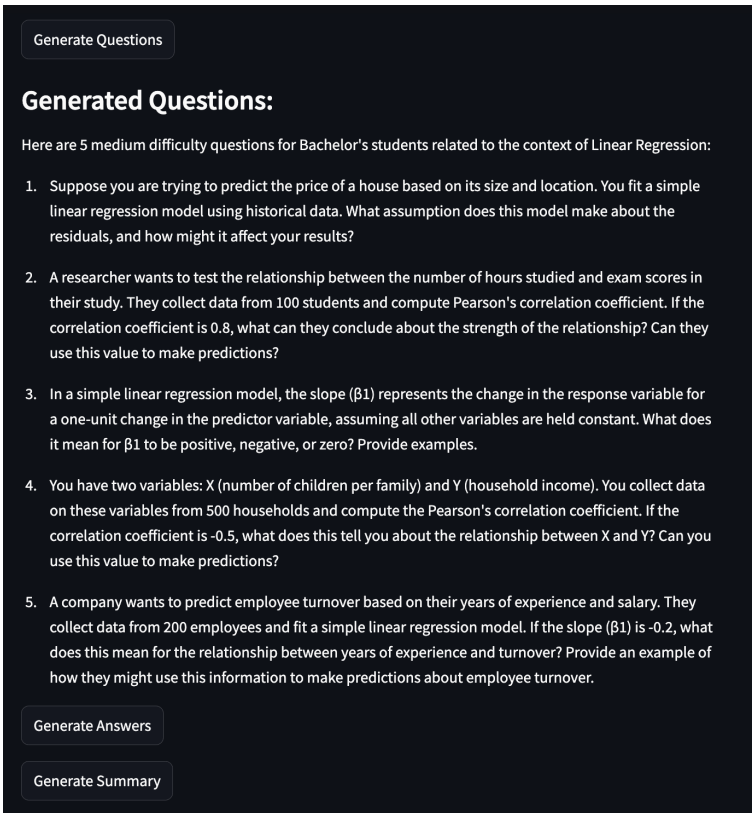

Figure 3: Sample of Questions Generated by the application

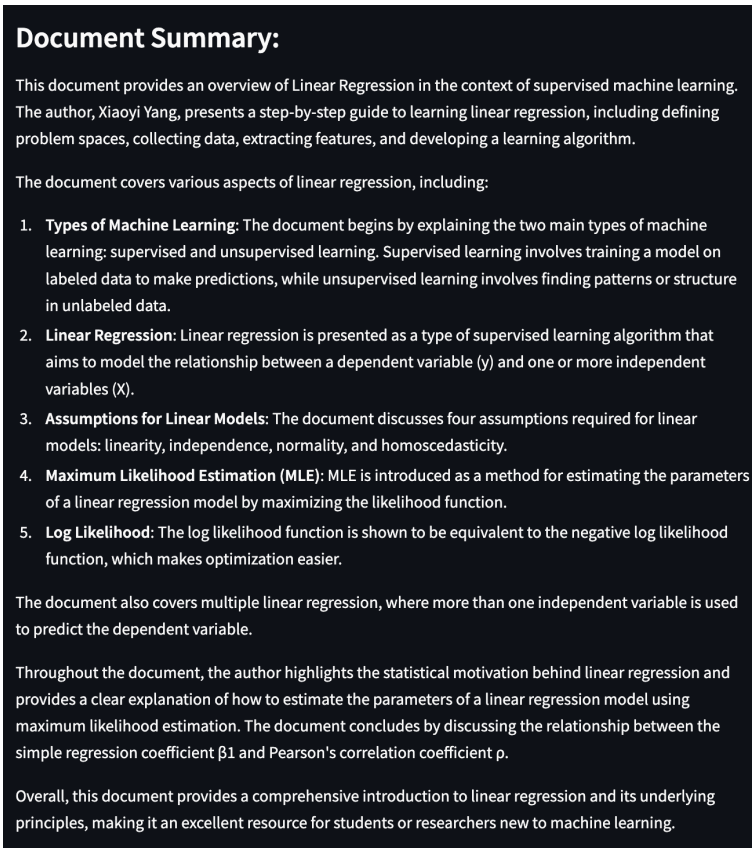

Figure 4: Sample of Summary Generated by the application

Evaluation and Comparison

Evaluation Metrics

To compare model performance:

Metric ↓ / Model →	GEMMA	T5
ROUGE-1 F1 SCORE	**0.59**	**0.04**
ROUGE-2 F1 SCORE	**0.28**	**0.03**
ROUGE-L F1 SCORE	**0.32**	**0.04**

Table 1: Summarization Metrics Comparison

- **Summarization Tasks:** ROUGE score was used to measure the quality of summaries generated by GEMMA and T5. As you can see, T5 is under performing when compared to GEMMA, hence we moved forward with GEMMA to perform the summarisation.

- **Question Generation Tasks:** We have used human as a judge to check the quality of questions and it is giving pretty good results so far.

Discussions

Model Consideration

For this project, transformer-based models were prioritized over traditional machine learning models due to their superior performance in natural language processing (NLP) tasks such as summarization and question generation. Specifically, the following models were considered:

- **GEMMA (Sequence-to-Sequence Model):** GEMMA was selected for summarization tasks due to its ability to condense large volumes of text while preserving contextual meaning. Its sequence-to-sequence architecture is well-suited for tasks requiring input-output mapping, such as summarization.
- **T5 (Sequence-to-Sequence Model):** T5 was considered for the same reasons as GEMMA due to its sequence-to-sequence architecture being well-suited for tasks like summarization.
- **LLAMA 3.2 (Decoder Model):** LLAMA 3.2 was chosen for question generation because decoder models excel at generating coherent and contextually relevant text. LLAMA's architecture supports fine-tuning for generating questions across different difficulty levels and educational contexts.

Final Model Selection

The final models were selected based on the following criteria:

1. **Task-Specific Performance:** GEMMA's summarization capabilities and LLAMA's proficiency in generating questions made them ideal candidates.
2. **Scalability:** Both models are scalable and can handle diverse input types (text or PDF).
3. **Ease of Fine-Tuning:** GEMMA and LLAMA offer pre-trained versions that can be fine-tuned on specific datasets to improve performance.

Transfer Learning

Both GEMMA and LLAMA leveraged pre-trained weights from publicly available datasets.

Future Work

1. **Multilingual Support with Hybrid On-Device Models:** To cater to a diverse user base, we plan to extend language model support beyond English leveraging hybrid on-device models that dynamically switch between languages as needed.
2. **AI-Powered Voice-Based Interactive Learning:** To make learning more accessible and engaging, we will integrate voice interaction using technologies like Whisper and Eleven Labs, enabling a natural and immersive learning experience.
3. **URL-based Knowledge Ingestion:** Instead of requiring PDF downloads, we aim to enhance the assistant's ability to directly process knowledge from web sources through integrated search and web scraping, making information access more seamless and efficient.

Conclusion

The project aimed at developing a Personalized Learning Assistant through the integration of on-device LLMs for quiz generation and summarization from PDFs. After thorough experimentation, several key findings emerged, which are significant both for the current project and the broader field of machine learning.

Key Findings

1. **Model Variability:** Different models exhibited varied responses to the same queries, with prompt design heavily impacting their performance. This finding highlights the importance of carefully choosing and fine-tuning models for specific tasks, as well as adjusting prompts to optimize outcomes. The experimentation revealed that no single model was universally superior, but some models worked better for summarization, while others excelled in question generation.

2. **On-Device Processing Efficiency:** The hypothesis that on-device models could match cloud-based models in terms of latency while being more cost-effective and privacy-conscious was confirmed. Models like LLaMA and GEMMA, when deployed locally, provided competitive performance, balancing efficiency and data privacy, especially in environments where connectivity may be limited.

3. **Question Difficulty Generation:** The models were able to generate quiz questions of varying difficulty. It was found that prompting techniques significantly influenced the model's ability to tailor question difficulty, suggesting that prompt engineering plays a crucial role in controlling output characteristics. This insight can be valuable for building adaptive learning tools that cater to individual learning levels.

Contribution to the field of ML

The integration of on-device models for personalized learning pushes the boundaries of privacy, accessibility, and cost-effectiveness in the educational technology space. By demonstrating that local models can function effectively in real-time learning scenarios, this project paves the way for more practical applications of LLMs in settings where privacy and offline capabilities are essential.

The findings on prompt engineering and model variability contribute to the growing body of knowledge on how to fine-tune models for specific tasks. This opens up new avenues for optimizing large language models in personalized learning systems and other specialized applications.

Reflection on the Experimentation Process

The experimentation process revealed the importance of model selection, prompt engineering, and system design in creating an effective learning assistant. Several challenges were faced, such as ensuring compatibility across platforms, maintaining low latency, and meeting varied user needs. However, through iterative development, these challenges were mitigated, and we learned valuable lessons about the complexities of deploying LLMs on-device.

One of the key challenges we faced during the experimentation process was the lack of a clear, standardized metric for evaluating the quality of generated quiz questions. Unlike summarization tasks, which have well-established evaluation metrics like ROUGE scores, assessing question quality is more subjective and multifaceted, involving factors such as relevance, clarity, difficulty, and alignment with learning objectives. Without a universal metric, we had to rely on human evaluation and heuristic rules, making it difficult to

objectively compare model performance and ensure consistent output quality across different models. This highlighted the need for further development of formalized metrics to assess question quality in personalized learning applications.

In the context of the overall project goals, the outcomes validate the hypothesis that personalized learning tools can be effectively built using on-device LLMs, paving the way for further development in this area. The team's approach to ensuring user-centric design and personalized learning paths was a success, as it contributed to building an application that caters to individual needs based on their educational level.