Chatbot for the examination office

Benjamin Brünau
HTWG Konstanz
Konstanz, Germany
be391bru@htwg-konstanz.de

Khadidja Kebaili
HTWG Konstanz
Konstanz, Germany
kh871keb@htwg-konstanz.de

Joel Merath
HTWG Konstanz
Konstanz, Germany
joel.merath.fh@gmail.com

Jan Schmidt
HTWG Konstanz
Konstanz, Germany
ja163sch@htwg-konstanz.de

***Abstract*—The aim of this team project is to develop a chatbot adapted to the requirements of the examination office (Prüfungsamt) to automatically answer frequently asked questions based on provided documents (e.g. studies and examination regulations). Given the increasing demand for quick and accurate information access in academic institutions, this research focuses on designing and implementing a chatbot to streamline administrative processes. The performance and accuracy of the chatbot will be evaluated using various metrics, to measure its reliability and effectiveness.**
These metrics will furthermore be used to measure how much of an improvement is achieved by finetuning a Large Language Model for this task and domain.

I. Introduction

In recent years, the use of artificial intelligence has gained significant attention in several areas. One of the emerging applications is the deployment of chatbots to handle routine inquiries, such as answering frequently asked questions of customers, employees, etc.
This paper focuses on developing a chatbot for the HTWG examination office that can automatically respond to frequently asked questions using existing documentation. Our aim is to improve the efficiency and accessibility of information for students and staff. The motivation behind this work lies in reducing the workload of administrative staff and providing timely, accurate information. Additionally, this research serves as a proof of concept, demonstrating the feasibility of implementing such a system in an academic administrative environment. The performance of the chatbot will be thoroughly evaluated using a variety of metrics to measure its effectiveness and reliability.

II. Design and Implementation

A. Document retrieval

Retrieving the correct document passage is crucial for ensuring accurate and relevant responses from the chatbot. Many search algorithms are using keywords as searching method. While these methods are effective for simple queries, they often fail to capture the contextual meaning, leading to suboptimal results.

To overcome these limitations, we used a semantic search method. This technique involves embedding both the queries and the documents into a high-dimensional space using a neural network, where semantically similar texts are represented by vectors in close proximity. To increase the chances of finding the correct information, multiple documents are retrieved. The count of how many documents are retrived is described with the K-nearest-neigbours (KNN) value. The retrieved documents are then passed to a Large Language Model (LLM) to generate the final response.

The document retrieval process requires several key steps. First, text extraction from various document formats such as PDF and DOCX is performed. We discovered that manual tuning of the extraction process—such as removing headers, footers, and combining the text into coherent passages—significantly improved the accuracy of the retrieval. The processed text is then divided into chunks using unique characters, ensuring that the passages are meaningfully segmented.

To optimize our retrieval system, we conducted extensive testing with various embedding models and configurations. The goal was to verify whether the model could correctly identify relevant sections within the text. This was assessed by defining unique keywords and ensuring all keywords were present within the retrieved passage. The performance of different models and settings was rigorously evaluated, and the results are detailed in the subsequent section III-A.

B. Evaluation

To evaluate the performance of the chatbot, a combination of custom and established metrics was employed. Three custom metrics were developed: the Keyword Score, which assesses the presence of essential keywords in the chatbot's responses; the Correct-Page Score, which determines whether the chatbot accurately identifies the correct source document; and the ChatGPT Score, which utilizes ChatGPT to further evaluate the quality of the responses.

In addition to these custom metrics, several established metrics were used to analyze the chatbot's performance on both semantic and textual levels. The semantic evaluation, which was given higher priority, is particularly critical for the chatbot's application within an examination office, where precise and contextually relevant responses are required. The Keyword Score plays a central role in ensuring that key terms are present in the responses. Cosine Similarity was employed to measure the semantic alignment between the generated responses and the reference answers, thereby verifying the accuracy and relevance of the content.

For the analysis at the word and text level, which was weighted less heavily, Edit Distance, ROUGE-1, ROUGE-2, ROUGE-L, and Jaccard Similarity were utilized. Although

these metrics were assigned a lower weight, their importance remains, as they assess the textual precision and consistency of the responses. These metrics ensure that the chatbot's outputs are not only semantically correct but also well-formulated and closely aligned with the reference texts. This is particularly significant in formal contexts, such as within an examination office, where the structural integrity of responses is critical.

Furthermore, Promptfoo was utilized to test the Retrieval-Augmented Generation (RAG) system with a diverse set of large language models (LLMs). A wide range of reference questions and answers were compared with the outputs generated by the LLMs, and the respective scores were calculated. This approach enabled the identification of LLMs that are most effective in conjunction with the RAG system and are likely to produce the most accurate responses. The average of the weighted metrics, including correctness, specificity, and relevance, serves as a key indicator in selecting the optimal model, providing a clear assessment of each model's performance.

Prompt engineering played a crucial role in refining the chatbot's interactions, ensuring that prompts were crafted to elicit the most accurate and relevant responses. The outcomes of these metrics were visualized using Promptfoo, providing a comprehensive overview of the chatbot's performance.

C. Finetuning

Two primary approaches were taken for finetuning an LLM:

- Teaching the LLM the content of documents by finetuning it on specific paragraphs.
- Enhancing the Retrieval Augmented Generation (RAG) process by training the model to extract relevant context and accurately answer questions based on it.

Given our computational constraints, we opted for parameter-efficient finetuning (PEFT) using LoRA (Low Rank Adaptation) instead of the more resource-intensive Full Fine Tuning. LoRA significantly reduces the number of parameters that need to be trained by approximating the weight update matrix with lower-rank matrices (as illustrated in Figure 1). The training process was executed using Huggingface Transformers and tracked via wandb. Datasets were synthetically created through API requests to *ChatGPT4o*.

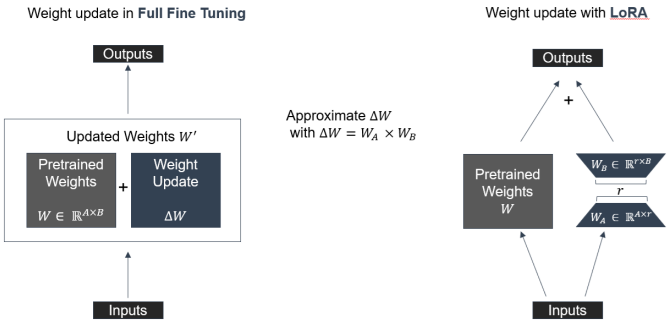

Fig. 1. Weight update: Full Fine Tuning vs LoRA

III. Results

A. Retrieval

The effectiveness of our document retrieval process was evaluated using several metrics, as detailed in Table I. We tested multiple embedding models, including all-MiniLM-L6-v2 (Model A), all-mpnet-base-v2 (Model B), and paraphrase-multilingual-MiniLM-L12-v2 (Model C). The models were evaluated across different context sizes (1024, 2048, and 4096) and varying KNN settings (1, 3, 5, and 10). As demonstrated in Table I, Models B and C exhibited the most optimal performance outcomes. Given that Model C yielded superior outcomes solely in consideration of the keywords, irrespective of the passage, in comparison to Model B, this model was selected for the subsequent tests. Furthermore, the chunk size of 2048 exhibited slightly enhanced performance metrics in comparison to 1024 or 4096. The optimal chunk size for text segmentation was determined to be 2048 tokens, with a chunk overlap of 256 tokens. All processed text chunks were stored in a ChromaDB, an open-source vector database, thereby facilitating efficient retrieval during chatbot operation.

Model	Knn	Context size 1024	Context size 2048	Context size 4096
A	1	0.24	0.28	0.21
B	1	0.41	0.48	0.45
C	1	0.52	0.52	0.48
A	3	0.38	0.45	0.38
B	3	0.62	0.62	0.66
C	3	0.59	0.59	0.59
A	5	0.62	0.72	0.66
B	5	0.72	0.72	0.72
C	5	0.69	0.72	0.72
A	10	0.90	0.86	0.86
B	10	0.79	0.79	0.79
C	10	0.86	0.90	0.86

TABLE I
ACCURACY OF MODELS WITH DIFFERENT CONTEXT SIZES AND K-NEAREST NEIGHBORS (KNN) FINDING THE CORRECT CONTEXT PASSAGE. THE BEST MODEL (HIGHEST AVERAGE OF SCORES) IN EACH KNN GROUP IS HIGHLIGHTED WITH A GRAY BACKGROUND.

A = all-MiniLM-L6-v2, B = all-mpnet-base-v2, C = paraphrase-multilingual-MiniLM-L12-v2

Model	Knn	Context size 1024	Context size 2048	Context size 4096
A	1	0.32	0.34	0.28
B	1	0.45	0.55	0.52
C	1	0.56	0.56	0.51
A	3	0.55	0.65	0.70
B	3	0.72	0.78	0.72
C	3	0.75	0.75	0.70
A	5	0.69	0.78	0.73
B	5	0.78	0.78	0.79
C	5	0.78	0.80	0.85
A	10	0.95	0.91	0.88
B	10	0.84	0.84	0.85
C	10	0.94	0.95	0.91

TABLE II
ACCURACY OF MODELS WITH DIFFERENT CONTEXT SIZES AND K-NEAREST NEIGHBORS (KNN) FINDING PASSAGES WITH KORRECT KEYWORDS. THE BEST MODEL (HIGHEST AVERAGE OF SCORES) IN EACH KNN GROUP IS HIGHLIGHTED WITH A GRAY BACKGROUND.

A = all-MiniLM-L6-v2, B = all-mpnet-base-v2, C = paraphrase-multilingual-MiniLM-L12-v2

B. Finetuning

Finetuning an LLM via PEFT on document paragraphs with the goal of teaching it new knowledge and expecting accurate question answering does not yield satisfying results in practice. Although the LLM may learn to mimic the style of the documents, it tends to hallucinate frequently, producing names and details that are no more likely to be correct than incorrect.

In contrast, instruction finetuning aimed at improving the model's performance in Retrieval Augmented Generation leads to significant performance gains when answering questions based on a given context. This improvement can be further enhanced by training the LLM on a dataset where answers are structured as Chain of Thought (COT) responses, guiding the model to work through problems step by step. The dataset used for finetuning with COT answers was approximately twice the size, which may have also contributed to the improved performance (192 vs. 452 rows). The model's enhanced ability to extract relevant excerpts, identify key elements, and formulate answers is evident in Figure 2, where finetuned models trained on COT answers outperformed others.

Interestingly, models trained on noisy, synthetic datasets that were not manually cleaned or validated before finetuning performed slightly better in some metrics, such as Keyword Score. For example, `mistral-rag-instruct.HF-NEW` (finetuned on a cleaned dataset) and `mistral-rag-instruct.HF-OLD` (finetuned on a noisy dataset) in Figure 2 show that the latter had a slight edge, though this could be attributed to various factors such as dataset size or random variation.

The semantic correctness of the generated responses was nearly on par with proprietary models like `GPT-3.5-turbo` and the finetuned models even outperformed it in keyword extraction. However, they scored lower in metrics comparing text similarity, since they generate more verbose responses through their Chain of Thought answer process, leading to greater divergence from the reference answers. This is reflected in Figure 2 (note that while cosine similarity is normalized to 1.0, the actual cosine similarity values can be seen in Figure 3).

IV. CONCLUSION AND RECOMMENDATION

Our findings suggest that developing a chatbot for the examination office is feasible, especially given the rapid advancements in this technology. However, there remains a significant level of uncertainty due to the tendency of large language models (LLMs) to generate incorrect or fabricated information when they lack sufficient knowledge. This issue persists even in advanced LLMs like ChatGPT, despite the substantial financial investments made in their development, making it likely that our local model will struggle to match their performance. Given the importance of accurate information in the context of an examination office, we recommend using the chatbot with caution, ensuring that it always provides references to the source documents and specific page numbers. Additionally, further exploration of options such as integrating multiple LLMs or other advanced techniques should be pursued to enhance the reliability of the system.

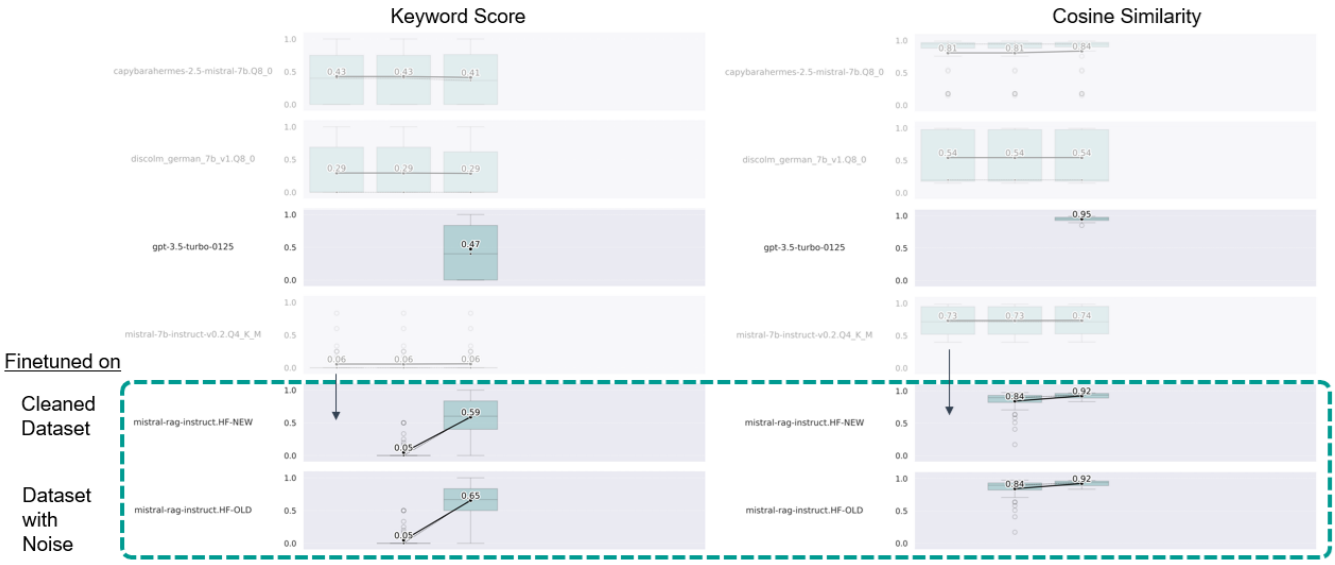

Fig. 2. Finetuning - Semantic Performance over several evaluation runs

Fig. 3. Finetuning - Comparison for all metrics on one evaluation run