Enhancing Student Support Services with Large Language Models: Developing a Chatbot and Virtual Assistance System for the University of Padova

Mohammad Khosravi
mohammad.khosravi.1@studenti.unipd.it

SINA DALVAND
sina.dalvand@studenti.unipd.it

LORENZO ROSALES VASQUEZ
lorenzo.rosalesvasquez@studenti.unipd.it

NAZANIN GHORBANI
nazanin.ghorbani@studenti.unipd.it

MOHAMMAD HOSSEINIPOUR
mohammad.hosseinipour@studenti.unipd.it

Abstract

In this paper, we explore the development of a chatbot for the University of Padova aimed at enhancing the student experience by providing accurate information on courses, fees, accommodations, bureaucratic procedures and more. To achieve this, we evaluated five state-of-the-art pre-trained Large Language Models (LLMs): Llama3, Gemma, Qwen, Mixtral, and Mistral. Our methodology involved a comprehensive comparison based on metrics such as RAG score, hallucination, bias, and toxicity, as well as human evaluation, multilingual capabilities, ease of fine-tuning, and usage rights. After selecting the most suitable model, we conducted a preliminary training phase using the university's computer science master's program catalog. Given the constraints of time and computational resources, we propose a future work plan involving extensive training on the entire university dataset and implementing a Retrieval-Augmented Generation (RAG) pipeline for continuous knowledge base updates. Our findings suggest that while existing chatbots at the university are limited in accuracy and scope, our proposed solution shows promise for providing a more reliable and comprehensive service. Future work will focus on scaling the model, refining the RAG architecture, and addressing practical deployment challenges.

1 Introduction

The rapid advancement of artificial intelligence (AI) and natural language processing (NLP) technologies has opened up new possibilities for enhancing the educational experience. Universities are increasingly looking to leverage these technologies to improve the services they offer to students, streamline administrative processes, and ensure timely dissemination of information. The University of Padova, one of Italy's oldest and most prestigious institutions, is no exception. In this project, we aim to develop an intelligent chatbot designed to enhance the student experience at the University of Padova by providing accurate and timely information on courses, fees, accommodations, bureaucratic procedures and so on.

The primary goal of this project is to identify and implement the most suitable large language model (LLM) that can meet the specific needs of the University of Padova. To achieve this, we have conducted a thorough evaluation of five state-of-the-art pre-trained LLMs: Llama3, Gemma, Qwen, Mixtral, and Mistral. Our evaluation is based on several key factors, including performance metrics, human evaluation, ease of fine-tuning, multilingual capabilities, and usage rights. This comprehensive analysis will guide our selection of the optimal model for developing the chatbot.

The dynamic nature of university resources presents a significant challenge in maintaining the accuracy and relevance of the information provided by the chatbot. To address this, we propose a two-pronged strategy:

First, fine-tuning and training the selected LLM on the current information resources available at the University of Padova. And the second, implementing a Retrieval-Augmented Generation (RAG) pipeline to ensure continuous updates to the LLM's knowledge base with new and changing information.

Given the constraints of time and computational resources available for this NLP course, our initial efforts will focus on training the selected LLM on a smaller dataset, specifically the master's degree program in computer science course catalog. This approach allows us to demonstrate the feasibility and effectiveness of our model on a manageable scale. In the future, we plan to expand this training to encompass the entire dataset of university information.

The use of a RAG pipeline to update the knowledge base of the LLM is a novel approach that has seen limited application. This method involves integrating technologies such as vector databases to find embeddings and similarities of concepts queried by students. While we will only assess the RAG capability of the LLM in the current methodology section, a more detailed exploration of this approach is planned for future work.

In the following sections, we will discuss related works that have been done, our methodology for evaluating and selecting the most suitable LLM, present the experiments

conducted to train and test the chosen model, and outline the future work required to fully develop and deploy the chatbot. Through this project, we aim to contribute a significant improvement to the digital services offered by the University of Padova, ultimately enhancing the student experience and facilitating smoother administrative interactions.

2 Related Work

The development and deployment of chatbots in educational institutions have gained significant traction globally. Leveraging advanced NLP models like GPT-3 and GPT-4, chatbots have revolutionized how people access information and support. These models are capable of understanding and generating human-like text, making them suitable for a wide range of applications including answering queries, providing personalized learning experiences, and supporting administrative functions.

Globally, several universities and educational platforms have integrated chatbots to enhance student engagement and streamline services. For instance, Stanford University uses a chatbot to help students with course selection and scheduling, while the Georgia Institute of Technology employs an AI teaching assistant named Jill Watson to support students in an online course. These implementations have demonstrated the potential of chatbots to significantly reduce the workload on administrative staff and improve the efficiency of information dissemination.

In the context of the University of Padova, we conducted a thorough investigation into existing chatbot services. Our findings reveal several initiatives aimed at improving student services through automated systems, though with varying degrees of success:

Department of Economics Management Chatbot: This chatbot operates with predefined questions but exhibits limited performance and accuracy when handling more complex or relevant student queries. Its responses often lack the depth and specificity required to address detailed inquiries effectively.

Career Services Chatbot: Available only in Italian, this chatbot struggles with accuracy and comprehension, often failing to provide meaningful or correct answers to user queries. This limitation significantly reduces its utility for non-Italian speaking students or those seeking precise information about career opportunities.

Telegram Chatbot: Intended to offer a convenient platform for student interaction, this chatbot is currently nonfunctional. Its inability to operate effectively further underscores the need for a more reliable and robust chatbot solution within the university's digital ecosystem. Our analysis indicates that while the University of Padova has taken steps to implement chatbot technology, the existing solutions are plagued by issues such as low accuracy, poor user comprehension, and limited functionality. These deficiencies highlight the necessity for a more advanced and capable chatbot system that can comprehensively understand and respond to student queries across multiple domains.

In this project, we aim to address these shortcomings by evaluating and implementing a state-of-the-art LLM-based chatbot. By leveraging advanced NLP models and innovative techniques like the Retrieval-Augmented Generation (RAG) pipeline, we intend to create a chatbot that not only meets but exceeds the current capabilities of the university's existing systems. This approach promises to deliver a more accurate, efficient, and user-friendly service, ultimately enhancing the student experience at the University of Padova.

3 Methodology

Our methodology focuses on identifying and implementing the most suitable large language model (LLM) for developing a chatbot that addresses the specific needs of the University of Padova. To select the best model, we evaluated five state-of-the-art pre-trained LLMs:

- Llama3 8b, 70b
- Gemma 2b, 7b
- Qwen/Qwen1.5-0.5B-Chat
- Mixtral 8×22b, 8×7b
- Mistral 7b

These models were chosen based on their high performance in the Hugging Face open LLM leaderboard and their potential to meet the requirements of our application.

In the "Models Comparison Factors" section, we introduce the criteria used to evaluate the LLMs under consideration. Subsequently, in the "Model Selection" section, we discuss our final decision on the most suitable LLM for training and fine-tuning, based on the comprehensive evaluation of these factors.

3.1 Models Comparison Factors

The following factors have been employed as our measurement criteria to select the final LLM to proceed with:

3.1.1 Metrics

Commonly used metrics for evaluating LLMs include:

- F1 Score: This metric is the harmonic mean of precision and recall, providing a single measure of a model's accuracy. Precision indicates the percentage of relevant results among the retrieved documents, while recall indicates the percentage of relevant documents that were retrieved.

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): This set of metrics evaluates the quality of summaries by comparing them to reference summaries. It measures the overlap of n-grams, word sequences, and word pairs between the generated summary and the reference summary.

- Perplexity: Perplexity measures how well a probability distribution or probability model predicts a sample. In the context of language models, perplexity quantifies how well the model predicts a sequence of words. A lower perplexity indicates that the model is better at predicting the next word in a sequence, suggesting a higher level of understanding and coherence in its generated text.

However, these traditional metrics are not entirely suitable for evaluating LLMs in our context. When evaluating responses from an LLM, these metrics rely on word-by-word comparisons between the model's output and the expected answers. If the responses do not match exactly, the scores from these metrics can be misleadingly low. This approach does not account for the semantic correctness or the relevance of the responses, which is crucial for our chatbot's effectiveness.

Therefore, we employed the following metrics instead:

- RAG Scores:

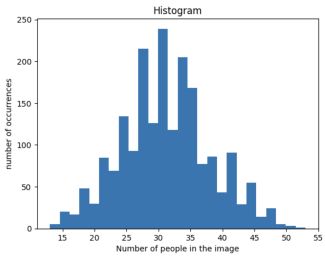

Figure 1: Lorem ipsum dolor sit amet

- Hallucination:

- Bias:

- Toxicity:

3.1.2 Benchmarks

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

- MMLU:

- TruthfulQA:

3.1.3 Human Evaluation

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

3.1.4 Ease of Fine-tuning

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

3.1.5 Multilingual Capability

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

3.1.6 usage rights

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

3.2 Model Selection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

4 Experiments

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Eleifend mi in nulla posuere. Laoreet non curabitur gravida arcu ac tortor dignissim. Ultricies tristique nulla aliquet enim tortor at auctor urna. At urna condimentum mattis pellentesque id nibh tortor id. Consectetur libero id faucibus nisl tincidunt eget.

4.1 Data Preparation

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

4.2 Model Training

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

4.3 Model Testing

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

5 Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

6 Future Work

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

7 References

References

[1] Mall Dataset. (Updated Apr 2014). Crowd Counting Dataset. Retrieved from `https://personal.ie.cuhk.edu.hk/~ccloy/downloads_mall_dataset.html`

[2] C. C. Loy, S. Gong, and T. Xiang, "From Semi-Supervised to Transfer Counting of Crowds," in Proceedings of IEEE International Conference on Computer Vision, pp. 2256-2263, 2013 (ICCV)

[3] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative Attribute Space for Age and Crowd Density Estimation," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2467-2474, 2013 (CVPR, Oral)

[4] C. C. Loy, K. Chen, S. Gong, T. Xiang, "Crowd Counting and Profiling: Methodology and Evaluation," in S. Ali, K. Nishino, D. Manocha, and M. Shah (Eds.), Modeling, Simulation and Visual Analysis of Crowds, Springer, vol. 11, pp. 347-382, 2013

[5] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature Mining for Localised Crowd Counting," British Machine Vision Conference, 2012 (BMVC).