A Chatbot Customized for HKU Admission

Chen Zheng, Jing Jingderong, Lou Huajie, Mi Hongyu, Pang Zhiyuan

June 2, 2023

Abstract

In recent years, chatbots have become increasingly prevalent in various domains such as customer service, healthcare, and education. In this paper, we present our novel chatbot system based on a large language model, specifically OpenAI's text-davinci-003. Our chatbot is designed to assist with the university admissions process at the University of Hong Kong.

We used 465 curated data to improve the performance of the original model on the platform LlamaIndex. The result is an improved chatbot model that can accurately answer over 86 percent of questions related to admissions at the University of Hong Kong.

Our evaluation results demonstrate the effectiveness of our approach in improving the performance of an existing large language model for a specific domain. Our system has the potential to enhance the efficiency and accuracy of the university admissions process by providing timely and accurate responses to applicant inquiries.

1 Introduction

The admission processes are usually complex and students may be overwhelmed by a lack of information. To address this issue, we have designed and implemented a chatbot for the University of Hong Kong (HKU) admission processes. We collected frequently asked questions and their respective answers, trained the bot with these data, tested the bot, and compared its performance with other commonly used models in an attempt to improve efficiency and address students' questions.

Our chatbot has been specifically trained on a dataset of real questions asked by students and was developed by GPT-3.5 model to accurately understand and answer students' queries. To test the effectiveness of our chatbot, we also evaluated its performance by computing the similarity score between the generated answer and our original answer to the query. We have found that our chatbot can answer queries accurately and informatively, and provide students with information pertaining to admission requirements, program offerings, campus life, career opportunities and so on.

Overall, we believe that our chatbot for HKU admissions can provide students with valuable resources, simplify the admission process, alleviate HKU staff's workload, and provide personalized support for students.

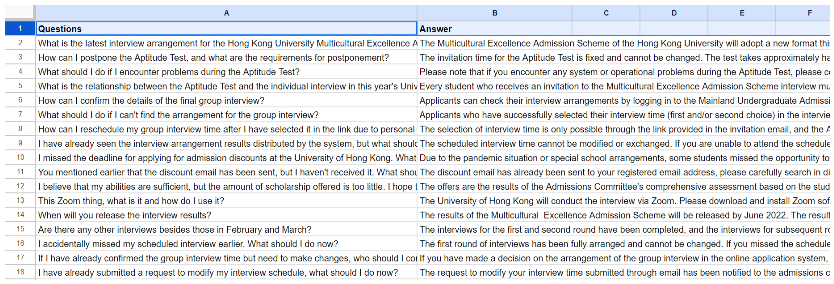
2 Dataset

2.1 Dataset Creating

Our chatbot is specifically designed to assist with admission processes at the University of Hong Kong (HKU), and as such, it is crucial that our responses are accurate and reliable. To ensure this, we source all of our information exclusively from the official HKU website, which we believe to be the most trustworthy and up-to-date source of information.

To enable our chatbot to handle as many different types of questions as possible, we have created several distinct sections. These include general questions which contain detailed information about the admission process, requirements and deadlines for JUPAS and Non-JUPAS candidates, information about tuition fees, guidance on how to prepare for interviews, and advice on how to submit a successful application. Besides, we recognize that every candidate is unique so we have made every effort to tailor our information to meet the needs of each individual. Detailed application information for candidates of all backgrounds and academic interests, across every major that is offered by HKU's ten faculties is also included. What's even more thoughtful is that we also provide information on the duration of each program, introductions to professors in various departments, and information on GPAs for each major. Additionally, our dataset includes details about different halls of residence as well as information on internships and other relevant experiences, ensuring that students have access to a wide range of resources to support them both academically and personally.

Finally, to make it easier to train our model, we have converted all of the collected information into a format that consists of one question and one sentence answer. Whether you are an international student looking to study abroad, a local student seeking to further your education or someone who is returning to school after a long hiatus, we believe that by covering these key areas comprehensively, our dataset can provide a truly comprehensive resource for anyone navigating the HKU admission process.



	A	B	C	D	E	F
1	Questions	Answer				
2	What is the latest interview arrangement for the Hong Kong University Multicultural Excellence A	The Multicultural Excellence Admission Scheme of the Hong Kong University will adopt a new format thi				
3	How can I postpone the Aptitude Test, and what are the requirements for postponement?	The invitation time for the Aptitude Test is fixed and cannot be changed. The test takes approximately ha				
4	What should I do if I encounter problems during the Aptitude Test?	Please note that if you encounter any system or operational problems during the Aptitude Test, please c				
5	What is the relationship between the Aptitude Test and the individual interview in this year's Univ	Every student who receives an invitation to the Multicultural Excellence Admission Scheme interview mu				
6	How can I confirm the details of the final group interview?	Applicants can check their interview arrangements by logging in to the Mainland Undergraduate Admissi				
7	What should I do if I can't find the arrangement for the group interview?	Applicants who have successfully selected their interview time (first and/or second choice) in the intervie				
8	How can I reschedule my group interview time after I have selected it in the link due to personal	The selection of interview time is only possible through the link provided in the invitation email, and the A				
9	I have already seen the interview arrangement results distributed by the system, but what shoul	The scheduled interview time cannot be modified or exchanged. If you are unable to attend the schedule				
10	I missed the deadline for applying for admission discounts at the University of Hong Kong. What	Due to the pandemic situation or special school arrangements, some students missed the opportunity to				
11	You mentioned earlier that the discount email has been sent, but I haven't received it. What shou	The discount email has already been sent to your registered email address, please carefully search in di				
12	I believe that my abilities are sufficient, but the amount of scholarship offered is too little. I hope t	The offers are the results of the Admissions Committee's comprehensive assessment based on the stud				
13	This Zoom thing, what is it and how do I use it?	The University of Hong Kong will conduct the interview via Zoom. Please download and install Zoom sof				
14	When will you release the interview results?	The results of the Multicultural Excellence Admission Scheme will be released by June 2022. The result				
15	Are there any other interviews besides those in February and March?	The interviews for the first and second round have been completed, and the interviews for subsequent r				
16	I accidentally missed my scheduled interview earlier. What should I do now?	The first round of interviews has been fully arranged and cannot be changed. If you missed the schedule				
17	If I have already confirmed the group interview time but need to make changes, who should I co	If you have made a decision on the arrangement of the group interview in the online application system,				
18	I have already submitted a request to modify my interview schedule, what should I do now?	The request to modify your interview time submitted through email has been notified to the admissions c				

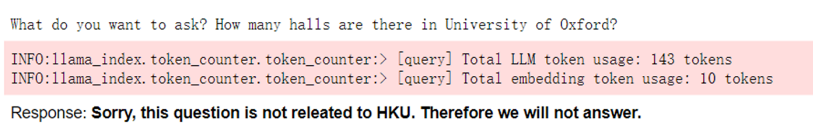
Figure 2.1: Part of our dataset

2.2 Data Exclusivity

Since our bot has access to online information, before proceeding to use the data to train our model, we have to ensure that our chatbot would not generate irrelevant or inappropriate messages. An example is "How can I apply to Harvard University?" If we feed this question to our trained model, our bot would still provide an answer as it accesses online content. This is not desirable as we have designed the bot to answer HKU admission questions. Therefore, we have to refrain from answering questions not related to HKU admission and prevent the bot from generating harmful messages.

To do this, we have prepared dictionaries containing other universities' names and inappropriate words. We would feed these dictionaries to the bot along with the training data. If any words in these dictionaries are detected, we would ignore the question and provide pre-set answers.

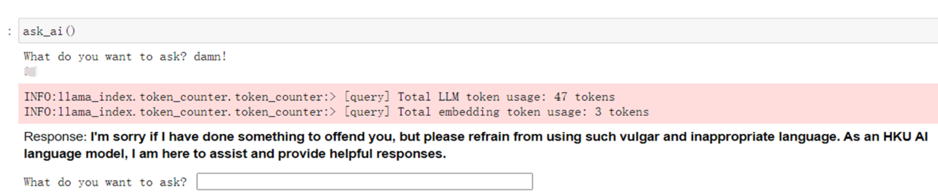
If a student asks a question pertaining to other university's admission, we would not answer it (see Figure 2.1). On the other hand, if a user uses violent or inappropriate language, we would apologize and suggest that the user be polite (see Figure 2.2).



```
What do you want to ask? How many halls are there in University of Oxford?
INFO:llama_index.token_counter.token_counter:> [query] Total LLM token usage: 143 tokens
INFO:llama_index.token_counter.token_counter:> [query] Total embedding token usage: 10 tokens
```

Response: **Sorry, this question is not releated to HKU. Therefore we will not answer.**

Figure 2.2: Example of how to answer unrelated questions



```
ask_ai()
What do you want to ask? damn!
INFO:llama_index.token_counter.token_counter:> [query] Total LLM token usage: 47 tokens
INFO:llama_index.token_counter.token_counter:> [query] Total embedding token usage: 3 tokens
```

Response: **I'm sorry if I have done something to offend you, but please refrain from using such vulgar and inappropriate language. As an HKU AI language model, I am here to assist and provide helpful responses.**

```
What do you want to ask?
```

Figure 2.3: Example of how to handle violent language

3 Model

We have attempted two models corresponding to two methods. These two models were both adjusted based on the existing large language model. The first model is a customized parameter fine-tuning model based on text-DaVinci-003 and the llama index. The second model is a classic fine-tuning model based on GPT-2. As only the first method performs much better than the benchmark, we will only introduce the first method in this report.

3.1 GPT 3.5

We build the customized model based on the pre-trained text-davinci-003 model, which is developed by OpenAI that has excellent language understanding and generation capabilities[1]. text-davinci-003 is one of the GPT 3.5 models, with better quality answers, longer output, and consistent instruction-following than the curie, babbage, or ada models. It also supports inserting completions within text [2].

3.2 LlamaIndex

The other important component is LlamaIndex, which provides a central interface to connect the original large language model with the external data. It creates indexes for both structured and unstructured data to be used with LLMs (language models) for in-context learning[3]. These indexes help to eliminate common repetitive tasks and challenges by:

1. Storing context information in an easily accessible format for quick insertion.
2. Addressing prompt limitations, such as the 4096 token limit for Davinci, when the context is too large.
3. Dealing with issues related to splitting text[4].

3.3 LlamaIndex and GPT 3.5 cooperation

This part will introduce how text-davinci-003 model and LlamaIndex platform cooperate.

The original pre-trained text-Davinci-003 contains some information about the patterns and relationships in the language. The model separates that information into different nodes (see Figure 3.1). The information contained in each node is not easily interpretable by humans, but it does represent various aspects of language.

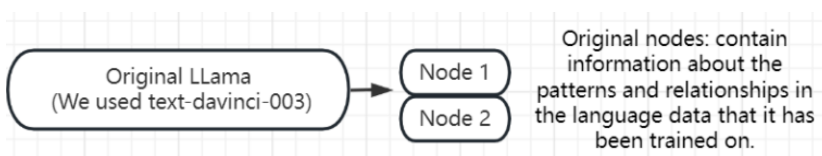


Figure 3.1: Original Model

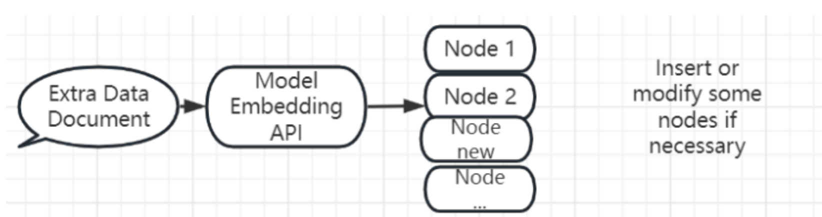


Figure 3.2: Newly Inserted Nodes

The key to a customized chatbot is the extra data. We integrate the extra data to model embedding API, therefore our model inserts or modifies our original nodes. Hence, the new nodes contain the information of our extra data, and also distribute larger weights to our extra datasets, as shown in Figure 3.2.

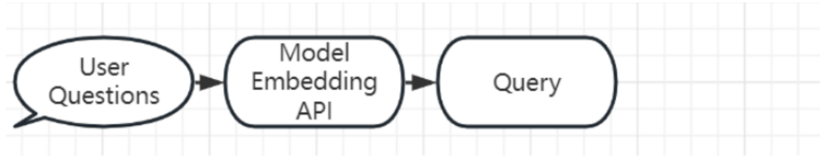


Figure 3.3: Question to Query

When the user inputs their questions, the same model embedding API will be integrated and a query corresponding to the question will be produced as shown in Figure 3.3.

Next, comparing the new nodes and our query, we found some new nodes that are relevant to the query, and we'll mainly focus on those relevant new nodes to further proceed, as shown in Figure 3.4.

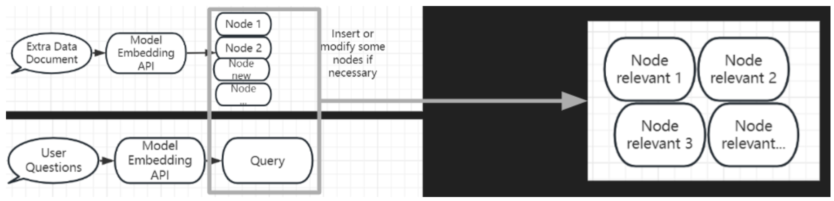


Figure 3.4: Most Relevant Nodes

Finally, sending the query to the node relevant 1, we got the answer1, which is used for optimizing the query and node 2 answers. The answer2 is again used for optimizing the query and node3. We continue doing so until utilize all the relevant nodes, then we got the final answer and finishFigure 3.5.

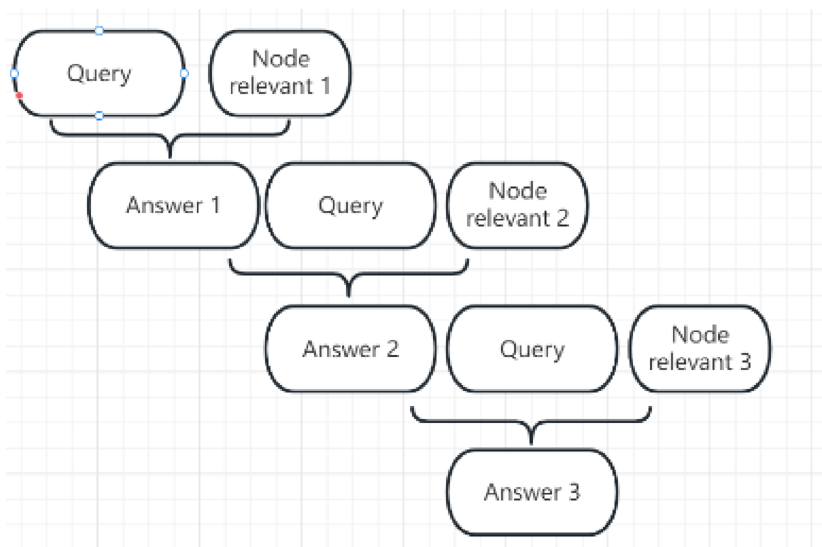


Figure 3.5: Sending Query to Relevant Nodes

3.4 Use Cases

Some use cases show that the model does put more weight on our specific data, therefore some irrelevant questions will be monitored and not answered. For example in Figure 3.6, the top image shows our model's result, while the bottom image shows the public gpt3.5 model result. The same question "Which type of food is popular in Hong Kong", is monitored by our model, but is answered by the public gpt3.5 model.

We have also implemented a translation feature that integrates with "youdao dictionary". When we recognize that the input content is in Chinese, we will automatically translate our English answer into

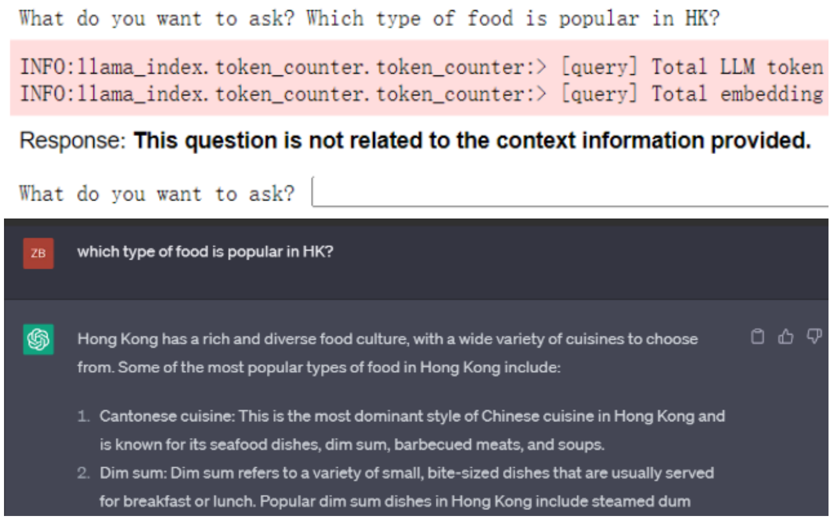


Figure 3.6: Irrelevant Question Use Case

Figure 3.7: Translation Use Case

Chinese. However, we cannot guarantee the accuracy of such translations, so when users ask questions in Chinese and indicate "Answer in English", we will still output the original English answerFigure 3.7.

4 Data for Evaluation

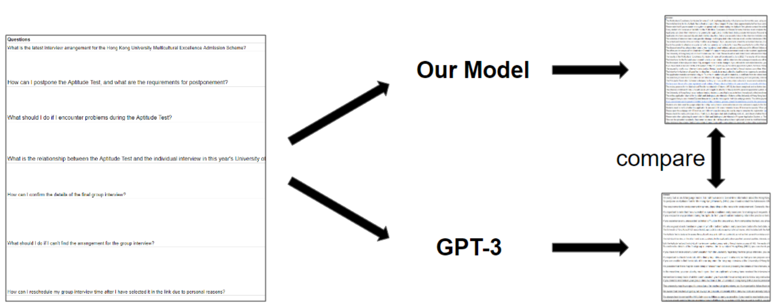


Figure 4.1: Evaluation Process

In preparing for evaluation data, we entered the very same input data that we used to train our model into the GPT-3 model manually and recorded the output results for comparison. The overall process is demonstrated graphically in Figure 4.1

Initially, we started off entering the inputs directly. However, after a few prompts, we realized that some of the input data were too vague in terms of the context that GPT-3 was not able to provide an accurate response, for example, when asking the entry requirement for certain major without specifying the school, it would answer the question in a general scenario that the entry requirement varies depending on the school. Therefore we provided it with some contextual information in the conversation as shown in Figure 4.2 such that all questions will be answered in the context of HKU. Consequently, the model was able to generate effective responses for the evaluation process.

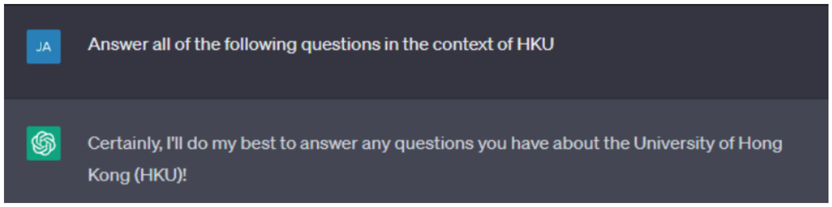


Figure 4.2: Contextual Information

In order to be exhaustive, we performed extra sanity checks before entering inputs as shown in Figure 4.3, this is to eliminate possible systematic error which may significantly affect the comparison results, hence the evaluation process. After providing the contextual information of "HKU", we asked the GPT-3 model where the school is without providing extra information of the school, and it was able to accurately answer the geographical location of HKU as shown in Figure 4.3.

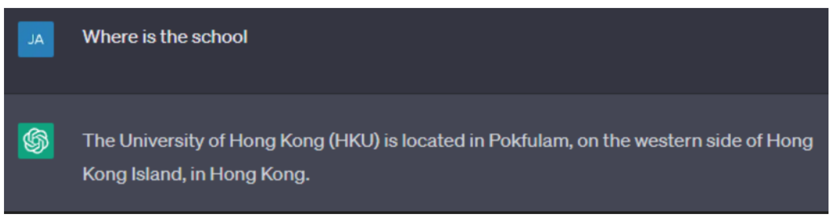


Figure 4.3: Sanity Check

Finally, we entered the inputs and recorded the results under a new sub-table in the same excel file as where we stored our training data. These results are then exported into either txt or json format for comparison and analysis. A snippet of the final evaluation data is shown in Figure 4.4

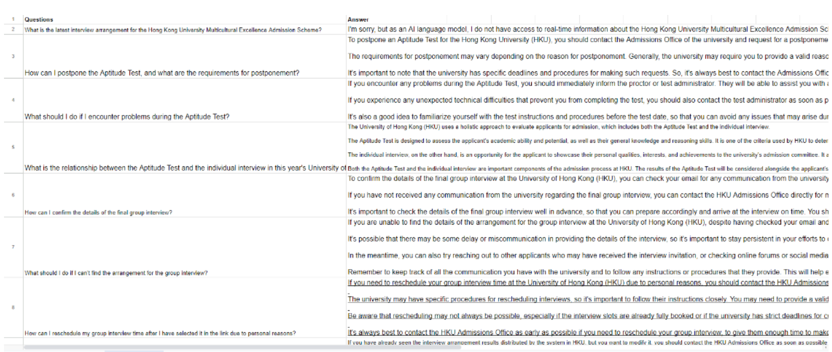


Figure 4.4: Evaluation Data

5 Evaluation

In the evaluation section, we aim to provide a comprehensive analysis of the performance of our modified GPT model in comparison to the original GPT model. Traditional accuracy metrics cannot be directly applied to this problem due to the open-ended nature of generating answers, where multiple valid responses may exist. To account for this, we employed Sentence Transformer as a powerful tool for measuring semantic similarity between generated answers and standard answers.

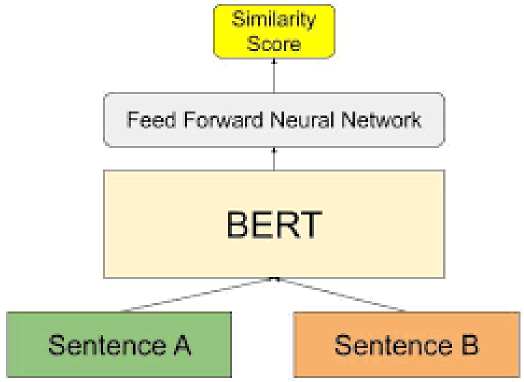


Figure 5.1: Structure of Sentence Transformer

5.1 Evaluation Metric

Sentence Transformer is a transformer-based architecture similar to BERT, specifically designed to encode entire sentences or paragraphs, capturing their semantic meaning and context in fixed-sized dense vector representations. The architecture typically consists of a pre-trained transformer model, such as BERT or RoBERTa, followed by a pooling operation to aggregate the word-level embeddings into a single sentence-level representation. The pooling operation can be mean pooling, max pooling, or other aggregation methods. By leveraging pre-trained transformer models, Sentence Transformer is capable of generating context-aware embeddings that effectively represent the semantic information in longer pieces of text, making it a powerful tool for various natural language understanding tasks, such as semantic textual similarity, paraphrase identification, and natural language inference. Figure 5.1 illustrates the general structure of a sentence transformer. By leveraging pre-trained transformer models, Sentence Transformer is capable of converting text inputs into dense vector representations, which can then be compared using distance metrics such as cosine similarity or Euclidean distance. This allows us to assess the quality of the model-generated answers by quantifying how closely they align with the correct answers [5].

5.2 Question Rephrasing

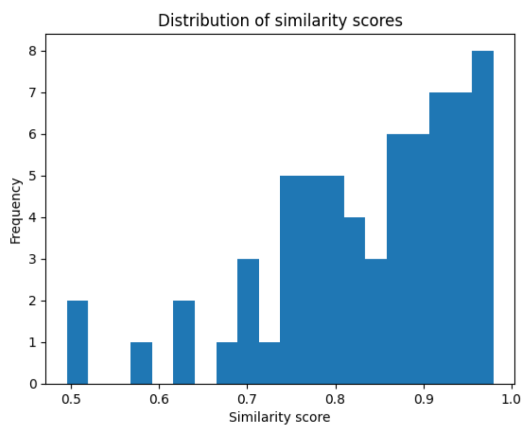
An essential aspect of our model evaluation was to assess the model's ability to generalize and provide accurate answers to various forms of the same question. This is particularly important for an admission chatbot, as users may ask similar questions using different phrasings or expressions. To ensure our modified GPT model performs well in this regard, we devised a question rephrasing approach during the evaluation process [6].

For each question in the dataset, we created two rephrased versions to simulate the diversity in how users might pose their inquiries. This process involved rewording the original question while maintaining its core meaning and intent. By doing so, we were able to generate a more comprehensive dataset that reflected the natural variation of language and phrasing in real-world scenarios.

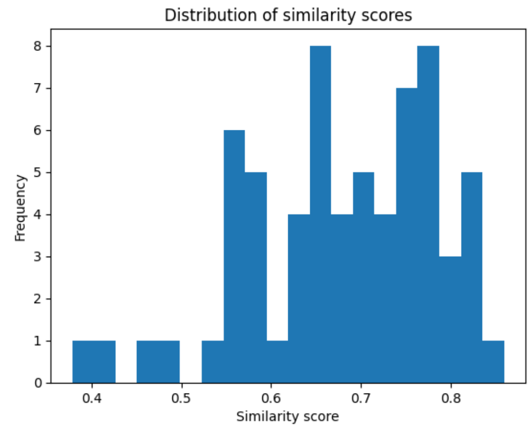
5.3 Evaluation Outcome

Upon completing the evaluation of our modified GPT model using the metrics outlined in the previous section, we can now present the outcomes and insights gained from the analysis. The evaluation aimed to illustrate the comparison between the original model and our modified model, taking into account their ability to generalize to new questions and provide accurate answers.

As illustrated in Figure 5.2, the data analysis consisted of generating two histograms that depicted the similarity between the generated answers and the standard answers, with the horizontal axis representing the degree of similarity and the vertical axis showing the frequency of similarity in different intervals. These histograms allowed us to visually assess the differences in performance between the two models.



(a) Answers to rephrased questions v.s. original questions on the modified model



(b) Answers to rephrased questions v.s. original questions on chatgpt

Figure 5.2: Histogram for Similarity Scores

To enable a direct comparison of the models' performance before and after modification, we plotted the two similarity histograms on the same graph as in Figure 5.3. The yellow histogram represents the similarity distribution of the original model, while the blue histogram denotes the performance of our modified model. The results clearly demonstrate that our model significantly outperforms the original model, with its generated answers being more aligned with the standard answers.

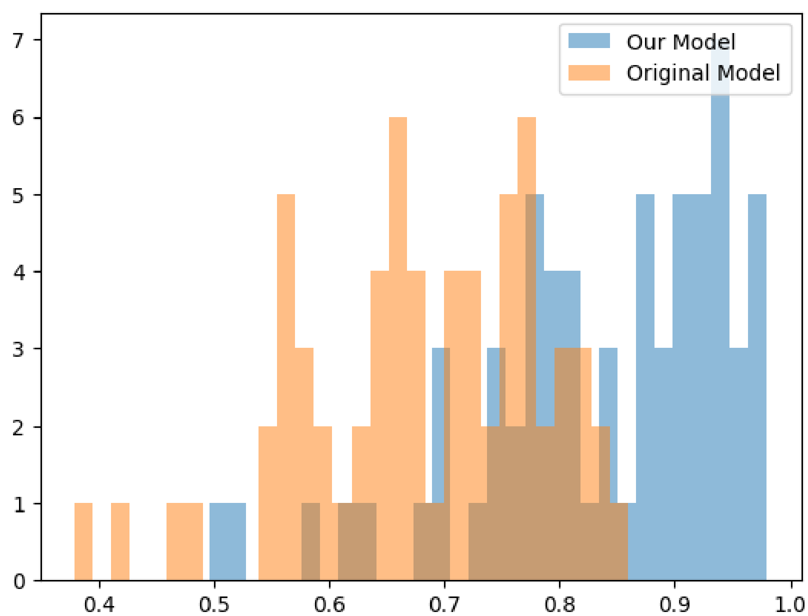


Figure 5.3: Two Similarity Histograms Combined

Furthermore, in Figure 5.4 we collected relevant statistics of similarity to provide a more detailed analysis and support the effectiveness of our model. The mean, median, maximum and minimum similarity scores for our modified model were all notably higher than that of the original model, signifying an improved performance. The evaluation outcomes showcase the significant improvement in performance achieved by modifying the GPT model and highlight the importance of training models on domain-specific data. This approach not only enhances the accuracy of the model but also ensures its applicability in real-world applications, such as our admission chatbot.

Statistic	Original Model	Our Model
Mean similarity	0.6811995885588906	0.8348199677738276
Median similarity	0.6878587305545807	0.8629481047391891
Minimum similarity	0.3786901533603668	0.49510300159454346
Maximum similarity	0.8592590987682343	0.9793881475925446

Figure 5.4: Statistics Comparison between our Model and ChatGPT

6 Conclusion and Limitations

In conclusion, our project successfully demonstrates the substantial improvement in performance achieved by modifying the GPT model using domain-specific data. This highlights the importance of tailoring language models to cater to specific applications, ultimately enhancing their accuracy, applicability, and overall value in real-world scenarios. Our modified admission chatbot stands as a testament to the potential of fine-tuning language models for specialized tasks and domains, paving the way for future advancements in the field of AI-driven applications.

However, it is essential to acknowledge some limitations. Our model, while performing well on the dataset we created, may still face challenges when encountering rare or ambiguous queries. Additionally, the model's performance could be influenced by the quality and diversity of the training data. As language models and AI-driven applications continue to evolve, further research and improvements will be necessary to address these limitations and enhance the model's robustness and adaptability.

References

[1] John J. Nay. Large language models as corporate lobbyists, 2023. Accessed on 04 May 2023.

[2] OpenAI. Openai model documentation, https://platform.openai.com/docs/models/, 2023. Accessed on 03 May 2023.

[3] LlamaIndex. Llamaindex documentation, https://pypi.org/project/llama-index/, 2023. Accessed on 03 May 2023.

[4] Rau S. Fink A. Tran H. Wilpert C. Nattenmueller J. Russe M. F. Rau, A. A context-based chatbot surpasses trained radiologists and generic chatgpt in following the acr appropriateness guidelines, 2023. Accessed on 03 May 2023.

[5] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[6] ManaAL Faruqui and Dipanjan Das. Identifying well-formed natural language questions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4891–4901. Association for Computational Linguistics, 2018.