Enhancing Chatbot Performance for Domain-Specific Data

Chenmeinian Guo
cg3972@nyu.edu

Jicheng Yu
jy2575@nyu.edu

Taoming Liu
tl3109@nyu.edu

ABSTRACT

The ability of chatbots to handle domain-specific queries is crucial as they become more prevalent in specialized fields. Traditional chatbot models often lack the nuanced understanding required for specialized fields, leading to inaccurate or irrelevant responses. This paper explores the enhancement of chatbot capabilities through two methodologies: Fine-Tuning and Retrieval-Augmented Generation (RAG). These techniques seek to improve the accuracy and contextual relevance of the chatbot's responses by incorporating proprietary and domain-specific data. We describe our methodology, which uses RAG to dynamically incorporate external knowledge during query processing and Fine-Tuning to refine the model's parameters to better fit particular domains. Our results demonstrate a significant improvement in the precision and reliability of chatbot responses across domain-specific datasets, emphasizing the potential of these techniques in transforming chatbot interactions in focused fields.

The accompanying code and further details can be found at https://github.com/guochenmeinian/Llama-Langchain-RAG.

Keywords: Chatbots, Domain-Specific Adaptation, Fine-Tuning, Retrieval-Augmented Generation, Large Language Models, Contextual Understanding.

1 INTRODUCTION

In recent years, the use of chatbots across various areas such as education, e-commerce, and personal assistance has grown significantly. While these systems are highly effective in general inquiries, their performance significantly deteriorates when dealing with domain-specific questions. This shortfall is mostly caused by the generic training datasets used during the development of these models, which often lack domain-specific depth. This paper studies not only the implementation of methods to enhance the capability of chatbots to handle specialized queries more effectively but also evaluates how these methods contribute to improved model performance, thereby reducing response errors and improving user satisfaction.

2 BACKGROUND AND RELATED WORKS

The introduction of machine learning and particularly deep learning has revolutionized the field, enabling chatbots to understand and generate human-like responses. Large Language Models (LLMs) like GPT-3 and others have been crucial in improving the capabilities of conversational bots. However, their use in specialized domains frequently requires further context-specific adaptations. To address these limitations, the research community has explored several strategies. Fine-tuning pre-trained models on domain-specific data has shown promise in enhancing chatbot relevance and accuracy. Retrieval-Augmented Generation (RAG), which dynamically integrates external knowledge sources during the conversation, offers another way to enhance chatbot responses with accurate and relevant information. Each of the two methods has its strengths and trade-offs, which we will discuss in detail.

2.1 LLaMA-2

LLaMA-2 represents a refinement in the series of transformer-based models that focus on efficiency and performance in natural language processing tasks [8]. Despite its robust training on diverse datasets, the model often requires further optimization to effectively address domain-specific queries.

2.2 Fine-Tuning

Fine-tuning involves adapting a pre-trained model on a smaller, domain-specific dataset to refine its responses relevant to that domain. This process helps the model to understand and generate responses that are contextually appropriate for the domain-specific interactions.

Principles of Fine-Tuning: Fine-tuning adjusts the weights of the LLM by continuing the training phase on a specialized corpus, thereby embedding deeper domain-specific knowledge into the model. This method has proven effective in enhancing the relevance and accuracy of responses in targeted domains [9].

2.3 Retrieval-Augmented Generation (RAG)

RAG addresses the limitation of fixed knowledge by dynamically pulling in external data during the conversation. This method enriches the chatbot's responses by integrating retrieved data relevant to the ongoing interaction, allowing it to provide informed and accurate answers based on the latest available information.

Mechanism: RAG combines the generative power of LLMs with a retrieval system, where the model queries a database of information to retrieve data relevant to the user's query before generating a response. This approach enhances the model's ability to provide responses that are not only relevant but also up-to-date and factually accurate [5].

3 METHODS

3.1 Model Selection

We selected the LLaMA-2-13B-chat model as our foundational language model due to its proven balance of computational efficiency and performance capabilities [8]. This selection provides a robust baseline that is well-suited for domain-specific adaptations necessary for enhancing chatbot performance, especially when benchmarked against other models in similar categories.

3.2 Domain Selection

Our project's domain centers on trivia about popular TV shows, particularly the series "Friends." This domain was chosen based on its broad appeal and the availability of diverse content for data collection, making it ideal for showing the effectiveness of our methodology in a real-world, entertainment-focused scenario.

3.3 Data Collection and Preprocessing

Data for this project was mainly collected through manual scraping of trivia content from websites like Seventeen [7] and BuzzFeed [1], and through the generation of content by prompting techniques with specific episodes' plots using ChatGPT-4, such as sourcing from the Friends Wiki [4]. This combined approach ensures both relevance and quality of data, consisting of trivia questions and answers related to "Friends."

For instance, a trivia about Monica's cooking from the website was transformed as follows:

- **Original**: What review was written about Monica's cooking? A food critic upsets Monica by writing a review saying, "The mahi-mahi was awful-awful.
- **Transformed QA Pair**: {"prompt": "What review was written about Monica's cooking?", "completion": "A food critic upsets Monica by writing a review saying, 'The mahi-mahi was awful-awful.'"}

Similarly, content generated from the Friends Wiki was processed. Below is a snippet of the summary of the character Rachel Greene's plot in Season 2:

- **Original**: Rachel goes to the airport to greet Ross when he comes back from China, then tries to leave in a hurry when she sees he has come back with a woman he is intimate with, and falls over, causing the couple to notice her. Rachel's so upset she has a one night stand with Paolo. **She acts incredibly cold to Julie and does everything she can to sabotage her relationship with Ross, including making Phoebe cut her hair "Planet of the Apes" style and trying to convince Ross to avoid sleeping with her, saying it makes women really hot for a guy if he makes her wait.** In an attempt to move on from her feelings for Ross, Rachel goes on a blind date but spends the whole evening drinking wine and talking about the situation with Ross. Rachel gets drunk and leaves a message on Ross' answering machine, claiming that she is over him. Ross listens to the message the next morning, becoming confused and a little angry. He goes to Central Perk later that night to confront Rachel, and they share their first passionate kiss.
- **Generated QA Pair**: {"prompt": "In Season 2, what extreme measures does Rachel take to sabotage Ross and Julie's relationship?", "completion": "Rachel tries to sabotage Ross and Julie's relationship by influencing Phoebe to give Julie a bad haircut and advising Ross to delay being intimate with Julie, claiming it will make Julie more attracted to him."}

We curated the information into high-quality question-answer pairs as illustrated above. For data generated via ChatGPT-4, each entry was rigorously fact-checked by our team to ensure accuracy. We utilized Python to transform the trivias into QA pairs in the JSONL format. This format, consisting of {"prompt": , "completion":} pairs, is the format required for fine-tuning the LLaMA-2 model. The use of Python was crucial in automating the conversion process, ensuring the data was properly structured for model training.

3.4 Model Enhancement

3.4.1 Fine-Tuning. The curated question-answer pairs were utilized to fine-tune the LLaMA-2-13B-chat model on the Replicate platform, using an 8x A40 GPU setup. The data, stored in JSONL format, included prompts and completions for training and optional validation. To enhance the model performance, we increased the number of training epochs from the recommended three to five. This adjustment allowed for deeper learning and finer adaptation to the nuances of "Friends" trivia.

The training was executed with a global batch size of 4 and a micro batch size of 4, using gradient accumulation when necessary. The validation, integrated into the training process, involved 50 samples from the dataset. This validation helped monitor the model's performance and adjust parameters accordingly.

To optimize the training process, the learning rate was set at 0.0001, the LoRA rank at 8, and the dropout rate at 0.05. This configuration aims to strike a balance between training speed and model robustness. Full single-precision floating point (FP32) was used for computations to ensure the precision needed for the complexity of the model.

3.4.2 Retrieval-Augmented Generation (RAG). Together with fine-tuning, we implemented a RAG component leveraging the same dataset. We employed an advanced retrieval mechanism that transforms text into embeddings via the OpenAIEmbeddings class from the langchain-openai library. These embeddings help the model semantically understand documents that are preprocessed and categorized into formats like text or JSONL.

For indexing, we used the Chroma vector store to manage and retrieve document embeddings. Documents are segmented into manageable chunks using a RecursiveCharacterTextSplitter, providing detailed granularity in retrieval. This chunking allows for a more focused and relevant fetch of information during the query phase.

During retrieval, the system fetches semantically similar document chunks from the Chroma database based on the query's embeddings. These chunks form a comprehensive context along with the user's query, resulting in enriched prompts for the model. Hosted on the Replicate platform, the model processes these prompts to generate contextually aware and precise responses. This method ensures that responses are not only relevant but also rich in accurate contextual details.

3.5 Interface Implementation

The Streamlit frontend of our LLaMA2 chatbot offers an interactive interface for users to engage with different configurations of the model, including the base LLaMA-2-13B-chat, finetuned one of the same, base with RAG, and finetuned with RAG. The chat interface manages session history, allowing users to input questions related to the "Friends" TV show, and dynamically generates responses based on the selected model. Responses are produced by constructing prompts that are limited to the current question only, enhancing accuracy and ensuring contextually precise interactions. This interface supports clear history functionality and error handling for API token validation, providing a user-friendly environment to explore the model's capabilities with real-time feedback.

4 EXPERIMENTAL SETUP

To evaluate how both fine-tuning and RAG techniques improve our models' answer performance, and due to the nature of the domain where many new trivia elements can only be memorized but not deduced from preexisting trivia or plots it has already learned, we selected 20% of our dataset entries and rephrased the questions part (answers must not be changed) as our evaluation dataset. We used GPT-4 to paraphrase the original questions. Each paraphrased question was manually checked to ensure they maintained the original intent, testing the models' abilities to recognize queries. This rephrasing aimed to ensure that the evaluation focuses on understanding the questions rather than memorization, making our evaluation dataset more robust against superficial text matching.

Additionally, GPT-4 was employed to assess the quality of the answers. Recent studies, such as those by Chiang and Lee [2], Dubois et al. [3], and Min et al. [6], suggest that large language models often align closely with human judgment and can effectively evaluate responses, regardless of whether the context is provided or not.

We tested the four models mentioned earlier: the base LLaMA-2-13B-chat, the fine-tuned LLaMA-2-13B-chat, the base model with RAG, and the fine-tuned model with RAG.

4.1 Grading Criteria

We introduced our three main grading criteria: Accuracy, Succinctness, and Relevancy, allowing us to measure improvements in model performance due to fine-tuning and RAG techniques quantitatively and qualitatively, across multiple dimensions of answer quality.

4.1.1 Accuracy. For each question and answer in the evaluation dataset, GPT-4 was prompted to generate an detailed evaluation metric. For example:

- **Question**: Who offers Ross comfort over his unreturned feelings for Rachel in Season 1, and what incident occurs between them?
- **Answer**: Chandler's mom, Nora Tyler Bing, comforts Ross about his situation with Rachel, and they end up kissing, which upsets Chandler.
- **Evaluation_guideline**: The answer should include that Nora Tyler Bing offers Ross comfort, they share a kiss, and this incident upsets Chandler.

Provided with such guidelines for each question in the evaluation dataset, GPT-4 scored the accuracy of the models' answers from 0% to 100%.

4.1.2 Succinctness. We created the following scoring sheet, which GPT-4 used to assess the answers:

Verbose : The answer includes unnecessary details, background information, or reiterations that do not directly contribute to addressing the question. For instance, detailing the entire scene or adding irrelevant character backstories.

Somewhat Verbose : The answer is more detailed than necessary, including some superfluous elements or minor details that, while related to the question, are not essential for understanding the response.

Neutral : The answer strikes a balance, providing enough detail to satisfactorily answer the question without venturing into unnecessary specifics. It may contain a brief explanation or a concise background when relevant.

Somewhat Succinct : The answer is compact and focuses strictly on the key elements of the question, omitting most extended explanations but still clear and complete.

Succinct : The answer directly addresses the question with minimal words, focusing solely on the core information needed. It provides the answer in the shortest form while still being accurate.

4.1.3 Relevancy. The following scoring sheet was developed:

Relevant Answer : Fully addresses the prompt: The response directly answers the question asked without omitting any critical information. Accurate details: The details provided in the response are factually correct and match the events or descriptions from the show. Contextually appropriate: The answer should be within the context of the question, reflecting an understanding of the specific scene, event, or character dynamics mentioned in the prompt.

Partially relevant Answer : Addresses the prompt but is incomplete: The response answers the main question but omits significant details that contribute to a full understanding of the event or context. Minor inaccuracies: Contains slight factual errors that do not completely distort the understanding of the event but affect the accuracy of the response. Lacks nuance or depth: Provides a correct general answer but fails to capture the nuances or emotional tones that are pivotal to the scenes or character interactions.

Irrelevant Answer : Does not address the prompt: The response fails to answer the question asked, either by misunderstanding the prompt or discussing unrelated content. Factually incorrect: Contains major factual errors regarding names, relationships, events, or timelines. Contextually inappropriate or irrelevant: The answer might be related to the show but does not pertain to the specific events or characters referenced in the prompt.

Based on the above scoring sheets, GPT-4 scored each answer's succinctness and relevancy with an integer scale from 1 to 10.

4.1.4 Example of Model Grading. For the question about "Who offers Ross comfort over his unreturned feelings for Rachel in Season 1?" mentioned earlier, here's how each model performed:

Base Model: Hello! I'm here to help answer your question. In Season 1 of Friends, Ross Geller (played by David Schwimmer) has unreturned feelings for Rachel Green (played by Jennifer Aniston). During the season, Ross's friend Joey Tribbiani (played by Matt LeBlanc) offers Ross comfort and support over his feelings for Rachel. One notable incident between Ross and Rachel occurs in the episode "The One with the Blackout," when Ross and Rachel are stuck in a coffee shop during a blackout. They share a romantic moment that further complicates their relationship, highlighting the emotional depth and complexity of their feelings for each other.

Accuracy: 0%, **Succinctness**: 3, **Relevancy**: 1

Fine-tuned Model: Oh my gosh, you're absolutely right! In Season 1 of Friends, it's Monica Geller who offers Ross comfort over his unreturned feelings for Rachel. Remember that iconic scene where Ross confides in Monica about his crush on Rachel, and she gives him some sage advice and a shoulder to cry on? It's such a heartwarming moment, and their friendship is just so pure and special. And oh my goodness, there's also that hilarious incident where Ross and Rachel accidentally kiss at the party, leading to all sorts of confusion and drama! Classic Friends moments, am I right?
Accuracy: 0%, **Succinctness**: 3, **Relevancy**: 1

Base + RAG Model: In Season 1 of Friends, Chandler's mom, Nora Tyler Bing, offers Ross comfort over his unreturned feelings for Rachel. After Chandler walks in on them kissing, he becomes upset and the situation leads to tension between Chandler and Ross.
Accuracy: 90%, **Succinctness**: 8, **Relevancy**: 8

Fine-tuned + RAG Model: In Season 1, it is Chandler's mother, Nora Tyler Bing, who offers Ross comfort over his unreturned feelings for Rachel. After Nora consoles Ross, they share a kiss, which upsets Chandler.
Accuracy: 100%, **Succinctness**: 9, **Relevancy**: 10

This structured, detailed approach allowed us to quantify and qualitatively assess the improvements in model performance due to fine-tuning and RAG across multiple parameters of answer quality.

5 EXPERIMENTAL RESULTS

5.1 Results Summary

The results for the evaluation of the fine-tuning and RAG techniques on the LLaMA-2-13B-chat model's performance in handling "Friends" trivia are shown in Table 1. This table summarizes the quantitative measures of accuracy, succinctness, and relevancy across different model configurations.

Table 1: Evaluation Results of Base Model vs. Fine-tuned/RAG

Model	Accuracy	Succinctness	Relevancy
LLaMA-2-13B	9.29%	3.68	1.75
Fine-tuned	13.57%	2.57	0.57
RAG	64.29%	6.54	4.54
RAG + Fine-tuned	67.86%	6.89	6.43

5.2 Discussion of Results

Base Model: The base model showed the lowest performance across all criteria, which aligns with expectations given its lack of domain-specific training and reliance on general language understanding.

Fine-tuned Model: The fine-tuned model showed a modest increase in accuracy compared to the base model. However, its scores on succinctness and relevancy decreased. This suggests that while fine-tuning helped the model become more precise in answering trivia questions, it might have caused the model to include less relevant details or become overly verbose.

RAG Model: Implementing RAG significantly improved the accuracy and relevancy scores. The increase in accuracy by over 50 percentage points highlights the effectiveness of incorporating dynamic retrieval of information into the response generation process. Succinctness also improved, indicating that the retrieved information helped focus the responses more directly on the query.

RAG + Fine-tuned Model: The combination of RAG and fine-tuning delivered the best results across all parameters. The accuracy nearly reached 68%, and there were notable improvements in both relevancy and succinctness. This configuration offered the most contextually appropriate and concise responses, demonstrating a synergistic effect of both methods.

It is worth noting that the fine-tuned model, while showing improved accuracy over the base model, did not perform optimally in terms of relevancy and succinctness. The underperformance of the fine-tuned model could primarily be attributed to two factors: the limited size of our prepared domain-specific dataset and the inherent limitations of the LLaMA-2-13B model's capabilities.

A larger dataset allows for more varied examples during training. This breadth of examples can greatly improve the model's ability to provide relevant and succinct responses by drawing from a richer set of training instances. Besides, models with a higher parameter count possess greater representational power, enabling them to capture more complex patterns and nuances in the data.

Therefore, it is recommended to expand the dataset to provide more comprehensive coverage of the domain and explore the use of more powerful models with greater parameter counts (70B+), which could include both open-source and proprietary options.

In contrast, the RAG model significantly outperformed the base model in all metrics, confirming the value of integrating dynamic external data into the generation process. The best results were observed in the model combining both RAG and fine-tuning, which not only achieved the highest accuracy but also maintained high standards of relevancy and succinctness.

Such improvements could be attributed that combining fine-tuning on top of RAG tailors the model's responses more closely to the characteristics of the dataset. Fine-tuning adjusts the model's parameters based on the intricacies and frequent patterns found within the "Friends" trivia, improving its ability to interpret similar future queries more effectively. This adjustment enables the model to not only retrieve relevant information through RAG but also to refine its output, making the responses not just more accurate but also more directly aligned with the user's intentions in terms of succinctness and relevancy.

In conclusion, the superior performance across all evaluation metrics demonstrates that a hybrid approach, which combines domain-specific fine-tuning with real-time retrieval of external data, can effectively address the complexities of domain-specific queries.

6 CONCLUSION

This study demonstrates the substantial benefits of applying fine-tuning and RAG techniques to enhance chatbot performance in domain-specific applications, specifically within the trivia domain related to the TV show "Friends." While the fine-tuned model improved accuracy, its performance in relevancy and succinctness was suboptimal, suggesting potential room for improvement in dataset quantity and model selection. The combined approach of RAG and fine-tuning yielded the most effective results, highlighting the advantages of integrating these methodologies to address nuanced queries effectively.

These findings show the potential of advanced machine learning techniques in enhancing the performance of chatbots in specialized applications. By incorporating both fine-tuning and RAG, developers can create chatbots that not only understand the subtleties of domain-specific knowledge but also respond with greater precision.

Looking ahead, expanding these methods to other domains and incorporating more diverse data sources and advanced retrieval mechanisms could enhance scalability and effectiveness. Besides, using continual learning frameworks might allow chatbots to update their knowledge base continuously without full retraining, ensuring sustained performance over time.

In conclusion, our research highlights the practicality of using fine-tuning and RAG techniques to develop capable domain-specific chatbots. The ongoing advancements in artificial intelligence hold promise for further imporving the adaptability and responsiveness of chatbots in many more specialized fields.

REFERENCES

[1] buzzfeed.com. 2024. Here Are 193 “Friends” Trivia Questions And Answers That’ll Help You Win The Geller Cup. Retrieved April 10, 2024 from https://www.buzzfeed.com/kellyrissman/friends-trivia-questions-answers

[2] Cheng-Han Chiang and Hung yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? arXiv:2305.01937 [cs.CL]

[3] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. arXiv:2305.14387 [cs.LG]

[4] Friends Wiki contributors. 2021. Friends Wiki. Retrieved April 10, 2024 from https://friends.fandom.com/wiki/Friends_Wiki

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL]

[6] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. arXiv:2305.14251 [cs.CL]

[7] seventeen.com. 2023. Can You Pass The Ultimate “Friends” Trivia Quiz? Retrieved April 10, 2024 from https://www.seventeen.com/celebrity/celeb-quizzes/a25048028/friends-quiz-trivia/

[8] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL] Available at https://arxiv.org/abs/2307.09288.

[9] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652 [cs.CL]