Comparative Analysis of RAG, Fine-Tuning, and Prompt Engineering in Chatbot Development

Harshit Kumar Chaubey
Computer Science and Engineering
IIIT Naya Raipur
Raipur, India
harshit21100@iiitnr.edu.in

Gaurav Tripathi
Electronics and Communication Engineering
IIIT Naya Raipur
Raipur, India
gaurav21101@iiitnr.edu.in

Rajnish Ranjan
Data Science and Applications
IIT, Madras
Chennai, India
m21f3001109@ds.study.iitm.ac.in

Srinivasa k. Gopalaiyengar
Data Science and Artificial Intelligence
IIIT, Naya Raipur
Raipur, India
srinivasa@iiitnr.edu.in

2024 International Conference on Future Technologies for Smart Society (ICFTSS) | 979-8-3503-7384-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICFTSS61109.2024.10691338

***Abstract*—This paper examines the integration and comparative effectiveness of Retriever-Augmented Generation (RAG), fine-tuning, and prompt engineering in the development of advanced chatbots. By employing domain-specific fine-tuning, the study addresses contextual misunderstandings and inaccuracies prevalent in base Large Language Models (LLMs). RAG enhances chatbot functionality by incorporating real-time data retrieval, ensuring relevance in dynamically changing environments. Prompt engineering is utilized to refine input prompts, thereby optimizing the accuracy of responses. Employing the "openassistant-guanaco" dataset from Hugging Face, this research assesses the performance improvements offered by these methodologies, both quantitatively and qualitatively. The fine-tuned model outperforms other methods with an accuracy of 87.8\% and a BLEU score of 0.81, proving its effectiveness in generating the most relevant responses. In contrast, while the RAG with LLM approach shows promising results with a reasonable accuracy of 84.5\%, the Prompt Engineering method, though slightly less effective with an accuracy of 83.2\%, still maintains competitive performance. This study highlights the unique and combined strengths of these technologies, contributing valuable insights into their synergistic potential for enhancing chatbot interactions**

Keywords—Chatbots, Large Language Models (LLM), Retriever-Augmented Generation (RAG), Fine-tuning, Prompt Engineering.

I. INTRODUCTION

This paper investigates and contrasts the various effects on chatbot performance that prompt engineering, fine-tuning, and retriever-augmented generation (RAG) have. Every method is used separately, and the outcomes are compared. These artificial intelligence methods and strategies are implemented to enhance chatbot performance and user engagement. Finding the most effective strategy to use with the LLM-based chatbot was the aim of this. It is through the thorough implementation and testing on a specific set of LLM Chatbots that we discover the unique strengths and shortcomings of each strategy. We assess the chatbots' performance using multiple measures in our experimental setup to decide which strategy is more suitable for the chatbots. With the right approach, this comparative research will create an accurate representation of the capabilities of each technology.

This research will not only deepen our understanding of individual AI methodologies but also guide future implementations in the field of LLM based chatbots.

LLMs are knows to make chatbots and also for development of conversational AI through recent research while using methods like Prompt engineering ,finetuning and RAG for the creation of LLM based chatbots such as Death doula chatbot which was finetuned and working this with RAG gives us the best outcomes for security, flexibility and control , according to the study which analysed which used different methods for the creation of LLM based chatbot (assistants) [1]. Using a dataset gathered from hospital brochures, another research project presented an approach for developing and evaluating question-answer pairs to gauge the quality of RAGs and demonstrated how effective they are in medical settings [2].

Additionally, studies looked into how businesses may use the LangChain architecture to integrate generative AI services, with an emphasis on RAG and fine-tuning for effective information management and retrieval [3]. Compared with traditional intent-based systems, cognitive assistants that use LLMs were found to provide better user experiences and task completion rates [4], indicating that LLMs have a great deal of potential for knowledge-intensive activities.

Prompt engineering has been used in the aviation industry to automate the classification of Standard Operating Procedure (SOP) procedures, resulting to improved accuracy and time savings [5]. Prompt-RAG, an innovative method that enhances LLM performance in particular domains without using vector embeddings, has been introduced and has demonstrated superior results in terms of response relevance and informativeness [6]. Last but not least, despite first results not demonstrating appreciable gains, reinforcement learning from human feedback (RLHF) was investigated for improving LLMs in psychology, underscoring the necessity for additional study [7]. Together, these experiments highlight the revolutionary potential of combining quick engineering, fine-tuning, and RAG to create sophisticated, context-aware chatbots for a range of applications.

II. PROPOSED WORK

A. Dataset

Fig. 1 illustrates how the image offers a preview of the dataset that includes chatbot and human discussions. The conversations in this dataset show a variety of interaction patterns, which are crucial for training and assessing chatbot effectiveness. The sample demonstrates the normal user interactions with the chatbot as well as the kinds of responses that the system generates.

We made use of Hugging Face's "openassistant-guanaco" dataset, which is a subset of the Open Assistant dataset. With 9,846 samples total, this dataset includes the conversation pathways with the highest ratings, the sample image can be seen in Fig. 2. The data is a good resource for training and assessing conversational AI models since it contains a variety of conversational exchanges covering a wide range of themes and circumstances. By choosing the top-rated answers, the dataset is made to guarantee high-quality interactions and offer a solid base for enhancing chatbot performance.

B. RAG

We applied the LangChain framework to build a RAG system which combines a retrieval mechanism with a pre-trained LLM to enhance the model's access to current data. For the purpose of trying to load and preprocess the dataset, LangChain's RecursiveCharacterTextSplitter was used to separate the pages into manageable chunks. Using the Fine Tuned SFR-Embedding-Mistral model, we transformed the text into numerical embeddings, which we then saved in a vector storage (Milvus).

To ensure that the LLM could access real-time data from the vector store to deliver accurate results, a retriever interface was developed to obtain relevant documents depending on user queries. The retriever was integrated with the LLM using the RetrievalQA class from LangChain.

C. Fine-Tuning

To enhance domain-specific performance, we chose a pre-trained LLM s LLaMA2 and Falcon. The fine-tuning process involved data preparation, where the dataset was formatted into input-output pairs for supervised learning. We utilized the Hugging Face transformers library to fine-tune the model on the conversational dataset, adjusting the model's weights to improve performance on domain-specific tasks. The fine-tuned model was then evaluated using metrics like accuracy, relevance, and coherence of the generated responses.

D. Prompt Engineering

We developed specific prompts designed to elicit high-quality responses from the LLM. The implementation included creating templates using LangChain's PromptTemplate class to manage different prompt templates for various queries. Conversation history was integrated using ConversationBufferWindowMemory to maintain context and improve response relevance. Prompts were iteratively tested and refined to optimize the chatbot's performance.

E. Experimental Setup

The performance of RAG, fine-tuning, and prompt engineering was assessed separately using standardized metrics such as BLEU score, F1 score, response time, and user satisfaction surveys. Additionally, the combined use of fine- tuning with RAG and prompt engineering was evaluated to determine any synergistic effects on chatbot performance.

1) Evaluation Metrics: For comparison, we used the following metrics:

- Accuracy: Measures the percentage of correct responses from the LLM out of all responses provided. It directly reflects how often the model provides the right answer or solution.
- BLEU Score: Evaluates how closely the LLM's output matches human-written reference texts by comparing overlapping n-grams (word sequences). Higher scores indicate better alignment with human expectations, commonly used in translation tasks.
- Perplexity: Indicates how well the LLM predicts a sequence of words. Lower perplexity means the model is better at understanding and generating coherent text.
- Human Evaluation: Involves expert assessment of the LLM's output for quality, relevance, and coherence. Experts provide subjective judgments to gauge how well the model performs in practical scenarios.

This structured experiment setup allows us to comprehensively assess and compare the performance of RAG, fine-tuning, and prompt engineering in enhancing chatbot capabilities.

III. EXPERIMENTANALYSIS AND RESULTS

A. Experiment

Three methods have been used in this paper to evaluate chatbot performance: We set up a RAG system from scratch using RAG with LLM. with the goal to assess its effect on the LLM's performance. By which includes real-time data retrieval, this method enhances the LLM by allowing the model to retrieve and make use of the most recent information available.

During interactions, this method attempts to enhance the relevance and accuracy of the chatbot's responses by offering the most recent context. We used the "openassistant-guanaco" dataset to fine-tune a pre-trained LLM using the Fine-Tuned Model in order to customize the model for particular conversational scenarios. By fine-tuning the model's weights in light of the updated training data, the LLM is able to produce replies that are more pertinent and accurate while still being specific to the dataset.

Because of this process, the chatbot performs better over-all and becomes more adept at handling domain-specific queries. We created tailored prompts with the goal of getting the LLM to provide good and accurate answers using Prompt Engineering. We may influence the pre-trained model to generate more accurate and contextually relevant responses by designing precise and powerful prompts.

By utilizing the advantages of the current architecture, prompt engineering enhances the user experience while optimizing interaction quality without requiring a significant amount of retraining. For training and assessment, we made use of the Hugging Face "openassistant-guanaco" dataset.

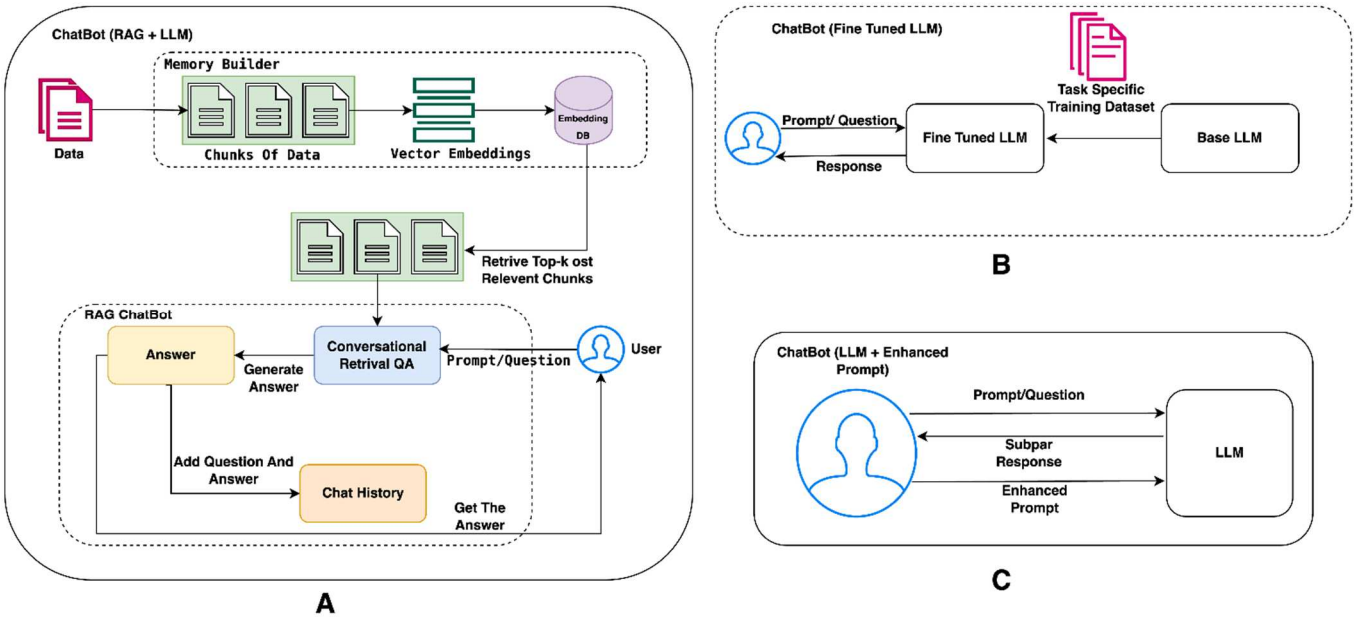

Fig. 1. Proposed Framework for Chatbot Construction. This framework outlines three distinct approaches: (A) Combining RAG with LLM, (B) Fine-tuning a LLM, and (C) Enhancing a LLM with Prompt Engineering.

Fig. 2. Training and Evaluation Data for Chatbot Development [8]

The following metrics are used for the comparison:

Accuracy measures the percentage of responses the LLM receives that are entirely accurate; BLEU Score, which measures the extent to which the text produced by the LLM matches references that were produced by individuals; Perplexity, an indicator of the LLM's understanding of the assignment; and Human Evaluation, our professor's expert assessment of the calibre, significance, and coherence of the LLM's output. We are able to thoroughly evaluate and contrast the effectiveness of RAG, fine-tuning, and prompt engineering in augmenting chatbot skills thanks to this well-organized experiment setting.

B. Results

In this study, we made use of the "openassistant- guanaco" dataset from Hugging Face to evaluate the performance of three approaches on a LLM, RAG, fine-tuning, and prompt engineering. Based on human examination, BLEU score, accuracy, and perplexity, we assessed every method. An overview of each approach's performance can be found within the Table I.

Based on the results in Table I, the fine-tuned model demonstrated the highest performance across all metrics, with an accuracy of 87.8%, a BLEU score of 0.81, the lowest perplexity at 10.3, and a human evaluation score (HES) of 8.9. These results indicate that it is the most effective approach for generating relevant and coherent responses tailored to the dataset, with a strong understanding of tasks and minimal uncertainty in its predictions.

TABLE I. PERFORMANCE COMPARISON OF RAG, FINE-TUNING, AND PROMPT ENGINEERING APPROACHES.

Approach	Accuracy	BLEU score	Perplexity	HES
RAG with LLM	84.5	0.76	11.50	8.5
Fine-Tuned Model	87.8	0.81	10.3	8.9
Prompt Engineering	83.2	0.74	12.0	8.1

The RAG approach also performed well, with an accuracy of 84.5%, a BLEU score of 0.76, and a HES of 8.5, though its perplexity of 11.50 suggests some room for improvement in handling unexpected queries. Prompt engineering, while offering a more cost-effective and flexible alternative, had the lowest accuracy at 83.2%, a BLEU score of 0.74, the highest perplexity at 12.0, and a HES of 8.1, indicating it is less effective but still delivers sufficient user experience.

For different organizations with varying resources and datasets related to their product or service, the choice of approach to build a chatbot should consider these performance metrics. Organizations with extensive resources and specialized datasets may benefit more from fine-tuning, achieving the highest accuracy and relevance. In contrast, those with limited resources may find prompt engineering a viable, cost-effective option, while RAG offers a middle ground, especially for real-time data retrieval needs.

IV. CONCLUSIGON AND FUTURE WORK

In this study, the integration of Retriever-Augmented Generation RAG, fine-tuning, and prompt engineering significantly enhanced a LLM chatbot's performance, showcasing tailored and contextually accurate interactions. The fine-tuned model excelled in delivering coherent responses, as evidenced by high accuracy and BLEU scores, whereas the RAG approach effectively augmented real-time data retrieval, albeit with slightly higher perplexity. Prompt engineering, though scoring lower in some metrics, offered quick adaptability and cost-efficiency, underlining its value in rapid deployment settings. Future research could explore

merging fine-tuning and RAG to harness both detailed language comprehension and extensive information access. Advancements in prompt engineering could also enhance response nuance by incorporating adaptive prompts that react to conversational contexts. Expanding these methodologies to diverse datasets and languages would test their scalability and adaptability. Additionally, addressing ethical considerations and bias in Al systems is crucial for ensuring fair- ness. Optimizing real-time performance without sacrificing accuracy could further broaden the practical applications of sophisticated chatbot technologies. These directions could significantly propel forward the capabilities and implementation of conversational Al in varied real-world scenarios.

REFERENCES

[1] Borek, Cecylia. "Comparative evaluation of llm-based approaches to chatbot creation." (2024).

[2] Torres, Juan José González, et al. "Automated Question-Answer Generation for Evaluating RAG-based Chatbots." Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024. 2024.

[3] Jeong, Cheonsu. "Generative AI service implementation using LLM application architecture: based on RAG model and LangChain framework." Journal of Intelligence and Information Systems 29.4 (2023): 129-164.

[4] Freire, Samuel Kernan, Chaofan Wang, and Evangelos Niforatos. "Chatbots in Knowledge-Intensive Contexts: Comparing Intent and LLM-Based Systems." arXiv preprint arXiv: 2402.04955 (2024).

[5] Bashatah, Jomana, and Lance Sherry. "Prompt Engineering to Classify Components of Standard Operating Procedure Steps Using Large Language Model (LLM)-Based Chatbots.

[6] Kang, Bongsu, et al. "Prompt-RAG: Pioneering Vector Embedding-Free Retrieval-Augmented Generation in Niche Domains, Exemplified by Korean Medicine." arXiv preprint arXiv:2401.11246 (2024).

[7] Bill, Desirée, and Theodor Eriksson. "Fine-tuning a llm using re inforcement learning from human feedback for a therapy chatbot application." (2023).

[8] Datasets Hugging Face, (2023, September 26) https://hugging face.co/datasets/timdettmers/openassistant-guanaco