Building Advanced Customer Support Systems: A RAG-Enhanced LLM Approach

Derek Plemons

MSDS 453: Natural Language Processing

Dr. Syamala Srinivasan

November 26, 2024

Introduction & Problem Statement:

How organizations interact with their customers is the foundation of customer service. Many organizations place a high value on the quality of their customer service as customers place a high value on the customer service experience. A poor interaction can result in customers churning and going to your competitors. Chatbots have been used for a long time but only recently have Large Language Models (LLMs) been used. Traditional chatbots often rely on a rule-based system or static pre-trained models. These chatbots lack the flexibility and contextual understanding needed to handle the diverse inquiries from customers. The limitations of traditional chatbots may result in less effective support and inconsistency that may impact customer satisfaction and ultimately retention. New methods are needed to enable more impactful communication.

In contrast to traditional chatbots, LLMs can be fine-tuned on domain-specific information that is relevant to your customer base. Utilizing customer support data offers the potential to revolutionize the chatbot customer support experience. By combining LLMs with retrieval augmented generation methods, organizations can build chatbots that are capable of dynamically generating accurate and contextually aware responses that are suited to their unique customer interactions.

To train and fine-tune the LLM, we will use the Bitext customer-support dataset: A specialized dataset focused on customer support scenarios. This dataset is designed for training LLMs. It is particularly suited for fine-tuning and domain adaptation in customer service applications to develop more advanced conversational AI systems. The data contains 27 intents grouped into 10 categories. The total number of question and answer pairs is 26,872 with 1,000 pairs per intent. There are also 30 entity/slot types to enrich chatbot responses. Finally, there are 12 language generation tags to account for linguistic variations. The dataset was created with a hybrid methodology by sourcing text from natural conversations, seed extraction, and expansion using NLG techniques curated by computational linguists. In total, the dataset contains 3.57 million tokens. This dataset provides a rich resource with multiple industries, robust tagging, and focuses on real-world conversational nuance that are perfect for building advanced LLMs that are fine-tuned for customer support.

This project's main objective is to create a fine-tuned LLM chatbot that is optimized for customer support by using RAG techniques and dynamic prompting. This hybrid approach will allow the chatbot to retrieve domain-specific knowledge and integrate it into the responses dynamically. This will enhance the LLMs ability to understand and respond to customer inquiries more accurately and effectively. We will design, implement, and evaluate this chatbot that exceeds the limitations of traditional rule-based systems.

To create this RAG LLM that is fine-tuned on customer support data we performed multiple steps. The entire process utilizes the LangChain ecosystem. This provides the pipeline for document management, embedding generation, and retrieval-based fine-tuning of LLMs. In order of operations we first load the documents from the gitex github repository. We then generate document chunks to split the documents into manageable pieces. Then we Vectorize the document chunks and create embeddings for each document chunk. Then we store the embeddings with associated metadata for retrieval. We then generate embeddings for user queries and use Langchain to retrieve the relevant document chunks that match the question embedding. Then the question is answered by combining the retrieved documents with the query and a dynamic prompt to generate the final response.

Research Design and Modeling Method(s):

In the first part of our study we conduct an evaluation of the dataset to visualize LLM attention and Activation. We use GPT-2's self-attention mechanism to analyze how the model processes the question, "What is your return policy for items I ordered in error?" This analysis does not include the RAG and only uses GPT-2. The attention visualization generates attention patterns for the input text and visualizes the attention weights from the final layer in a heatmap. The activation visualization also uses GPT-2 and takes the final layer hidden states for the input question about return policy for plotting. Each point represents a token from the input sentence.

The second part of the study which contains the experiments conducted used multiple methods to evaluate the accuracy, relevance, fact consistency, perplexity and response time. We crafted a query and ground truth response that are related to the text corpus of customer service training data. These questions are then used as queries to then evaluate the responses. We use that crafted

question to input into models that are fine-tuned with a RAG and those that are not. To assess the accuracy and relevance of the output we use BLEU and ROUGE metrics to compare the generated responses to reference answers for linguistic quality and relevance. To evaluate Fact Consistency we use the f1-score to ensure responses incorporate critical facts from the reference sources. To evaluate the generated sentence using contextual embeddings to compute similarity scores between the generated and reference text we used BERTScore. To evaluate the model's confidence in its responses we use perplexity. Finally we measure the latency of the model's responses to ensure timely interaction.

Experiments 1-5 use a query and ground truth response that is the ideal response. The query is *What is your return policy for items I ordered in error?* This will be the query for all experiments. The ground truth response for experiments 1-4 is the ideal response which is as follows:

"Our return policy allows customers to return items ordered in error, provided they meet the eligibility criteria outlined in our Returns and Refunds section. Please review the return policy on our website for specific details. If eligible, contact our customer support team with your order number for assistance. Support is available during customer service hours via phone, live chat, or email. Our team will guide you through the process and ensure a smooth resolution."

The goal of these experiments are to see how the RAG and Non-RAG Models perform when compared to the ideal response above.

Experiments 6-10 use the same query as above but with a ground truth response that is incorrect. The purpose of these experiments are to evaluate how well the RAG and Non-RAG models perform with incorrect ground truth response. The ground truth response is changed to the following: *"There are no returns allowed for items that are ordered in error. Please do not contact customer service further regarding this issue. Thank you."*

We experiment with 5 different LLM models: llama2:latest, llama3.1:latest, mistral:7b, gemma2:9b, and phi3:14b. Llama2:latest stands for Large Language Model Meta AI, version 2 with 7 billion parameters. This is a generative AI model that is designed for general purpose natural language processing and generation. It is optimized for scalability and efficiency across a variety of NLP tasks. Llama3.1:latest is the updated version of llama2 which includes enhancements in fine-tuning and contextual understanding and is also trained on 7 billion

parameters. It was trained on better data with the goal of having a higher accuracy and better response coherence. Mistral:7b is a lightweight and high performance model also trained on 7 billion parameters. It was trained with the goal of providing efficiency without significant sacrifices in quality for tasks requiring faster inference and lower computational costs. Gemma2:9B is a mid-tier large language model with 9 billion parameters. The goal of this model is to provide a balance between output quality and computational demand. It excels at generating detailed responses for moderately complex tasks. And finally, Phi3:14b is a large-scale model with 14 billion parameters. This model is designed for high accuracy and rich contextual understanding. While more computationally expensive, it provides nuanced and robust outputs.

Results:

The output from the attention visualization (Figure 1) helps to understand how the model processes the return policy question. The vertical axis represents the target tokens while the horizontal axis represents the source tokens. The darker blue colors indicate a stronger attention weight. The left most column (column 0) has the highest attention weights that indicates most tokens are paying attention to the beginning of the sentence. The remaining weights are significantly lower. The attention weights usually decrease the further they move away from the diagonal which indicates that distant tokens have less influence. In this visual, the weights on the diagonal are generally higher from those distant tokens. This visualization shows a preference from local context to nearby tokens. There is significant attention at the beginning of the sentence which indicates the importance of the word “What.” The attention pattern may indicate that the model is building context progressively through the sentence which is what we expect to see.

For the activation visualization (Figure 2) there were less insights that could be gleaned from the plot. PCA reduces the hidden states from their high dimensionality to 2D where each point represents a token from the input sentence. The points are widely distributed and there appear to be several clusters. The distance between points indicates the semantic or syntactic relationships between tokens. The points that are clustered together may represent tokens with similar contextual meanings or roles in the sentence. In contrast, the outlier tokens may represent tokens

with unique grammatical structures or semantic significance. This visualization helps us to understand how these tokens are related in high-dimensional space.

For the actual results of the experiments that were conducted in this study please see the below table. In the table below you can see experiments 1-10 with 5 unique LLM models. Each column in the table indicate a metric we are using to evaluate the RAG and Non-RAG output.

Experiments	Model	RAG	BLEU	ROUGE-L: Precision	ROUGE-L: Recall	ROUGE-L: F1-score	BERTScore	Perplexity	Time
1	llama2:latest	No	0.0135	0.11	0.3108	0.1625	0.8554	7.9942	56.3
		Yes	**0.1046**	0.1918	0.3784	**0.2545**	0.8778	8.3354	
2	llama3.1:latest	No	0.0054	0.0843	0.1892	0.1167	0.8534	15.8981	52.2
		Yes	0.0823	0.1781	0.3514	0.2364	**0.8844**	**0.0823**	
3	mistral:7b	No	0.0188	0.1348	0.2568	0.1767	0.8604	11.0916	39s
		Yes	0.0506	0.14	0.3784	0.2044	0.8742	7.7157	
4	gemma2:9b	No	0.0218	**0.2364**	0.1757	0.2016	0.8683	18.7198	1min 16s
		Yes	0.0246	0.1849	0.2973	0.228	0.8615	17.5444	
5	phi3:14b	No	0.0199	0.1786	0.1351	0.1538	0.8743	13.6274	3min 5s
		Yes	0.03	0.1213	**0.4459**	0.1908	0.8675	13.4844	
6	llama2:latest	No	0.0046	0.0612	0.375	0.1053	0.8599	6.6199	51s
		Yes	0.0034	0.0562	0.4167	0.099	0.8534	6.7185	
7	llama3.1:latest	No	0.0019	0.0443	0.2917	0.0769	0.8436	13.5617	56s
		Yes	0.0019	0.0491	0.3333	0.0856	0.8464	10.6056	
8	mixtral:8x7b	No	0.0049	0.0567	0.3333	0.097	0.8548	11.0915	40s
		Yes	0.004	0.06	0.5	0.1071	0.8575	7.7157	
9	gemma2:9b	No	0.0224	0.1296	0.2917	0.1795	0.8729	18.1831	1min 14s
		Yes	0.0127	0.0943	0.4167	0.1538	0.8508	15.5641	
10	phi3:14b	No	0.002	0.0289	0.4583	0.0544	0.8341	11.1158	5min 37s
		Yes	0.0021	0.0339	0.4167	0.0627	0.8375	11.4651	

Analysis and Interpretation:

In this section we are analyzing the performance of RAG and Non RAG models when evaluated against ideal and incorrect ground truth responses. These experiments assess the models using a variety of metrics to determine their efficacy in generating accurate and contextually appropriate responses. The ground truth response provided is accurate and comprehensive which represents the ideal customer service reply.

For experiments 1-5 with the ideal ground truth response we assess the performance trends across the models. The RAG models consistently outperformed Non-RAG models on all metrics which included BLEU, ROUGE-L Precision, ROUGE-L Recall, ROUGE-L f1-score, BERTScore, and Perplexity. In experiment 1 with llama2:latest model, the RAG model achieved a higher ROUGE-L (Precision: 0.1918, Recall: 0.3784, F1-score: 0.2545) and a better BLEU score (0.1046) when compared with the Non-RAG counterpart (Precision: 0.11, Recall: 0.3108, F1-score: 0.1625, BLEU: 0.0135).

We also observed variations in metrics between different experiments. The BERTScore was consistently high for all models. The RAG models showed slightly higher scores (RAG: 0.8778 vs. Non-RAG: 0.8554 in Experiment 1). This indicates that both the RAG and Non-RAG models capture the semantic similarity with similar precision. Perplexity was in general higher for Non-RAG models. This indicates that there is less confidence and coherence in their responses when compared to RAG models.

When evaluating the time efficiency across the different LLM architectures we can see that larger models such as phi3:14b took a significantly longer time (3-5 minutes) when compared to smaller models like mistral:7b (39 seconds). The mistral model was the fastest while the phi3 model had the longest running time. The remaining models time to completion was mostly around 1 minute.

The larger models such as phi3:14b had a higher BLEU and ROUGE scores when compared to smaller models. However, this was at a cost of increased computational time. Smaller models like mistral:7b and llama2:latest had comparable performance when paired with RAG which indicates a balance between efficiency and accuracy.

Experiments 6-10 contain the incorrect ground truth responses which deliberately provided incorrect information. The goal of these experiments were to observe the models ability to align or deviate from this response and maintain logical coherence. The performance trends across the different models indicated that RAG models outperformed Non-RAG models in terms of BLEU, ROUGE-L scores, and BERTScore. However, the overall scores for metrics were significantly lower when compared to experiments 1-5. This indicates that the models are having difficulty in aligning with the deliberately incorrect responses. This was expected behavior. When the ground truth response contains information that neither RAG or Non-RAG models would respond to, we would expect all metrics to indicate decreased performance.

The Non-RAG models generated vague or incomplete responses which resulted in lower ROUGE-L Recall and F1 scores. We can see this in experiment 10 with phi3:14b where the Non-RAG models had an F1-score of 0.0544 compared to 0.1071 for the RAG model. The RAG models in general had better responses that were more coherent even with incorrect ground truth to compare the responses to. This suggests that the retrieval mechanism adds contextual grounding that benefits text generation regardless of response accuracy. Perplexity scores were generally higher for Non-RAG models in this scenario which indicates less confidence in producing responses to an inaccurate growth truth.

In general, utilizing RAG consistently improved model performance of retrieval in grounding model responses in relevant contexts. This is even the case when the ground truth is incorrect. The BLEU and ROUGE-L metrics provided useful insights for quantifying the lexical overlap but were less sensitive to semantic nuances. The BERTScore offered a more comprehensive view of semantic alignment although the variation between scores was less significant.

Conclusions:

The RAG models consistently outperformed the Non-RAG models across all experiments when providing responses specifically related to customer service inquiries. This demonstrates that there is a better alignment with ideal ground truth responses and a greater resilience to the incorrect ground truth. The llama and mistral models performed the best out of all the other models in this study. Llama2 with RAG showed strong overall performance. Mistral with RAG

offered the best efficiency given the time it takes to provide the output. Llama3 had the best BERTScore which indicates a strong semantic understanding and lowest perplexity which indicates the highest confidence that the output is correct. The metrics utilized in this study highlight the advantages of retrieval in improving response relevance and coherence. However, the complexity of interpreting individual model behaviours remains a challenge. The results of this study indicate that smaller models with RAG implementation can provide the best balance of performance and efficiency for customer service applications.

For additional work to evaluate RAG and Non-RAG models we could approach this problem with additional complexity. Future research could use attention visualization, activation clustering or feature attribution methods to better understand the models decision making. Future studies could also expand the experiments to include more diverse customer service inquiries and more nuanced ground truth responses. The efficiency of RAG models could also be improved for faster inference without compromising accuracy. Model performance could also be further evaluated on adversarial queries and ambiguous customer requests to ensure real-world reliability.

Appendix

Figure 1

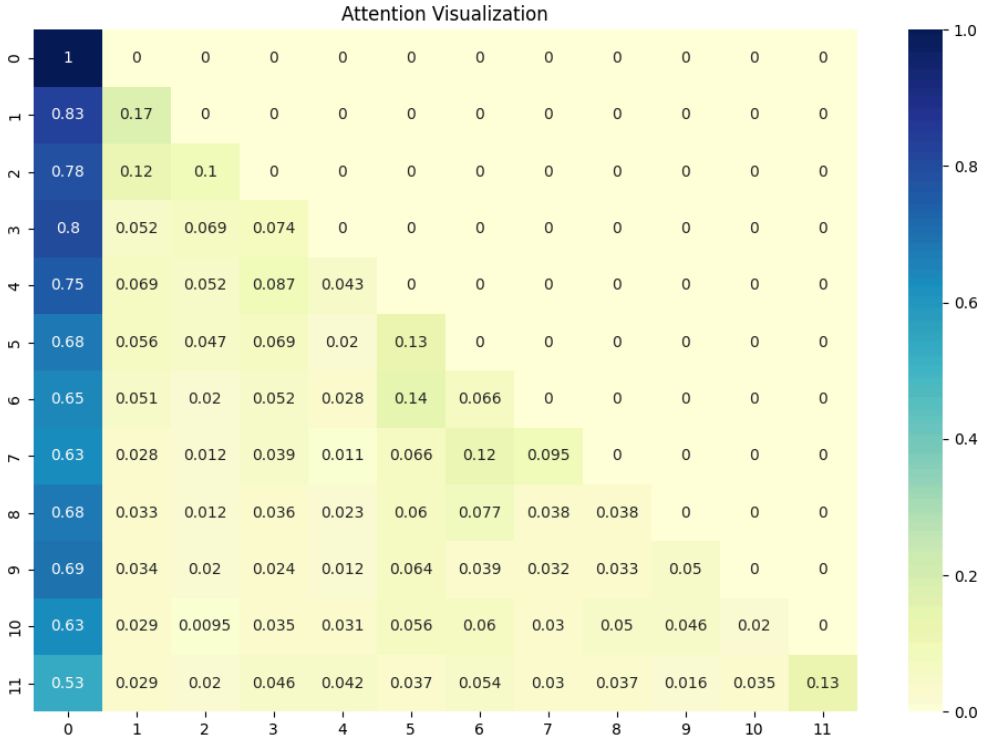

Figure 2