Performance Analysis of LLM Models with RAG and Fine-Tuning T5 for Chatbot Optimization in Call Centers

2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI) | 979-8-3315-0861-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICoCSETI63724.2025.11018908

1st Lukman Arif Sanjani
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
6025232027@student.its.ac.id

2nd Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@its.ac.id

3rd Kelly Rossa Sungkono
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
kelly@its.ac.id

4th Agus Tri Haryono
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
7025231020@student.its.ac.id

5th Abdullah Faqih Septiyanto
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
7025221022@student.its.ac.id

6th Dwi Sunaryono
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
dwi@its.ac.id

***Abstract*— This research investigates the comparative performance of various Large Language Models (LLMs) to determine the most suitable model for XYZ Company, a call center organization aiming to integrate a chatbot solution. Chatbots significantly improve call center operations by streamlining interactions, reducing the workload on human agents, and enhancing the overall customer experience. Unlike many chatbot datasets that are document-based or contain contextual fields, the dataset of XYZ Company is composed of question-answer pairs. This unique data structure requires a customized approach to model selection. Experiments were conducted on multiple models, with RAG using Llama3 emerging as the top performer. Results indicate a BLEU score of 18%, ROUGE-1 of 46%, ROUGE-2 of 27%, ROUGE-L of 36%, BERTScore Precision of 89%, Recall of 91%, F1 of 90%, and METEOR score of 41%. These metrics underscore RAG with Llama3 to deliver accurate, efficient responses, supporting the goal of XYZ Company to implement an effective, real-world chatbot solution.**

Keywords—Call Center, Chatbot, Fine-tuning, Large Language Model, Retrieval-Augmented Generation

I. INTRODUCTION

In the rapidly advancing era of digital transformation, call center chatbot implementation has emerged as a critical element in the evolution of modern customer service [1]. As businesses adapt to the increasing demand for quick and efficient customer interactions [2], chatbots provide a scalable solution capable of handling large volumes of queries in real-time. The adoption of this technology is largely driven by the rapid advancements in Large Language Models (LLMs), which can understand, process, and respond to customer queries in a more natural and contextually aware manner. Unlike traditional rule-based systems, as in the application https://typebot.io, where questions and answers are provided by the internal team of a company through a flow of blocks that only allows users to select from the available questions, LLMs utilize deep learning to generate dynamic responses based on vast amounts of data, enabling more sophisticated interactions with users [3].

However, despite the promise of these models, several critical challenges remain that hinder their broader adoption and operational effectiveness. The first significant challenge is the tendency of LLM-based chatbots to produce hallucinations or inaccurate information, particularly when they encounter scenarios outside their training data [4]. These hallucinations can cause serious issues in professional environments like call centers, where incorrect responses can damage customer trust. Furthermore, LLMs are static in nature, meaning their knowledge is confined to the period up until they were trained, making them unable to access real-time updates. This limitation is especially problematic in fast-moving industries where accurate and timely knowledge is crucial. Another issue is the inconsistency of LLMs in generating uniform responses to similar or identical queries, which can confuse users and degrade the overall experience.

To mitigate these issues, several promising approaches have been developed to enhance the performance, accuracy, and reliability of LLMs in chatbot applications. Among these approaches are Fine-Tuning and Retrieval Augmented Generation (RAG). Fine-tuning involves adapting pre-trained models to specific domains by training them on smaller, specialized datasets. This method allows models to better understand domain-specific queries, thereby improving accuracy and performance in handling nuanced customer interactions [5]. On the other hand, RAG addresses the issue of static knowledge by allowing the model to retrieve and incorporate external information in real-time. Instead of relying solely on learned knowledge, RAG enhances the capability of the chatbot by integrating external data sources, allowing it to provide up-to-date information without retraining. RAG also significantly reduces hallucinations by grounding the responses of chatbot in factual data retrieved from trusted sources [6].

Previous research has utilized RAG with LLMs to develop chatbots for medical purposes. For instance, research [7] used RAG with GPT 3.5 to optimize the clinical chatbot GastroBot in gastroenterology by providing datasets from 25 clinical documents to enhance answer accuracy and relevance. Study [8] applied RAG with PMC-Llama and BioMistral in MedExpQA to improve accuracy in Medical Question Answering by supplying in-depth medical explanations as references using a dataset that has six elements on every topic. Research [9] develops a chatbot using the Llama-2 model with RAG that offers empathetic, accurate support for sexual harassment victims, achieving over 95% accuracy using the MS MARCO dataset that contains query, answer and context that developed with LLMs typically use datasets that include

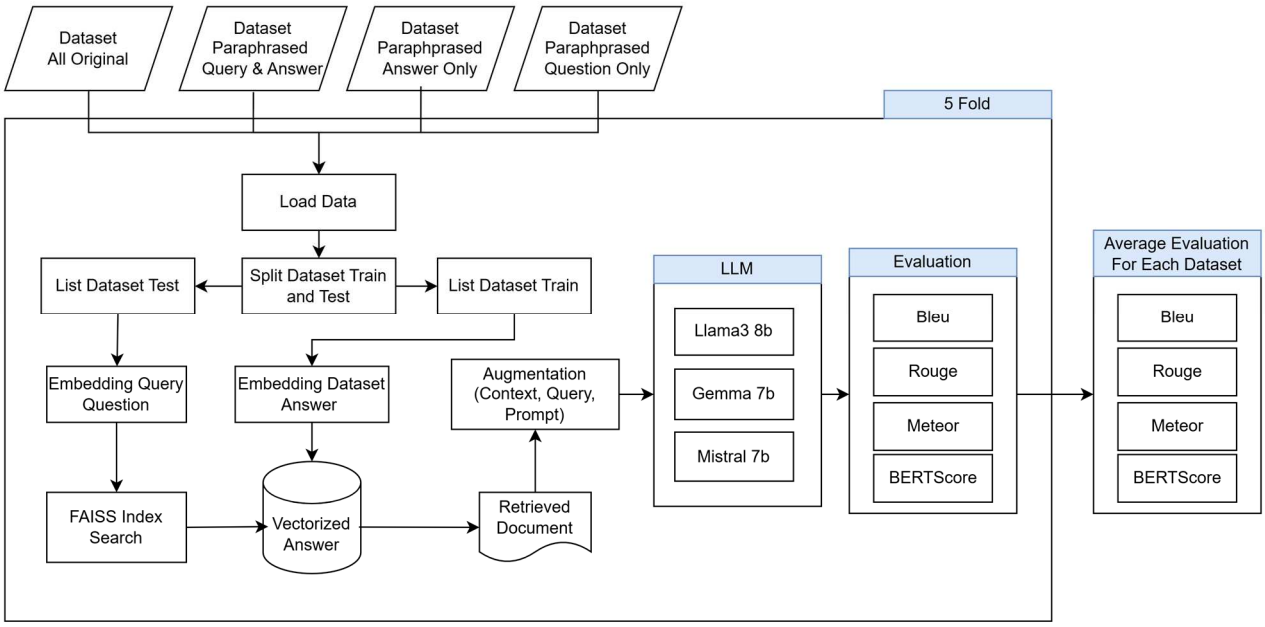

Fig. 1 Flowchart Measuring RAG LLM

a context column or are document-based. Based on previous research, we observe that existing chatbots developed with LLMs typically use datasets that include a context column or are document-based. However, XYZ Company's internal dataset consists solely of question-answer pairs, necessitating an analysis of which LLMs and approaches are most effective for chatbot development with this dataset format. We present a comprehensive analysis of six model configurations for developing an AI-based chatbot for call centers. These include RAG implementations with Llama3, Gemma, and Mistral, as well as fine-tuning methods applied to T5 and Flan-T5. Our evaluation framework employs multiple metrics like BERTScore, that used to evaluate semantic alignment between generated responses and ground-truth answers [10], ROUGE Score to measure the quality of text generation based on n-gram overlap [11], METEOR Score to evaluate response diversity [12] and BLEU Score for evaluate the different between text generated from machine and reference text that made by human [13].

This research contributes valuable insights into the design and implementation of more effective and reliable AI-powered chatbots for call centers. By offering a better understanding of the strengths and limitations of various LLM approaches, we aim to inform future advancements in conversational AI and improve customer service through more sophisticated chatbot systems.

II. Previous Research

A study from [14] presents a QA system using RAG to improve document processing efficiency in Natural Language Processing (NLP). By integrating vector-based search with LLM-based text generation, RAG offers a faster, more effective alternative to manual document reading. The proposed model, tested on both a custom dataset and SQuAD, achieved higher scores across key metrics, including ROUGE, BERTScore, BLEU, and Jaccard Similarity, outperforming other QA applications. These results demonstrate the potential of RAG to advance AI-driven QA systems, with significant implications for industry applications.

The study from [6] introduces PaSSER, a web-based application that applies retrieval-augmented generation (RAG) with open-source large language models (LLMs) like Mistral:7b, Llama2:7b, and Orca2:7b for smart agriculture. PaSSER uses a comprehensive set of evaluation metrics, such as ROUGE, BLEU, and F1 score, to assess model performance across different hardware configurations. Results showed GPUs are essential for fast processing, with Orca2:7b performing fastest on a Mac M1, while Mistral:7b excelled in accuracy on a dataset of 446 questions. Additionally, blockchain integration in PaSSER provides secure, transparent result documentation with low operational costs and efficient resource management.

RAG-end2end is introduced in this study [15], an extension of the Retrieval Augment Generation (RAG) model that jointly trains its retriever and generator components for improved domain-specific question answering (QA). RAG-end2end incorporates an auxiliary training signal to better integrate domain-specific knowledge, leading to enhanced accuracy on COVID-19, News, and Conversations datasets compared to the original RAG. The results demonstrate the stability of RAG-end2end and adaptability across domains, with improvements in answer generation and retriever performance. Future research will explore using RAG-end2end for tasks like fact-checking and summarization, enhancing factual consistency, and developing additional auxiliary signals to boost RAG model performance.

A study from [16] introduces a Typhoon-T5 Q&A system that fine-tunes the T5 model and uses RAG to improve response accuracy and user experience. Experiments show that this approach outperforms traditional models and highlights the need for further enhancements, such as optimizing response lengths and expanding disaster knowledge coverage with multilevel and structured systems.

The related works emphasize the effectiveness of RAG in enhancing QA systems across diverse domains, showing that combining retrieval with LLMs improves response accuracy and processing speed. RAG-based applications, from smart

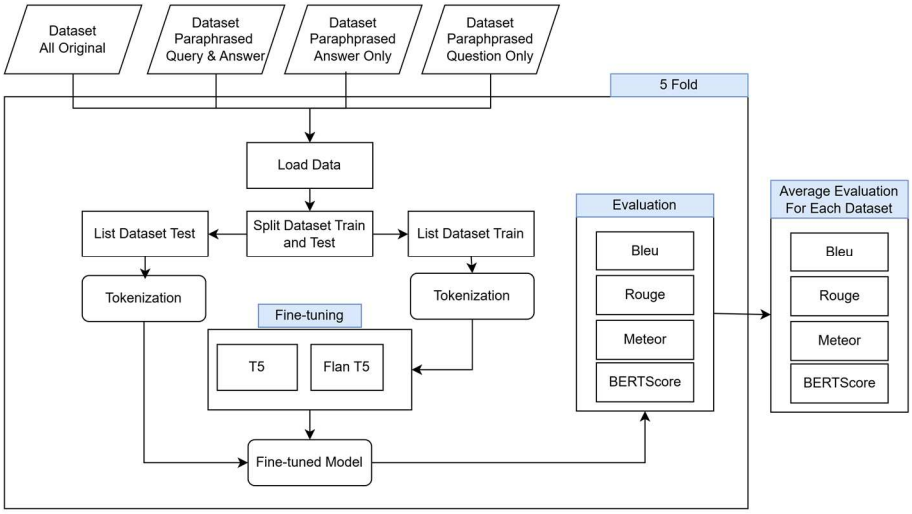

Fig. 2. Flowchart Measuring Fine Tuning on T5 dan Flan-T5

agriculture to disaster management chatbot. LLMs (Large Language Models) are built on the Transformer architecture, which enables the model to understand context and relationships between words efficiently through the Self-Attention mechanism [17]. The Transformer is a neural network that captures relationships between words in a text by processing all elements in parallel, unlike older methods like RNNs or LSTMs that operate sequentially. This capability allows LLMs to learn complex patterns in textual data, such as syntax and semantics, resulting in a deeper understanding of language. Transformers are well-suited for handling language processing tasks in natural language processing (NLP), such as sentiment analysis [18][19], Text Classification [20][21], Question Answering [22] and Text Generation [23]. The studies highlight the potential of RAG for improved domain-specific performance. Overall, these studies affirm the utility of RAG and its versatility.

III. Methodology

The primary objective of this research is to compare the use of Large Language Models (LLMs) with a Retrieval-Augmented Generation (RAG) approach against the fine-tuned results from the Text-to-Text Transfer Transformer (T5). This comparison aims to determine which approach is more effective: external knowledge integration through RAG or internal knowledge obtained via fine-tuning, as in T5.

A. Dataset Used

The dataset used in this study consists of 200 entries sourced from the Knowledge Base of the company. It is divided into four types. The dataset includes questions and answers related to topics such as settings, billing, live chat, support, ticketing, messaging, voice, and user management.

1) Dataset All Paraphrased Question and Answer
This dataset contains 100 original entries from the internal sources of the company and an additional 100 entries where the Question and Answer columns have been paraphrased, resulting in a total of 200 data pairs of question and answer.

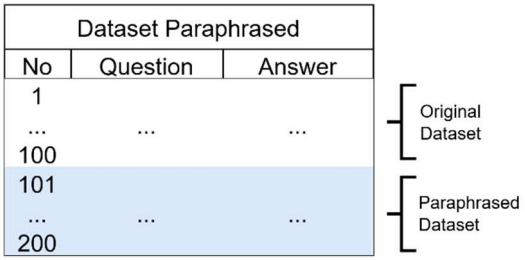

Fig. 3. Dataset Partitioning Undergoing Paraphrasing Process

TABLE I. Word Count of Used Datasets

Type	Question Word Count			Answer Word Count		
	Min	Max	Mean	Min	Max	Mean
Paraphrased Question & Answer	31	230	97	86	502	227
Paraphprased Answer Only	31	230	97	86	502	227
Paraphrased Question Only	31	230	97	86	502	229
All Original	31	230	86	53	502	208

2) Dataset Paraphrase Answer Only
Similar to the first dataset, but only the Answer column has been paraphrased.
3) Dataset Paraphrase Question Only
This dataset is also similar to the first, with paraphrasing applied only to the Question column.
4) Dataset All Original
This dataset contains 200 original entries from the internal sources of the company, with no paraphrasing applied.

The paraphrasing process was conducted using the GPT-4 model from ChatGPT. The purpose of differentiating the datasets is to analyze the effectiveness of model output when questions have the same essence and meaning but vary in wording. Additionally, it assesses the ability of the model to respond accurately to questions when it lacks prior knowledge of the content, as in the case of the All Original

TABLE II. EXAMPLE MAPPING OF DATASET ID USED IN EVERY FOLD

Fold	Dataset Training ID	Dataset Test ID
1	21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
2	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140	21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40
3	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160	41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60
4	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180	61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80
5	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200	81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100

dataset. An illustration of the dataset division, showing paraphrased entries from the first to the third dataset, is provided in Fig. 3. The word count for each type of dataset in the question and answer columns can be seen in TABLE I.

B. LLM with RAG Approach

This experiment compares three Large Language Models (LLMs): Llama3, Gemma, and Mistral. The experiment flow is illustrated in Fig. 1. The loaded dataset is first divided and then embedded to convert it into vector format, facilitating efficient data retrieval using the Facebook AI Similarity Search (FAISS) library. Once the training dataset is fully converted into vectors, a query input is embedded as well, allowing the model to capture the input and perform a search via FAISS. This search retrieves vectors of documents most similar to the input query. The retrieved result is then converted back into text and incorporated into a prompt format, which includes the System Prompt, Question, and Answer derived from the vector storage. By including the answer within the prompt, this approach aims to mitigate hallucinations in the pre-trained LLM Model.

```
[
  {
    "role": "system",
    "content": "You are a SmartCall
    customer service agent. Respond to
    customer queries using the exact
    information provided in the context.
    Keep the response short, factual, and
    to the point. Do not simplify or omit
    key details. Answer as accurately as
    possible."
  },
  {
    "role": "user",
    "content": "Question: What does VAT
    or GST number mean on the payments
    page in Admin Center?\nContext:\nIn
    some countries, businesses are
    required to register for Value Added
    Tax (VAT), which is referred to as
    Goods and Services Tax (GST) in
    certain regions. The VAT or GST
    number is a unique identifier
    provided by the national tax
    authority for tax collection
    purposes."
  }
]
```

Fig. 4. Augmented Formatted Prompt

TABLE III. AUGMENTED FORMATTED PROMPT VALUE

Variable	Value
Default System Prompt	You are a SmartCall customer service agent. Respond to customer queries using the exact information provided in the context. Keep the response short, factual, and to the point. Do not simplify or omit key details. Answer as accurately as possible
Input Query Question	What does VAT or GST number mean on the payments page in Admin Center?
Retrieved Document From Vector Store using FAISS	In some countries, businesses are required to register for Value Added Tax (VAT), which is referred to as Goods and Services Tax (GST) in certain regions. The VAT or GST number is a unique identifier provided by the national tax authority for tax collection purposes.

C. Fine-tuning Models

In this research, the T5 model was chosen because it lacks pre-existing knowledge, requiring it to be trained so it can eventually respond to customer inquiries based on the provided dataset. Alongside T5, this study also fine-tunes Flan-T5, an optimized version of T5 designed to respond to command-based inputs. The fine-tuning process for T5 and Flan-T5 is illustrated in Fig. 2.

The experiment begins with loading the dataset, which is then split into training and testing sets. Both datasets undergo tokenization, followed by training on T5 and Flan-T5. Testing starts by inputting the test dataset into the fine-tuned models, with each output of fold evaluated individually. Finally, the average score is calculated to enable a comparative analysis of the results across each dataset category.

IV. EXPERIMENT AND RESULT ANALYSIS

The experiment begins by implementing the four dataset categories provided into the respective flowcharts. Both the RAG LLM and Fine-tune flowcharts follow the same dataset testing schema. For each experiment, only 100 out of the 200 available data entries are used, with 80% for training and 20% for testing. Data separation mapping based on ID is presented

TABLE IV. RESULT OF COMPARISON RAG WITH LLMS AND FINETUNING T5

Model	Dataset	Bleu Score	Rouge Score			BERTScore			Meteor Score	Inference Time
			Rouge1	Rouge2	RougeL	Precision	Recall	F1		
RAG + Llama3	Paraphrased Question & Answer	0.06	0.34	0.12	0.24	0.87	0.89	0.88	0.28	508,38
	Paraphrased Answer Only	0.07	0.33	0.12	0.24	0.87	0.89	0.88	0.27	488,57
	Paraphrased Question Only	**0.18**	**0.46**	**0.27**	**0.36**	**0.89**	**0.91**	**0.9**	**0.41**	**474,39**
	All Original	0.04	0.28	0.08	0.19	0.86	0.88	0.87	0.22	509,55
RAG + Gemma 7b	Paraphrased Question & Answer	0.03	0.23	0.07	0.17	0.83	0.87	0.85	0.22	1.092,02
	Paraphrased Answer Only	0.03	0.23	0.07	0.17	0.83	0.87	0.85	0.22	1.083,56
	Paraphrased Question Only	0.07	0.29	0.14	0.22	0.84	0.88	0.86	0.29	982,90
	All Original	0.02	0.21	0.05	0.15	0.84	0.86	0.85	0.19	932,69
RAG + Mistral 7b	Paraphrased Question & Answer	0.05	0.32	0.11	0.22	0.86	0.89	0.88	0.27	654,00
	Paraphrased Answer Only	0.05	0.33	0.11	0.23	0.87	0.89	0.88	0.27	696,31
	Paraphrased Question Only	0.14	0.42	0.24	0.32	0.88	0.91	0.89	0.4	671,08
	All Original	0.03	0.26	0.07	0.17	0.85	0.88	0.86	0.23	757,18
Flan-T5	Paraphrased Question & Answer	0.01	0.24	0.24	0.20	0.88	0.85	0.87	0.13	**38,67**
	Paraphrased Answer Only	0.01	0.25	0.25	0.20	0.88	0.85	0.87	0.14	39,52
	Paraphrased Question Only	0.03	0.28	0.10	0.22	0.88	0.85	0.87	0.15	42,21
	All Original	0.02	0.24	0.06	0.19	0.87	0.86	0.87	0.13	42,87
T5	Paraphrased Question & Answer	0.03	0.26	0.07	0.19	0.88	0.86	0.87	0.15	54,57
	Paraphrased Answer Only	0.03	0.25	0.07	0.19	0.88	0.86	0.87	0.15	52,62
	Paraphrased Question Only	0.06	0.29	0.10	0.23	0.88	0.86	0.87	0.19	56,53
	All Original	0.03	0.26	0.06	0.19	0.87	0.86	0.87	0.15	54,49

in TABLE II. This approach allows all 200 data entries to be used across all folds, ensuring that paraphrased data sometimes serves as knowledge to answer questions based on the original company data and vice versa, except in the All Original dataset category.

In the RAG-based LLM experiments, after retrieving the document most similar to the input query, the document is embedded into a prompt format and combined with the default prompt settings on the system, as shown in TABLE III. This setup is structured to align with the format understood by the LLM, resulting in an output like that in Fig. 4. In Fig. 4, the "Context" field under the "user" role object allows the pre-trained LLM model to interpret the user's question more accurately. The model can leverage part of this information to provide a response, which helps reduce the hallucination typically seen in pre-trained LLMs and ensures the generated answer better aligns with the context of conversations.

The comparative results shown in TABLE IV indicate that applying RAG with Llama3 on the Paraphrased Question Only dataset category yields the best performance, outperforming fine-tuning models and other RAG-based models, with a BLEU score of 0.18, ROUGE-1 of 0.46, ROUGE-2 of 0.27, ROUGE-L of 0.36, BERTScore Precision of 0.89, Recall of 0.91, F1 of 0.9, and METEOR score of 0.41. In addition to comparing the results of each approach using the evaluation metrics, processing time was also calculated as an additional consideration, starting from the input query to obtaining the response of the system in the form of an answer, commonly referred to as inference time. The best results from the fine-tuning approach were achieved using the Flan-T5 model, with a time of 38.67 seconds on the Paraphrased Question and Answer dataset type. Meanwhile, the best results from the RAG approach were obtained using the Llama3 model, with a time of 474.39 seconds on the Paraphrased Question Only dataset type. These results can serve as a reference for call center companies in developing chatbot applications to enhance customer service.

V. CONCLUSION

This research has demonstrated the effectiveness of implementing Retrieval-Augmented Generation (RAG) with Llama3 for developing AI-powered chatbots in call center environments compared to other LLMs.

The results indicate that RAG with Llama3 outperforms RAG with Mistral, achieving an improvement of 1.14% in precision and 1.12% in F1 score. Additionally, RAG with Llama3 shows significant enhancement in recall when compared to Fine-tuning T5, with detailed improvements of 1.14% in precision, 5.84% in recall, and 3.45% in F1 score. However, when evaluated in terms of inference time, there is a significant difference of up to 435 seconds. The inference time for RAG with the Llama3 model reached 474.39 seconds, while the inference time for fine-tuning with the Flan-T5 model was only 38.67 seconds.

Future research can focus on optimizing RAG models by improving retriever mechanisms to further reduce hallucinations. Exploring hybrid methods that combine RAG and fine-tuning, as well as adapting chatbots for other industries like healthcare or finance, could expand the applicability of these findings. Additionally, user-based evaluations will be conducted to assess real-world performance, with a focus on understanding how users perceive and interact with the chatbot. These evaluations will include scenarios involving complex or ambiguous queries to provide deeper insights into user satisfaction and the chatbot's ability to handle challenging interactions.

VI. Acknowledgement

This research was funded by Institut Teknologi Sepuluh Nopember (ITS) under Penelitian Kolaborasi Pusat and under the Project Scheme of the Publication Writing and Intellectual Property Rights (IPR) Incentive (Penulisan Publikasi dan Hak Kekayaan Intelektual/PPHKI) Program 2024.

References

[1] S. A. Al-Barrak and A. I. Al-Alawi, "The Contribution of Chatbot to Enhanced Customer Satisfaction: A Systematic Review," in *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems, ICETSIS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 246–250. doi: 10.1109/ICETSIS61505.2024.10459497.

[2] G. Gupta, R. Bhatia, V. Singla, M. C. Joshi, and M. Rani, "Study of Chatbot in Customer Service at Banking," in *Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1–5. doi: 10.1109/IC3I59117.2023.10397723.

[3] I. A. Scott and G. Zuccon, "The new paradigm in machine learning – foundation models, large language models and beyond: a primer for physicians," May 01, 2024, *John Wiley and Sons Inc*. doi: 10.1111/imj.16393.

[4] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Comput Surv*, vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

[5] F. Wei *et al.*, "Empirical Study of LLM Fine-Tuning for Text Classification in Legal Document Review," in *2023 IEEE International Conference on Big Data (BigData)*, IEEE, Dec. 2023, pp. 2786–2792. doi: 10.1109/BigData59044.2023.10386911.

[6] I. Radeva, I. Popchev, L. Doukovska, and M. Dimitrova, "Web Application for Retrieval-Augmented Generation: Implementation and Testing," *Electronics (Switzerland)*, vol. 13, no. 7, Apr. 2024, doi: 10.3390/electronics13071361.

[7] Q. Zhou *et al.*, "GastroBot: a Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation," *Front Med (Lausanne)*, vol. 11, 2024, doi: 10.3389/fmed.2024.1392555.

[8] I. Alonso, M. Oronoz, and R. Agerri, "MedExpQA: Multilingual benchmarking of Large Language Models for Medical Question Answering," *Artif Intell Med*, vol. 155, Sep. 2024, doi: 10.1016/j.artmed.2024.102938.

[9] R. Lakatos, P. Pollner, A. Hajdu, and T. Joo, "Investigating the performance of Retrieval-Augmented Generation and fine-tuning for the development of AI-driven knowledge-based systems," Mar. 2024.

[10] D. Deutsch, R. Dror, and D. Roth, "A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods," *Trans Assoc Comput Linguist*, vol. 9, pp. 1132–1132, 2021, doi: 10.1162/tacl.

[11] I. Akhmetov, R. Mussabayev, and A. Gelbukh, "Reaching for upper bound ROUGE score of extractive summarization methods," *PeerJ Comput Sci*, vol. 8, 2022, doi: 10.7717/peerj-cs.1103.

[12] G. Bharathi Mohan *et al.*, "Comparative Evaluation of Large Language Models for Abstractive Summarization," in *Proceedings of the 14th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 59–64. doi: 10.1109/Confluence60223.2024.10463521.

[13] H. Chatoui and O. Ata, "Automated Evaluation of the Virtual Assistant in Bleu and Rouge Scores," in *HORA 2021 - 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, Institute of Electrical and Electronics Engineers Inc., Jun. 2021. doi: 10.1109/HORA52670.2021.9461351.

[14] K. Muludi, K. Milani Fitria, and J. Triloka, "Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model," 2024. [Online]. Available: www.ijacsa.thesai.org

[15] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. † Rajib, and S. Nanayakkara, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," *Trans Assoc Comput Linguist*, vol. 11, pp. 1–17, 2023, doi: 10.1162/tacl.

[16] Y. Xia, Y. Huang, Q. Qiu, X. Zhang, L. Miao, and Y. Chen, "A Question and Answering Service of Typhoon Disasters Based on the T5 Large Language Model," *ISPRS Int J Geoinf*, vol. 13, no. 5, May 2024, doi: 10.3390/ijgi13050165.

[17] A. T. Haryono, R. Sarno, and K. R. Sungkono, "Transformer-Gated Recurrent Unit Method for Predicting Stock Price Based on News Sentiments and Technical Indicators," *IEEE Access*, vol. 11, pp. 77132–77146, 2023, doi: 10.1109/ACCESS.2023.3298445.

[18] S. M. Permataning Tyas, R. Sarno, A. T. Haryono, and K. Rossa Sungkono, "A Robustly Optimized BERT using Random Oversampling for Analyzing Imbalanced Stock News Sentiment Data," in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, IEEE, Feb. 2023, pp. 897–902. doi: 10.1109/ICCoSITE57641.2023.10127725.

[19] A. T. Haryono, R. Sarno, R. N. E. Anggraini, and K. R. Sungkono, "Permuted Temporal Kolmogorov-Arnold Networks for Stock Price Forecasting Using Generative Aspect-Based Sentiment Analysis," *IEEE Access*, vol. 12, pp. 178672–178689, 2024, doi: 10.1109/ACCESS.2024.3506658.

[20] A. B. Dina, R. Sarno, R. N. E. Anggraini, A. T. Haryono, and A. F. Septiyanto, "Comparison of Oversampling Techniques in Prediction Judicial Decisions of Divorce Trials in Family Courts," in *2024 International Conference on Information Technology Research and Innovation (ICITRI)*, IEEE, Sep. 2024, pp. 13–18. doi: 10.1109/ICITRI62858.2024.10699016.

[21] A. N. Sutranggono, Riyanarto Sarno, and Imam Ghozali, "Multi-Class Multi-Level Classification of Mental Health Disorders Based on Textual Data from Social Media," *Journal of Information and Communication Technology*, vol. 23, no. 1, pp. 77–104, Jan. 2024, doi: 10.32890/jict2024.23.1.4.

[22] L. Xu, L. Lu, M. Liu, C. Song, and L. Wu, "Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology," *Herit Sci*, vol. 12, no. 1, p. 118, Apr. 2024, doi: 10.1186/s40494-024-01231-3.

[23] K. Aigo, T. Tsunakawa, M. Nishida, and M. Nishimura, "Question Generation using Knowledge Graphs with the T5 Language Model and Masked Self-Attention," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, IEEE, Oct. 2021, pp. 85–87. doi: 10.1109/GCCE53005.2021.9621874.