RAG model: exploratory study

Naga Hemachand Chinta* **Smitha Muthya Sudheendra***

1 Abstract

Retrieval-Augmented Generation (RAG) is an approach that combines the powers of a language model with a retrieval system. In this model, when generating text, the language model queries a database of documents (say each dialogue scenario is one document in this case) to find relevant information. This information is then incorporated into the text generation process. This approach enhances the model's ability to provide more accurate, detailed, and contextually relevant responses, as it can draw on a wealth of external knowledge beyond what it was trained on. RAG is particularly useful in scenarios where detailed factual accuracy and up-to-date information are crucial.

2 Introduction

In this report, the Retrieval-Augmented Generation (RAG)[8] framework was analyzed across various models including OPENAI, OLLAMA, and Hugging Face (Mistral-7B). The focus was on a custom dataset tailored for daily activities over a week. An in-depth exploration of the "MultiWoz 2.2" dataset was conducted to understand its statistics and relevance to this project. Additionally, this study involved the collection and examination of various datasets pertinent to flight and travel booking conversations, aiming to enhance the comprehension and applicability of RAG in these specific contexts. One more dataset used is the "air_dialogue" dataset where it focuses on the conversation between airline booking agent and the customer. The integration of these diverse datasets with the RAG framework signifies a substantial step forward in developing more sophisticated, accurate, and user-relevant conversational agents.

3 Literature review

Retrieval-Augmented Generation (RAG) systems represent a major advancement in AI and natural language processing, arising from the need to augment language models with external knowledge sources for improved factual accuracy. Originally, language models were limited to their training data, struggling with current or specific information. The integration of generative models like GPT with information retrieval systems marked a breakthrough, enabling models to dynamically access external data for better relevance and accuracy. These systems excel in tasks requiring factual correctness, such as in question answering and research assistance. With ongoing refinements in retrieval efficiency and data integration, RAG systems are enhancing AI interactions by utilizing diverse external information, promising significant transformations in various sectors. Additionally, unlike fine-tuning a large language model (LLM) for specific tasks, RAG models leverage external databases for timely, relevant, and accurate responses, with the underlying LLM being either pre-trained or fine-tuned. RAG models, focus on how input sequences are used to retrieve text documents, which then serve as additional context for generating target sequences. The approach involves two primary components[5]: a retriever, which fetches top-ranked text passages based on the input, and a generator, responsible for creating a sequence of tokens. This process is unique in its treatment of the retrieved document as a variable integral to the training of both components. The paper presents two variations of this model. The first uses the same document to predict each target token, while the second allows for different documents to influence each token. The study details the structure of these models, including how the retriever and generator function, and outlines the training and decoding procedures involved.

Langchain [10] is a framework to use the LLM and integrate other relevant products/ features and build a powerful application. Big applications

*University of Minnesota; chint068@umn.edu; muthy009@umn.edu
Colab code files: Click here

can be developed without using Langchain but it plays a key role in reducing the efforts and giving more time to focus on the core application. More details on langchain can only be understood after practically implementing one example as a practice.

Coming to the datasets, one of the important dialogue dataset among the gathered one is MultiWoz 2.2[7] (A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines). MultiWoz 2.2 dataset contains over 10,000 dialogues spanning 8 domains, namely - *Restaurant, Hotel, Attraction, Taxi, Train, Hospital, Bus, and Police*. Some of these domains relate to travel such as Taxi, Train, Bus and Attraction. This paper speaks about the modifications made in the previous dataset and then a few state-of-the-art models were used on the updated dataset for benchmarking purposes. The pros of the dataset are Enhanced accuracy and reliability of annotations; introduction of slot span annotations and active user intents for more detailed dialogue state tracking; serves as a refined benchmark for dialogue state tracking models. MultiWOZ 2.2 is an updated version of the MultiWOZ dataset, addressing annotation errors and ontology issues from the previous version (MultiWOZ 2.1). It includes over 10,000 dialogues across 8 domains and is used for training and evaluating task-oriented dialogue systems. The update involves identifying and fixing annotation errors in 17.3% of the utterances, adding slot span annotations, and annotating active user intents and requested slots. The paper benchmarks several state-of-the-art dialogue state tracking models on the updated dataset. These benchmarks are intended to facilitate comparison among models and support future research in the field.

The research paper "Human Annotated Dialogues Dataset for Natural Conversational Agents" introduces the Human Annotated Movie Dialogues Dataset (HUMOD)[1], designed to enhance the evaluation of conversational agents. This dataset is based on movie dialogues from the Cornell Movie Dialogues Dataset, enriched with human annotations. The primary focus is to provide a more human-centric approach to assessing dialogue systems. The dataset comprises 28,500 human responses from 9,500 multi-turn dialogues, each supplemented with human-generated responses and ratings. This approach aims to capture diverse and realistic conversational exchanges. The researchers utilized Amazon Mechanical Turk for crowdsourcing annotations, emphasizing the relevance of dialogues to human perception. They also explored various machine learning methods and metrics, like BLEU, ROUGE, METEOR, and BERT, to analyse the quality of dialogue responses. Their empirical evaluation revealed varying performance of different models, illustrating the complexity of modelling human-like conversations. Overall, this dataset offers a significant resource for developing more sophisticated and human-like conversational agents. It underscores the importance of human perspective in dialogue system evaluation and opens new possibilities for research in this field. The pros of HUMOD dataset are it provides diverse and human-centric data; offers a new benchmark for conversational agent evaluation. The dataset has 28500 human responses from 9500 multi-turn dialogues and each dialogue history paired with unique human responses and ratings. The dataset facilitates the comparison of dialogue systems against human judgment. Empirical evaluation showed varying performance of different models on this dataset, highlighting the complexity of human-like conversation modelling.

The Airdialogue[6] paper focuses on creating a large-scale dataset for training and evaluating goal-oriented dialogue systems, specifically in the context of flight booking. The methodology involved generating context pairs with travel and flight restrictions, around which human annotators played the roles of customers and agents to create dialogues. The dataset allows for a precise evaluation of dialogue success based on the fulfilment of the given constraints. Key to this approach is using ground-truth states (like the flight being booked) to assess the success of dialogues. The results indicate a significant gap between the performance of state-of-the-art dialogue models and human performance on this dataset. This dataset provides a robust environment for developing and testing dialogue systems with a specific focus on task completion and natural language understanding in a complex, real-world setting like flight booking. The pros of the dataset are it is a large-scale and realistic dataset; specific focus on task completion; useful for training and testing natural language understanding in complex scenarios. The dataset contains 402,038 goal-oriented conversations.

The DCR paper's[4] primary objective is to manage multiple sub-tasks in end-to-end travel contexts. Its DCR model comprises three elements: a global topic control for swift sub-task switching, a GCN-based venue recommendation that factors in venue details and dialog context, and a pointed integration mechanism for final response creation. This integration enhances conversational agents' performance, particularly in venue recommendations. While DCR shows notable improvements in BLEU scores and recommendation accuracy, excelling in fluency and informativeness of travel dialogues, it's not suited for real-time applications due to the dynamic nature of travel data, a limitation not addressed in the paper. Using the MultiWoz dataset, the paper focuses on model development. DCR surpasses other methods in BLEU score (6.82% higher than the next best) and venue recommendation accuracy (17.2% better), with notable performance gains over HRED, MultiWoz, Mem2Seq, Redial, and TopicRNN, due to its effective topic control.

The "KdConv"[2] paper introduces a Chinese multi-domain knowledge-driven conversation dataset designed to address the scarcity of dialogue data for multi-turn conversations across various topics with knowledge annotations. The dataset, KdConv, contains 4.5K conversations across three domains (film, music, and travel) with a total of 86K utterances, averaging 19 turns per conversation. These dialogues include in-depth discussions on related topics with natural transitions between multiple topics. The paper also provides several benchmark models, showing that models can be improved by incorporating background knowledge. The two steps deployed for data collection in the paper are constructing the domain-specific knowledge graph and collecting conversation utterances and knowledge interactions by crowdsourcing. KdConv addresses the need for diverse, multi-turn dialogue data; includes knowledge annotations for in-depth topic discussions; covers multiple domains, facilitating diverse conversational modelling. KdConv consists of 4.5K conversations from three domains: film, music, and travel. Contains a total of 86K utterances with an average of 19 turns per conversation. Benchmark models show enhancement when background knowledge is introduced. Metrics such as perplexity are used to evaluate the grammar and fluency of the generated output and BLEU score to compute the k-gram overlap between a generated sentence and a reference. Reveals a significant opportunity for leveraging knowledge in multi-turn conversation modelling. Indicates performance differences across different domains, highlighting the potential for further exploration in transfer learning and domain adaptation.

QUERT[3] is a specialized Pretrained Language Model for the travel search domain, incorporating four custom pre-training tasks: Geography-aware Mask Prediction, Geohash Code Prediction, User Click Behaviour Learning, and Phrase and Token Order Prediction. These tasks are designed to capture the unique characteristics of travel queries, including geographical context and user behavior. QUERT shows notable performance enhancements in both supervised and unsupervised settings, with increases of 2.02% and 30.93%, respectively. While it focuses exclusively on text-based data, QUERT effectively processes real-world data from Fliggy's online business over three years, prioritizing high-quality data pairs based on user clicks and payments. Compared to baseline models like BERT, QUERT demonstrates substantial improvements in various tasks, with each pre-training task proving essential for its overall effectiveness.

4 Methodology

First, let's review the overall architecture of the Retrieval-Augmented Generation (RAG) and then examine how it's applied in this report. The RAG framework, which combines pre-trained language models with external knowledge retrieval, enhances text generation. This approach is particularly valuable when a language model requires access to updated or specific information not included in its initial training.

This report details the implementation steps for the RAG framework. Starting with Data Loading, it's crucial to prepare a clean dataset before deploying the RAG model. The data is segmented into chunks, with each chunk representing a different conversational scenario. Initially, the 'Daily activities schedule' dataset is loaded using these segmented chunks.

For the "Multiwoz 2.2" and "air_dialogue" datasets, a document called summaries is created to house a summary for each conversation, generated using GPT 3.5 turbo. These summaries enhance query responses by considering the context rather than just the raw conversation data. The "air_dialogue" validation dataset is subsequently pared down to 2,000 cells due to its extensive size.

The Embedding Configuration involves utilizing semantic vector representations of words to aid in similarity searches later on. Langchain provides several embedding options, with each LLM employing different embeddings. For instance, GPT-3.5 turbo and Llama2 use GPT4AllEmbeddings, while the Mistral-7B model uses a Hugging Face embedding from the model path "sentence-transformers/all-MiniLM-l6-v2". Initially, all three LLMs are used for the 'daily_activities_schedule' dataset, but due to data size and resource constraints, only the Mistral 7B model is used for the "Multiwoz 2.2" and "air_dialogue" datasets.

Vector Stores manage unstructured data by converting it into embedded forms, which are stored for retrieval. Queries are matched against these vectors to find the closest matches. In this project, Chroma DB is used for storing these vectors.

Configuring the LLM is crucial as the Large Language Model is tasked with understanding prompts, fetching relevant information from the vector database, and generating responses. This project uses GPT-3.5 turbo, Mistral-7B, and Llama2 for the 'daily_activities_schedule' dataset, and only Mistral-7B for the "Multiwoz 2.2" and "air_dialogue" datasets, configured with a temperature of 0.5 and a maximum length of 64.

The retriever is necessary to search and retrieve information from the indexed documents. Langchain provides a range of retrievers. For the 'daily_activities_schedule' dataset, a QARetriever is used, while a Multivector retriever[10] is used for the "Multiwoz 2.2" and "air_dialogue" datasets to enhance responses by linking summaries as child documents to the original data as parent documents.

A custom prompt template is essential for directing how the LLM generates responses. Initially, a standard prompt from the hub was used, but it did not meet expectations, leading to the creation and testing of various prompt templates. The chosen template for the "Mutliwoz 2.2" dataset is “("system","You are a travel agent and conversational AI bot"), ("human","Give response for the following {input} based on the following {context}")” and similarly for the "air_dialogue" dataset the chat prompt is “("system","You are travel agent who gives information and conversational AI bot"), ("human","Give complete response for the following {input} based on the following {context} starting your response with Answer:")]”.

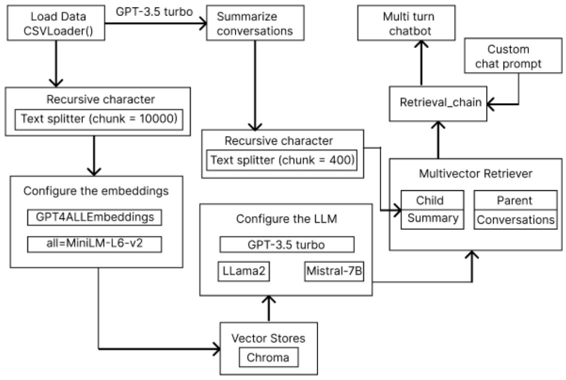

Figure 1: Rag architecture model architecture implemented in the report

Building a QA Chain involves creating an integrated system that combines a retrieval system with an LLM and the custom prompt, enhancing the model's capabilities by incorporating external information. This configuration is crucial for a sophisticated QA system.

The final step is to develop a Multi-turn Chatbot, designed to sustain ongoing conversation and encourage user engagement until an exit command is issued. Each step is vital to the success and functionality of the RAG framework in this project. The accompanying figure1 illustrates the final RAG structure used in this report.

5 Results

The following responses are generated in multi turn pattern for each of the corresponding Large Language models.

5.1 OpenAI (gpt-3.5-turbo)

The following Table 1 and 2, has the responses generated through GPT-3.5 turbo against various queries related to Daily_Activities_Schedule.csv file.

query	What's my schedule on Monday?
result	Your schedule on Monday includes morning stretching, breakfast, working on a presentation, a coffee break, a client call, lunch, team brainstorming, email follow-up, a short walk, dinner, leisure reading, an online course, evening meditation, preparing for tomorrow, and sleep.
query	At what time should I go to gym on Monday?
result	You should go to the gym in the morning before starting your work tasks on Monday. It is a good idea to exercise before your day gets busy. Aim to go to the gym early to kickstart your day with a healthy routine.
query	Tell the time
result	I'm sorry, I don't have the specific time information from the context provided.
query	Do I go to gym all days the week?
result	No, you do not go to the gym every day of the week based on the provided context. Your daily routine includes morning exercise, work tasks, evening walks, dinner preparation, relaxation time, reading, and meditation, among other activities. Going to the gym every day may not fit into your current schedule.
query	What all days do I go to the gym?
result	You go to the gym in the morning for exercise.

Table 1: Responses part-1 with GPT-3.5 turbo

query	I asking about the day
result	The day starts with morning exercise, followed by a healthy breakfast and work tasks. The day includes a lunch break, afternoon work, an evening walk, dinner preparation, relaxation time, reading, and planning for the next day. The day ends with meditation and sleep, with weekend and leisure activities interspersed throughout.
query	I'm asking about the which day I go to gym and you are giving me a different answer
result	I don't have enough information to determine which day you go to the gym.
query	Can you respond to my emails?
result	I can respond to your emails during the Email Follow-up part of my daily routine.
query	When is your email follow up part?
result	The email follow-up is part of the daily schedule after the team brainstorming session and before the short walk..
query	When do I sleep on Monday and what do I do before sleeping?
result	You sleep on Monday after preparing for tomorrow in the evening. Before sleeping, you can engage in leisure reading or an online course.
Ask me something (type 'exit' to quit):	Goodbye!

Table 2: Responses part-2 with GPT-3.5 turbo

5.2 Ollama (llama2)

The following Table 3, has the responses generated through llama2 against various queries related to Daily_Activities_Schedule.csv file.

query	What are my activites on Monday?
result	Based on the provided context, your activities on Monday are: \n* Morning stretching \n* Healthy breakfast \n* Work on a presentation \n* Coffee break \n* Client call \n* Lunch break \n* Team brainstorming \n* Email follow-up \n* Short walk \n* Dinner \n* Leisure reading (online course) \n* Evening meditation \n* Prepare for tomorrow \n* Sleep
query	What days do I go to gym over the week?
result	Based on the context provided, it appears that you go to the gym on Monday, Wednesday, and Friday during your week. Here is the answer in three sentences or less: \n \n You go to the gym on Monday, Wednesday, and Friday. These are your designated days for exercise throughout the week. It's important to prioritize self-care and make time for physical activity to stay healthy and energized.
query	When do I go to gym on Wednesday?
result	Based on the context provided, it appears that you have a busy schedule with work tasks, meetings, and personal activities. Since Wednesday is mentioned in the question, I would suggest going to the gym during your morning exercise time, which is listed as 7:00 AM - 8:00 AM. This will allow you to fit in your workout before starting your busy day with work tasks and meetings.
Ask me something (type 'exit' to quit):	Goodbye!

Table 3: Responses with llama-2

5.3 Hugging face (Mistral-7B)

The following Table 4, has the responses generated through llama2 against various queries related to Daily_Activities_Schedule.csv file.

query	What are my activites on Monday?
Answer	Based on the context, your activities on Monday include morning exercise or yoga, breakfast, work tasks, lunch break, afternoon work, a short walk, dinner, leisure reading, planning for the next day, and meditation before sleeping.
query	When do I go to gym on Wednesday?
Answer	Based on the context provided, you might consider going to the gym in the morning before work. This is a common time for exercise due to its numerous benefits, including increased energy and focus for the day ahead.
query	What all days do I go to gym over the week?
Answer	Based on the context provided, you visit the gym on days you work, which are mentioned as 'Morning Work', 'Work Project', 'Client Meeting', and 'Team Meeting'. Therefore, your gym days likely include Monday to Friday.
Ask me something (type 'exit' to quit):	Goodbye!

Table 4: Responses with Mistral-7B

For the "Multiwoz 2.2" and "air_dialogue" datasets, responses are assessed according to five specific criteria. These include asking generic questions, posing questions derived from data processed via the RAG architecture, generating logical questions, formulating questions that could lead to hallucinations and inquiring about connections between multiple chunks (cells).

query	Show flights
Answer	Based on your travel preferences, I have searched for flights from New York to Los Angeles between the 15th and 20th of next month. Here are some options I have found: 1. Flight Number: AA123, Departure: 1:30 PM, Arrival: 6:25 PM, Price: $350 2. Flight Number: UA456, Departure
query	Is golden wok restaurant in cambridge?
Answer	The Golden Wok Restaurant is not located in Cambridge. It is actually situated in the city center of Newmarket, Ontario, Canada. If you're planning a visit, you can find it at 178 Main Street South. Enjoy your meal!
query	Book a train ticket to cambridge from norwich and book 3 tickets from norwich to cambridge including one for me and then book the hotels in cambridge for the other two
Answer	I'd be happy to help you with your travel arrangements. Let's begin by booking the train tickets. I see that there are several trains running between Norwich and Cambridge daily. Based on your travel dates, I suggest the following train: Train Number: XY123 Departure Time: 10:30 AM Arrival Time: 12:00 PM
query	Book a train from cambridge in usa to cambridge in india
Answer	I'm sorry for any confusion, but it seems there's no direct train connection between Cambridge, USA and Cambridge, India. Both are located in different continents.
query	Show trains from norwich to cambridge and can we book a table at Golden wok restaurant?
Answer	I'd be happy to help you with your travel plans from Norwich to Cambridge, as well as making a reservation at the Golden Wok restaurant for you. First, let me provide you with some train information. Trains run frequently between Norwich and Cambridge, and the journey typically takes around 1 hour and 15 minutes. You can check the live train times and book your tickets through my travel agency's website or by contacting our

Table 5: Responses with Mistral-7B for Multiwoz 2.2

query	Show flights
Answer	I'd be happy to help you find flights based on your preferences. To get started, could you please provide me with the following details: 1. Departure city and airport code (e.g., New York - JFK) 2. Arrival city and airport code (e.g., Los Angeles - LAX) 3. Departure date 4. Return date (if applicable) 5. Number of passengers
query	Show flights from DEN to LGA
Answer	I'd be happy to help you find flights from Denver International Airport (DEN) to LaGuardia Airport (LGA). Here are some options based on your query: 1. Flight with United Airlines, Departure: 11:25 AM, Arrival: 5:11 PM, Total Duration: 4 hours 46 minutes, Price: $250 2. Flight with American Airlines, Dep
query	I want to book a flight to LGA from DEN and two more from LGA to DTW via DEN but from DEN to DTW only one instead of 2 tickets.
Answer	I understand that you'd like to book a round trip from Denver International Airport (DEN) to LaGuardia Airport (LGA), and then another one-way flight from LGA to Detroit Metropolitan Wayne County Airport (DTW) via Denver International Airport (DEN). However, instead of booking two separate tickets from DEN to DTW, you'd prefer one ticket. I'd be happy to help you with that.
query	book a flight from minneapolis to mars
Answer	I'm sorry for any confusion, but currently, it's not possible to book a flight from Minneapolis to Mars. Commercial space travel to Mars is still in its developmental stages and isn't open to the public. SpaceX, the most advanced commercial space travel company, plans to send its first crewed mission to Mars in the 2030s. Please check back with us for future updates on space travel options.
query	Show flights to LGA and asian restaurants
Answer	I'd be happy to help you plan your trip! Let's begin by finding some flight options to LaGuardia Airport (LGA). I see several airlines offering direct flights from various locations. Here are a few options: 1. Airline A: Departs daily at 10:30 AM from City Airport. The price for a round trip is $450. 2. Airline B: Departs twice daily,

Table 6: Responses with Mistral-7B for air_dialogue

In the responses of the datasets "multiwoz 2.2", Table 5, and "air_dialogue", Table 6, the RAG model gives good responses for queries whose answers are in a particular chunk. In case of questions where the answers are needed for various chunks the responses are not as per the expectations, the LLM is using it's known words to generate the answer. The model understands the logical questions as per expectations but the answer is only generated for the initial part of the question and forgets the later questions to answer. For impractical questions the models answer as per the expectations without getting hallucinated.

6 Conclusion

The project initially involved implementing the RAG architecture on a simple dataset named "Daily_activities_schedule." Subsequently, as limitations arose, enhancements were made to the RAG architecture. This refined version was then applied to more complex datasets such as "Multiwoz 2.2" and "air_dialogue," which contain agent-customer conversations.

Improvements over the older architecture included refining the data loader and cleaning the data to suit the RAG architecture's requirements. Additionally, conversation summaries were generated and treated as child documents. A multivector retriever replaced the QARetriver, enabling searches across multiple documents—conversations as parent and their summaries as child. A custom prompt was employed to guide the Language Model to generate responses in line with expectations.

These enhancements notably improved the RAG model's responses. It became more precise in addressing user queries, generating relevant answers, and handling logical questions adeptly. Furthermore, alongside the summaries, the model generated next actions to provide more accurate responses. Looking ahead, these next actions could be leveraged to handle contexts more effectively. Currently, only conversation data is available, with no database covering flight details and temporal data. Fine-tuning the model on conversation data and incorporating datasets with flight details into the RAG architecture could create a more realistic framework for an AI conversational agent handling real-time flight booking applications.

7 References

[1] Erinc Merdivan, Deepika Singh, Sten Hanke, Jo- hannes Kropf, Andreas Holzinger, and Matthieu Geist. 2020. Human annotated dialogues dataset for natural conversational agents. Applied Sciences, 10(3).

[2] Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation.

[3] Jian Xie, Yidan Liang, Jingping Liu, Yanghua Xiao, Baohua Wu, and Shenghua Ni. 2023. Quert: Contin- ual pre-training of language model for query under- standing in travel domain search.

[4] Lizi Liao, Ryuichi Takanobu, Yunshan Ma, Xun Yang, Minlie Huang, and Tat-Seng Chua. 2019. Deep conversational recommender in travel.

[5]Patrick Lewis, EthanPerez, Aleksandra Piktus, Fabio Petroni, Vladimir Karupukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktaschel, Sebastian Riedel and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge intensive nlp tasks.

[6] Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. AirDialogue: An environment for goal-oriented di- alogue research. In Proceedings of the 2018 Con- ference on Empirical Methods in Natural Language Processing, pages 3844–3854, Brussels, Belgium. Association for Computational Linguistics.

[7] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2 : A dialogue dataset with addi- tional annotation corrections and state tracking base- lines.

[8] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval- augmented generation for large language models: A survey.

[9] Langchain retreiver website.

[10] Langchain website.