Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Exploring Q&A Information Retrieval using LLMs

Progress Report

Pages 10

Contents

Motivation 2
Overall Goal and Objectives 3
Literature Review 3
Current status 5
Self-review 6
Project Management 6
References 8

Motivation

The transformative impact of advancements in LLMs highlighted by the advent of ChatGPT in 2022 is the main motivation behind this dissertation. The potential of LLMs in understanding and processing advanced queries and give intelligent responses has led to improved and interactive information retrieval. Inspired by the work done at Perplexity.ai in using LLMs for answering user queries on live and recent events by summarising the events with live information by leveraging LLMs, giving significantly more informed answers than a static search engine such as Google. The efficacy of different models, both open-source and closed-source models based on different criteria such as coherence, comprehensiveness, factual consistency and bias in their answers are going to be compared due to the reproducible nature of the open-source LLMs as compared to the closed-source LLMs. This will be done to formulate the right approach which leverages the capabilities of LLMs, combines the concepts of Information Retrieval and present an innovative method of interacting with and querying already existing as well as new datasets. A comparison between a query made by perplexity and Google on the same topic is shown. Perplexity leverages RAG and LLMs to produce a context and content rich response which encompasses many relevant topics associated with the query.

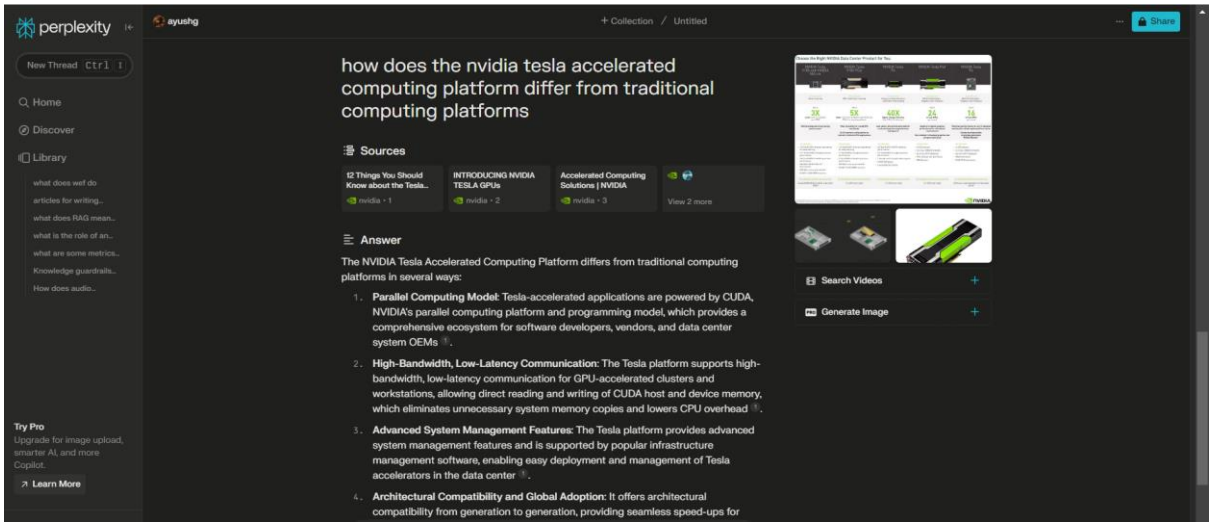

Figure 1: Query made using Perplexity.ai

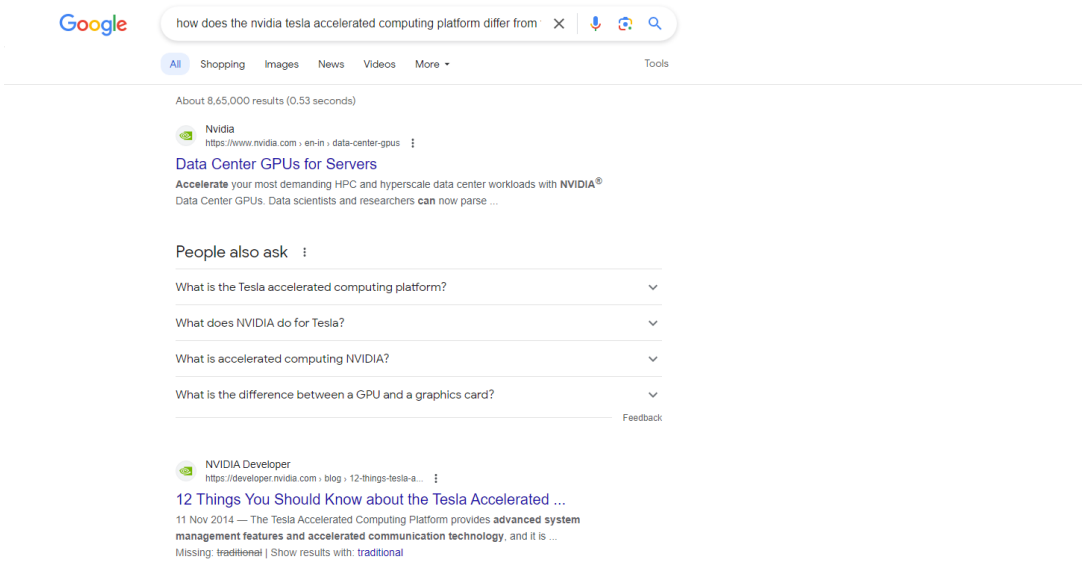

Figure 2: Same query made using Google.com

Overall Goal and Objectives

The main objective of this dissertation is to explore the questioning and answering capabilities of Information Retrieval systems using LLMs. The performance of different LLMs will be compared and contrasted against each other and would contain a good mix of closed source and open source models for comparison. Their performance will be tested on various shared tasks, which serve as a benchmark to assess the performance of different models by providing a uniform set of questions and a standard evaluation metric, enabling direct comparison of the efficacy of various LLMs in retrieval and presentation of relevant answers. The dissertation will have 3 major experiments which are as follows:

1. Working on Shared tasks using the closed source models such as OpenAI APIs and Claude's API.
2. Working on shared tasks using the open source models such as Llama and other present top performing models from Hugging Face.
3. Test out the efficacy of the models tested in the previous experiments on a new task. This new task will look at the ability of the proposed solution to query and answer an information corpus.

The above 3 experiments are laid out and would be completed through the course of this dissertation. There could be additional experiments which could be performed which are better tailored for the final application during the dissertation but would fall under the umbrella of these 3 experiments and can be regarded as sub experiments to better track the progress. The main objective is to assess the quality of answers given by the different open-source and closed source models, compare the limitations and strengths of these models based on metrics such as response comprehensiveness, factual consistency, coherence and bias in the answers. The metrics for evaluating LLM-enhanced Information retrieval system would include traditional IR metrics and LLM-specific metrics such as Recall, Precision and F1 score for traditional IR and Time To First Token(TTFT), Inter-Token Latency(ITL), End-to-End Latency and Completed requests per minute for LLM specific metrics. The limitations and challenges posed by working with open-source models as compared to closed-source models will also be extensively documented, making a comparison on the basis of the difference in computational requirements and cost.

Literature Review

LLM also known as Large Language Models exhibit natural language processing, reasoning and multi-step problem solving capabilities. They show impressive abilities in few-shot learning, instruction following and model calibration demonstrating the ability of the model to scale well. The attention mechanism and encoding strategies enable complex representation of input sequences and understanding of the sequence order. Pairing this with activation function and normalization techniques further refines an LLMs processing capabilities. Despite these astonishing advancements, using LLMs probes some challenges, a major one being that of hallucination: when a model generates non-existent or false content. This necessitates a careful oversight and validation of outputs from an LLM and should be regulated in sensitive domains like scientific research and medicine. [1] [2] [9]

RAG also known as Retrieval Augmented Generation (RAG) merges the functionalities of retrieval and generative models. The architecture allows a RAG to retrieve information from big databases with generation of context-rich text. RAG works very well in text summarization as it can provide concise and relevant summaries. RAGs are computationally very demanding due to the complexity of integrating retrieval and generative components. Data preparation poses another significant challenge, which requires data engineering to ensure relevance and utility for the generative model. These can be solved by regular updates to the data source, and comprehensive output evaluation. [3] [4] [6]

Information Retrieval has evolved significantly with the integration of LLMs and AI marking a new era in the field. What started with Vector Space Model, ranking documents based on the cosine similarity between the query and document vectors, progressed to probabilistic retrieval models such as BM25 and TF-IDF model, which incorporate probabilistic frameworks to estimate the relevance of a document to a given query. Recently, the emergence of neural information retrieval models and its integration with LLMs has improved query understanding, document retrieval, result ranking and relevance feedback mechanisms. This has led to a shift from keyword based querying to intent driven retrieval which signifies a shift in way new information is sought and processed. [3]

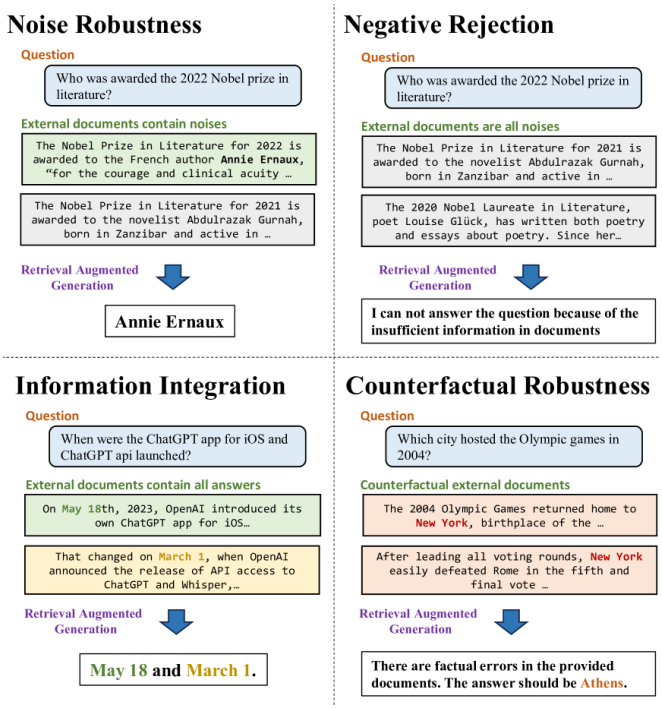

Figure 3: Abilities required for RAG of LLMs [3]

The use of LLMs have emphasized the utility of LLMs in augmenting retrieval processes as seen in Retrieval-Augmented Generation (RAG). The RAG framework evaluates LLMs based on Noise robustness which assesses the ability of the LLM to filter out irrelevant information from noisy inputs along with metrics such as negative rejection and information integration which are used to avoid propagation of misinformation from the document and comprehensiveness of responses respectively.

Additionally, the use of LLMs in Information Retrieval systems have transformed they key components of IR systems such as the query rewriters, retrievers, re-rankers and readers. A Query Rewriter is used to refine user queries to align with information requirements. the ability of the LLM to understand the context with which the query is written by the user, the LLMs can perform tasks like query expansion and pseudo relevance feedback which improve the precision and expressiveness of the queries. This capability is beneficial in personalized and conversational search. LLMs have enable the transition from traditional bag-of-words models like BM-25 to neural IR paradigms, where queries and documents are projected into high-dimensional vector spaces. This enhances the nuanced understanding of the query-document relationships by capturing semantic similarities and hence improving the efficiency and effectiveness of the retrieval process. [1] [5] [7] [8]

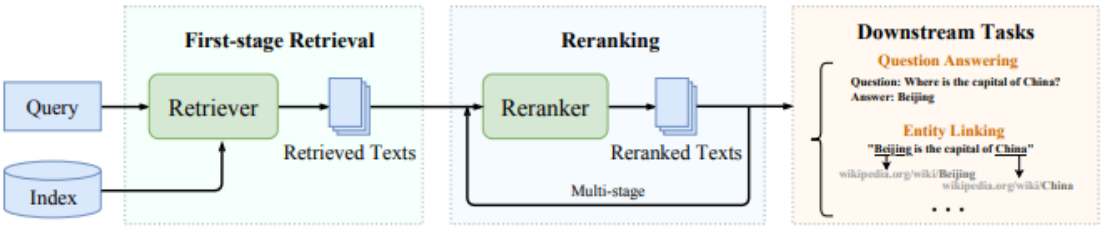

Figure 4: Overall pipeline of an Information Retrieval System [6]

Reranking is the process of fine grained reordering of documents within the retrieved set. Traditionally methods such as vector inner products was used for this task and have been improved upon by integration of LLMs as they provide richer matching signals to the Re-Ranker. The reader module has seen the most significant enhancements. Instead of preparing a list of candidate documents for a specific query as was done previously, LLMs organise answer text in more intuitive manner, resembling natural human information access while also integrating references into generated responses to enhance the credibility of the information provided. [3]

Current status

Initial experiments were conducted to see the viability of different developmental frameworks which are common practice in Information Retrieval and development of large language models. Upon conducting the experiments, LLamaIndex was decided upon to be used for indexing the corpus of documents that will be used, storing the index for later retrieval and querying the indexed file. In the first experiment a closed source model, the OpenAI davinci-002 API is leveraged to query the indexed documents. In this example an essay by **Paul Graham** is fed into the indexer which is indexed and stored. Query engine is invoked and queries are made. For each query, a reference context from the text is specified and stored to compare it against the response given by the queried LLM.

```
from llama_index import VectorStoreIndex, SimpleDirectoryReader

documents = SimpleDirectoryReader("data").load_data()
index = VectorStoreIndex.from_documents(documents)
```

Figure 5: Creating an index of the data

```
query_engine = index.as_query_engine()
response = query_engine.query("What did the author do growing up?")
print(response)
```

Figure 6: Querying data leveraging LLM

```
"query": "The author shares his work experience at a company called Interleaf. Describe the unique feature that Interleaf had added to their software and its
"query_by": {
    "model_name": "gpt-3.5",
    "type": "ai"
},
"reference_contexts": [
    "We actually had one of those little stoves, fed with kindling, that you see in 19th century studio paintings, and a nude model sitting as close to it as
],
"reference_answer": "Interleaf, the company where the author worked, had added a unique feature to their software. Inspired by Emacs, they incorporated a scri
"reference_answer_by": {
    "model_name": "gpt-3.5",
    "type": "ai"
}
```


Figure 7: Reference context and query response

The finally obtained responses are then compared against the reference context texts to obtain the standard metrics of Information retrieval and LLMs such as answer_similarity, retrieval_precision, Augmentation_accuracy and answer_consistency.

Self-review

The project has picked up pace over the winter break and significant progress has been made in formulating the strategy and the plan for a successful writing of the dissertation. Before the winter break, many frameworks and technologies were looked at to nail down on the final tech which would be used in the project. Since the advent of LLMs is very recent and the frameworks around it to support the development of technologies that leverage the capabilities of LLMs are nascent, it has been quite challenging to nail down on the frameworks as part of the developmental pipeline. As the developmental tools keep on evolving for both open-source and closed-source LLMs, it has been decided to use the current stable versions of Langchain and LlamaIndex which are open source libraries used to interact with the LLMs and enable RAGs.

As discussed earlier, the project has been divided into 3 main experiments which allows for better management and progress tracking. Currently the experiment1 is being worked on and as per the project plan, the progress is in line with formulated timeline. The nascent nature of frameworks will pose a challenge in development for the project due to the high number of bugs which could be encountered, as is the case with newly launched frameworks and libraries.

Project Management

As it can be observed from the Gantt chart the project has the following timeline:

	Task Mode	Task Name	Duration	Start	Finish
1		Experiment 1 - Shared task with OpenAI	10 days	Mon 08-01-24	Fri 19-01-24
2		Experiment 2 - Shared task with Closed-Source model	10 days	Wed 24-01-24	Tue 06-02-24
3		Experiment 3 - Application with the best model	15 days	Thu 08-02-24	Wed 28-02-24
4		presentation preperations	5 days	Thu 29-02-24	Wed 06-03-24
5		literature review writing	30 days	Mon 22-01-24	Fri 01-03-24
6		preparing ground truth for experiment 3	3 days	Mon 12-02-24	Wed 14-02-24
7		Miscellaneous bug fixing	35 days	Mon 22-01-24	Fri 08-03-24
8		Final writeup	10 days	Thu 07-03-24	Wed 20-03-24
9		Draft review cycle 1	4 days	Thu 21-03-24	Tue 26-03-24
10		Draft review cycle 2	4 days	Wed 27-03-24	Mon 01-04-24
11		Final touchups in report presentation	4 days	Tue 02-04-24	Fri 05-04-24
12		submission	1 day	Mon 15-04-24	Mon 15-04-24

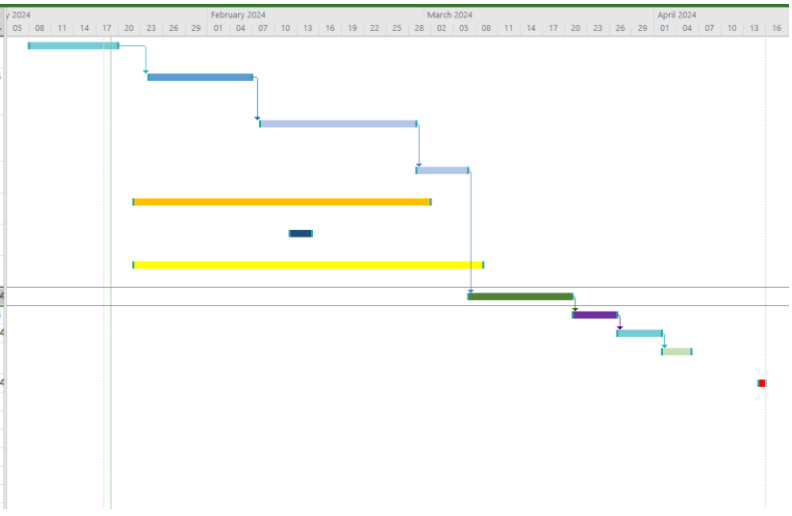

Figure 8: Project task Breakdown with corresponding Gantt Chart

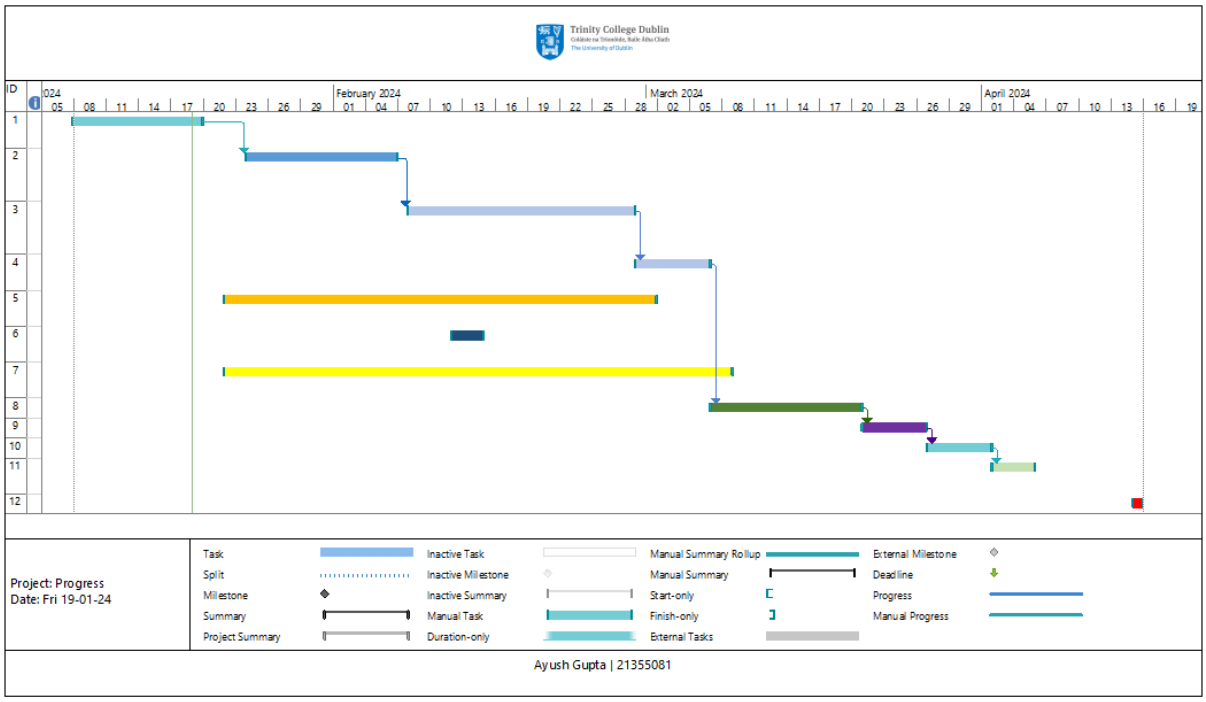

Figure 9: Project Timeline as illustrated by a Gantt Chart

	Task Mode	Task Name	Duration	Start	Finish	Predecessors
1		Experiment 1 - Shared task with OpenAI	10 days	Mon 08-01-24	Fri 19-01-24	
2		Experiment 2 - Shared task with Closed-Source model	10 days	Wed 24-01-24	Tue 06-02-24	1
3		Experiment 3 - Application with the best model	15 days	Thu 08-02-24	Wed 28-02-24	2
4		presentation preperations	5 days	Thu 29-02-24	Wed 06-03-24	3
5		literature review writing	30 days	Mon 22-01-24	Fri 01-03-24	
6		preparing ground truth for experiment 3	3 days	Mon 12-02-24	Wed 14-02-24	
7		Miscellaneous bug fixing	35 days	Mon 22-01-24	Fri 08-03-24	
8		Final writeup	10 days	Thu 07-03-24	Wed 20-03-24	4
9		Draft review cycle 1	4 days	Thu 21-03-24	Tue 26-03-24	8
10		Draft review cycle 2	4 days	Wed 27-03-24	Mon 01-04-24	9
11		Final touchups in report presentation	4 days	Tue 02-04-24	Fri 05-04-24	10
12		submission	1 day	Mon 15-04-24	Mon 15-04-24	

Figure 10: Project Tasks Breakdown

As it can be observed from the tasks breakdown, good amount of time has been allocated to the performing of the 3 main experiments of the project. Certain tasks are scheduled to be done which would be based on the outcomes of the experiments such as the bug fixes and the relevant literature review. Since we're meticulously comparing the viability of using Open-source and close-source LLM models in our application, various metrics will be recorded to make a comparison and hence a final recommendation on the best practices.

The proliferation of various LLM applications has ushered a new dimension in traditional information retrieval systems wherein the way of querying the documents and receiving appropriate response have changed significantly. The challenge posed by working with closed source LLMs vs open source one would be a difficult comparison to make because of the trade off in convenience and meaning they have to offer. While working with closed source models such as the OpenAI API or the Claude API, the model weights are not disclosed which means that the performance of the model is subject to its owner and is unlikely to remain constant and hence would be unreproducible. At the same time the closed source models APIs make the development process easier, due to better support for the API and more features while also taking away the complexity of setting up the model and running the inferences on a local machine which would be computationally very expensive.

On the contrary, open source model while offering the long term reliability in terms of reproducibility, often fall behind at the time of writing in various metrics which measure the performance of LLMs when compared to closed source LLM models. Nonetheless, over time the open source LLMs would match the capabilities of closed source models which also being freely and widely available and hence they would be preferred to use during the course of this work. The performance of the two models would be compared on various metrics as discussed earlier.

Given the nascent nature of this rapidly evolving space with new improved tools coming in everyday, as discussed earlier, its best to use stable releases of prominent tools and models which are open-source in nature and hence would allow great interoperability and reproducibility for our experiments. In case working with open-source models posses challenges beyond the scope of this project such as requirement of exorbitantly high computational power then the information retrieval system would be tested on the closed-source models more closely.

A timeline has been created as it can be observed from the Gantt chart and the task breakdowns in figure 7 and figure 8, it serves as a compass to guide and ground the project and is liable to delays and extensions. Hence, a 10 day contingency period is left at the end of the current version of the plan to accommodate and rectify any bugs and issues which might arise during the final stages of the project.

References

[1] S. G. a. M. Zakershahrak, “Adapting LLMs for Efficient, Personalized Information Retrieval: Methods and Implications,” [Online]. Available: https://arxiv.org/pdf/2311.12287v1.pdf.

[2] W. C. W. P. e. Derong Xu, “Large Language Models for Generative Information Extraction: A Survey,” 29 December 2023. [Online]. Available: https://arxiv.org/pdf/2312.17617.pdf.

[3] H. L. X. H. L. S. Jiawei Chen, “Benchmarking Large Language Models in Retrieval-Augmented Generation,” 20 December 2023. [Online]. Available: https://arxiv.org/pdf/2309.01431.pdf.

[4] K. Pakhale, “Large Language Models and Information Retrieval,” *International Journal for Multidisciplinary Research (IJFMR),* vol. 5, no. 6, 2023.

[5] T. B. Z. C. e. Qingyao AI, “Information Retrieval Meets Large Language Models: A strategic report from Chinese IR Community,” [Online]. Available: https://arxiv.org/pdf/2307.09751.pdf.

[6] J. L. R. R. a. J.-R. W. Wayne Xin Zhao, “Dense Text Retrieval based on Pretrained Language Models: A Survey,” 27 November 2022. [Online]. Available: https://arxiv.org/pdf/2211.14876.pdf.

[7] H. Y. S. W. e. Yutao Zhu, “Large Language Models for Information Retrieval: A Survey,” [Online]. Available: https://arxiv.org/abs/2308.07107v2.

[8] K. Hambarde and H. Proença, “Information Retrieval: Recent Advances and Beyond,” 2021. [Online]. Available: https://arxiv.org/pdf/2301.08801.pdf.

[9] K. Z. J. L. T. T. X. W. e. Wayne Xin Zhao, “A Survey of Large Language Models,” 24 November 2023. [Online]. Available: https://arxiv.org/pdf/2303.18223.pdf.