THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Project Proposal

Improving Temporal Awareness in LLM-based IR Systems

Anmol Arora

47945209

anmol.arora@uqconnect.edu.au

School of Electrical Engineering and Computer Science

University of Queensland

Supervisor

Professor Guido Zuccon

g.zuccon@uq.edu.au

School of Electrical Engineering and Computer Science

University of Queensland

Submitted for the degree of Master of Data Science

17 April 2024

Abstract

This research proposal aims to address the critical gap in temporal awareness within Large Language Models (LLMs), particularly in the context of the University of Queensland's development chatbot, ChatUQ. Despite the sophisticated capabilities of LLMs in information retrieval (IR) and response generation, their performance in processing time-sensitive queries remains markedly deficient, which can undermine their utility in dynamic and real-time environments. The overarching goal of this project is to enhance the temporal processing capabilities of ChatUQ by systematically identifying, developing, and integrating novel methodologies within its retrieval-augmented generation (RAG) framework.

To achieve this, the research will first quantify temporal awareness deficiencies through the creation of a robust dataset comprising diverse temporal queries. This dataset will facilitate the measurement and enhancement of the temporal accuracy of LLM-based IR systems. Furthermore, the project will review and adapt existing temporal awareness methods from current literature, and develop innovative approaches tailored to address the identified shortcomings. The efficacy of these methodologies will be evaluated through rigorous testing against the enhanced dataset.

Key project objectives include the development of a refined understanding of the temporal limitations present in current LLMs, and the creation of a framework that supports continuous improvement and adaptation of temporal awareness capabilities in IR systems. Through a comprehensive methodology encompassing incident tracking, experimental validation, and performance analysis, this project aims to significantly advance the state-of-the-art in LLM temporal comprehension.

Ultimately, this research not only seeks to improve the functional accuracy of ChatUQ but also aims to set a benchmark for future developments in LLM technologies, particularly for applications requiring acute temporal precision. This proposal outlines a detailed plan for addressing these challenges through innovation in methodology, dataset creation, and systematic evaluation.

Table of Contents

1 Introduction

1.1 Project Context

In this project, I will be working on a specific problem, that is, lack of temporal awareness in LLM-Based IR systems that incorporate RAG architectures such as ChatUQ, a chatbot under development for the University of Queensland.

RAG is Retrieval Augmented Generation. RAG systems use a combination of retrieval and generation to produce responses that are both relevant and contextually rich. These systems are particularly useful in scenarios where answers require external knowledge beyond the training data of the model. They are essential in enhancing the capabilities of large language models (LLMs) in real-world applications (Lewis et al., 2020).

Large Language Models leverage extensive training on diverse datasets to achieve an understanding of language nuances and generate coherent responses. These models, such as OpenAI's GPT series, are transformative in fields ranging from conversational AI to complex problem-solving, demonstrating profound capabilities in Information Retrieval (Naveed et al., 2023).

Information Retrieval(IR) systems are designed to extract relevant and precise information from large datasets quickly and efficiently, catering to specific user queries. These systems underpin the functionality of search engines, digital libraries, and data management services, providing critical support in navigating and accessing information (Zhu et al., 2023).

1.2 Problem Domain & Scope

Although this project aligns with the architecture and needs of ChatUQ, currently, direct integration into ChatUQ, or the use of its specific data, is out of scope. Preliminary analysis has identified a lack of temporal awareness as a critical gap within the ChatUQ system. The research will thus explore and develop methodologies in a broader context, mirroring the architecture used by ChatUQ—its LLMs and RAG pipelines—but will not utilize actual ChatUQ data due to privacy and ethical considerations. This approach ensures that ethical concerns regarding data access are managed, and software engineering complexities related to direct system integration are avoided. Although the research is not conducted on ChatUQ data directly, the findings and methodologies are expected to be fully transferable to the ChatUQ environment.

1.3 Motivation Behind the Project

The motivation for this research stems from the observed deficiencies in temporal awareness in LLMs used within the ChatUQ framework. Current models struggle to accurately address time-sensitive queries, often failing to recognize or properly contextualize temporal expressions. For ex. Table 1 contains 5 queries and with their answers that the current model produces.

Query	Generated Answer
When is my next DATA7901 assignment due?	Your next assignment due date is not available in the provided context.
What happened in the last student committee meeting at UQ?	Based on the new context provided, the most recent discussion topic in the last student committee meeting at UQ was not provided.
When is my next class?	The next class is not available in the provided context.
When will the election for the next Academic Board be held at UQ?	The next Academic Board election at UQ will not be held in April or July 2019, as previously stated. Instead, the election will be held in October 2019.
When will the Exams for Semester 1, 2024 be held at the University of Queensland?	The Exams for Semester 1, 2024 at the University of Queensland have not been announced yet.

Table 1: Examples of ChatUQ Failing to Capture Temporal Context in Queries

As I can see from Table 1, the model does not give time-aware answers. This deficiency not only diminishes user experience but also impacts the reliability of the chatbot in delivering timely and accurate information. Enhancing temporal awareness in LLMs will significantly improve their utility and accuracy. This improvement is crucial for maintaining the relevance and effectiveness of AI systems in educational environments, where up-to-date information is often critical.

1.4 Overview of the Project

In my research, I aim to identify and measure gaps in how Large Language Models (LLMs) handle time-related information. I will analyze how often and under what conditions these models misunderstand temporal data. To do this, I will create and refine a dataset of diverse temporal queries to test and improve the accuracy of LLM responses in information retrieval (IR) systems. Additionally, I will explore existing solutions for similar issues, adapting these methods to fit the specific needs of my research. I also plan to develop new strategies targeted at overcoming the current systems' limitations. A key part of my research involves a detailed evaluation of these solutions, not just to test their initial effectiveness but also to develop an ongoing improvement process for integration into broader applications. My goal is to forge a path for continuous progress, ensuring that our approaches remain adaptable and relevant in the ever-evolving field of LLMs and IR.

2 Background

2.1 RAG Architectures and LLMs

Research on retrieval-augmented generation (RAG) architectures in language models (LLMs) has predominantly focused on enhancing LLMs' ability to generate contextually relevant and information-rich content by leveraging external knowledge bases. The initial exploration presented an innovative approach by combining the transformer architecture with a dynamic external memory. The authors highlighted the model's proficiency in contextual understanding and its applications across various domains such as chatbots and information retrieval systems. However, a notable limitation identified was the latency introduced due to the retrieval process, which impacts the real-time response capabilities of the models (Lewis et al., 2020).

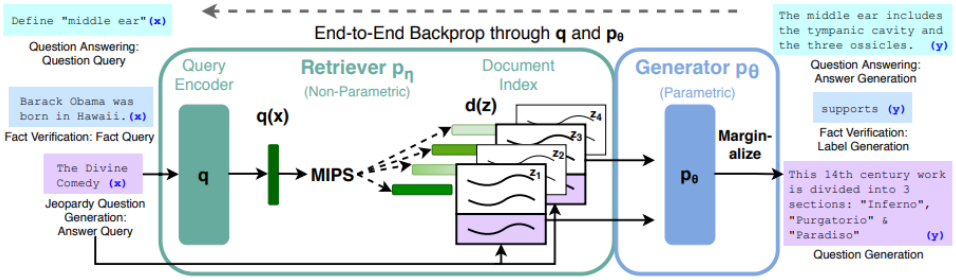

Figure 1: Overview of a Retrieval-Augmented Generation Model (Lewis et al., 2020)

Figure 1 gives an overview of the working mechanism of the RAG model. RAG models find and use relevant text (z) to help generate answers (y) based on a query (x). They consist of a retriever, which finds related content, and a generator that creates responses based on this content and the query context.

Building on this, researchers expanded the RAG architecture's application to more complex tasks including summarization and question answering. Johnson and colleagues refined the retrieval process to improve efficiency and introduced a multi-vector retrieval system that reduces the time overhead. Despite these improvements, the paper noted a persistent challenge with the scalability of the model when interfaced with significantly large data sets, which often led to decreased precision in retrieval outcomes (Gao et al., 2023).

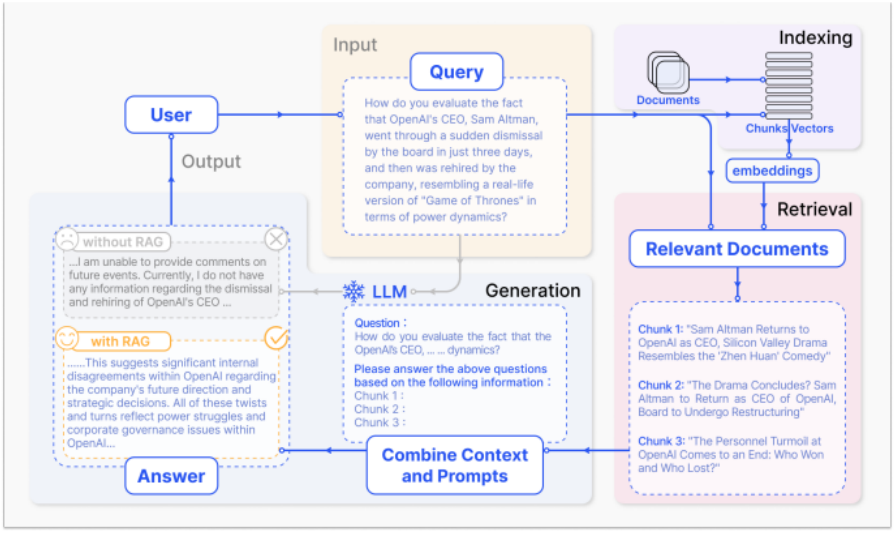

Figure 2: A representative instance of the RAG process applied to QA (Gao et al., 2023)

Figure 2 illustrates a RAG model for answering questions: It first creates a searchable index of document parts, then finds the most relevant parts to a question, and finally uses those parts to generate an answer.

2.2 Information Retrieval using LLMs

Further study delves into integrating information retrieval (IR) techniques directly with LLMs, without the intermediary of traditional RAG components. This integration aims to streamline the retrieval process by embedding IR functionalities within the LLM framework itself. The authors demonstrated that this direct integration allows for faster retrieval times and enhances the relevance of the information retrieved, contributing to more accurate and context-aware responses from LLMs. However, the study also highlighted a limitation in terms of adaptability, where the LLM struggled to dynamically adjust to new or evolving data types without manual interventions (Ai et al., 2023).

2.3 Temporal Awareness in LLMs

Advancements in temporal awareness within LLMs have been marked by a series of studies focused on embedding a sense of 'time' or 'history' into the models. Further research introduced a framework that allows LLMs to maintain stateful interactions over time, which significantly improves their utility in continuous interaction environments like digital assistants and monitoring systems. However, the authors noted that maintaining accuracy with increasing temporal distance remains challenging (Z. Chen et al., 2023).

Other researchers built on this by introducing methods to integrate these temporal frameworks with existing LLM architectures more efficiently. Innovations such as the use of a temporal attention mechanism were highlighted, which enables the LLM to focus on relevant parts of a stored conversation or document based on the time cues. Yet, these methods still face challenges in terms of computational efficiency and the complexity of training with temporal data. (Mann et al., 2020; Zhang et al., 2024).

Further exploration introduced more refined temporal encoding techniques, aiming to reduce the resource demands and improve the model's ability to generalize across different temporal contexts. Each step forward in the research demonstrated a progressive enhancement in handling temporal data, but also underscored the need for more robust training datasets that are temporally varied to train these models effectively (Dhingra et al., 2022; Kynoch & Latapie, 2023).

2.4 Temporal Awareness in IR

The exploration of temporal awareness in information retrieval (IR) systems has significantly advanced with research focused on integrating temporal dimensions into IR frameworks. Recent research introduced a novel approach for enhancing traditional IR systems with temporal tagging capabilities that allow for the sorting and retrieval of information based on time-specific criteria. This model showed promising results in improving the relevance of retrieved documents for time-sensitive queries, though it struggled with the accuracy of temporal tagging in noisy data environments (Moulahi et al., 2016).

"Building on this, another study extended the concept by incorporating a machine learning model that predicts the temporal relevance of documents based on their content and metadata. This system was able to dynamically adjust its retrieval strategies based on the temporal intent of the queries. However, the model required extensive training data to effectively understand and categorize temporal expressions, which limited its immediate applicability in smaller or less structured datasets (Whiting, 2016).

Other researchers implemented an advanced algorithm that combines the features of the previous studies with a real-time updating mechanism to handle streaming data. Their system demonstrated the ability to update its index with new information while maintaining an accurate temporal context. Despite these advancements, the complexity of the algorithm led to challenges in deployment, particularly in resource-constrained environments (Campos et al., 2014).

2.5 Temporal Awareness in QA Systems

The integration of temporal awareness into QA systems has been markedly advanced through the development of the Time-Sensitive QA dataset (W. Chen et al., 2021). This dataset is designed to challenge and benchmark QA systems on their ability to handle time-sensitive questions that require understanding and reasoning over temporal information. The dataset effectively highlights the shortcomings of current state-of-the-art QA systems, which typically achieve only about 46% accuracy compared to human performance at 87%, underscoring a significant gap in the ability to handle temporal reasoning. The authors suggest that the development of new models trained on this dataset could lead to significant improvements in automated systems' understanding of time-sensitive information.

3 Project Plan

3.1 Aim of the Project

The overarching goal of this research is to address the pervasive issue of temporal awareness in Large Language Model (LLM)-based information retrieval systems, specifically within the context of ChatUQ, a chatbot under development for the University of Queensland. The lack of temporal awareness in such systems manifests as a deficiency in processing and responding accurately to time-sensitive queries, crucial for effective user interaction. This project aims to systematically quantify the problem of temporal awareness, identify potential solutions from existing literature, and develop novel methodologies that can be integrated into the retrieval-augmented generation (RAG) framework of ChatUQ. Key objectives include:

- Quantifying Temporal Awareness Deficiencies: Determine the frequency and conditions under which temporal misunderstandings occur within LLMs.
- Development and Refinement of a Dataset: Create and utilize a dataset consisting of diverse temporal queries to measure and enhance the temporal accuracy of response generation in LLM-based IR systems.
- Exploration and Adaptation of Existing Solutions: Investigate current methodologies addressing similar issues in other contexts and adapt these solutions.
- Tailoring of New Methodologies: Develop and test new approaches that explicitly address shortcomings of existing solutions.
- Evaluation and Future Planning: Assess the effectiveness of implemented solutions and establish a road map for ongoing enhancement and integration.

By achieving these aims, the project will significantly advance the temporal processing capabilities of ChatUQ, setting a benchmark for LLM-based systems in handling time-critical information requests.

3.2 Methodology

3.2.1 Measuring Temporal Comprehension

- Incidence Tracking: Establish a systematic approach to track and analyze incidents of temporal misunderstandings within LLM interactions.

- Experimental Validation: Conduct controlled experiments to validate the findings from the incidence tracking.

3.2.2 Dataset Creation

- Query Generation: Utilize both manual creation and LLM-generated prompts to compile a comprehensive dataset of temporal queries relevant to the university context, such as event dates and historical data inquiries.

- Evaluation Criteria Establishment: Implement human annotation to determine the presence of temporal comprehension issues within responses. Deciding what evaluation tools could be used such as precision, Z-Tests for Proportions or binomial tests to validate the significance of findings based on the dataset's diversity and size.

3.2.3 Literature Review

- Source Identification: Conduct a thorough search on databases like Google Scholar using keywords related to temporal awareness in LLMs and RAG systems, focusing on fields like Natural Language Processing and Information Retrieval.

- Paper Categorization: Manually categorize identified papers based on their methodology's relevance and applicability to the project's goals. Methods will be classified as directly applicable, adaptable, inspirational, or not useful, depending on their potential to address the specific needs of ChatUQ.

- Analyze the Papers: Identify baselines that could be implemented and existing methods that could be adapted. Further, report shortcomings of current methods.

3.2.4 Refining Methodologies

- Improvement and Adaptation: Based on insights gained from the literature review, improve on existing methods by creating new methods or adapt existing ones to improve temporal accuracy in ChatUQ. This phase focuses on integrating effective strategies into the RAG framework, considering computational and data constraints.

- Prototyping: Implement these methods within a controlled environment to test their initial effectiveness and integration with existing systems.

3.2.5 Evaluation

- Performance Analysis: Compare and contrast the enhanced system using the specially created dataset with the existing system to evaluate improvements in temporal awareness.
- Gap Identification and Planning: Identify any remaining gaps in temporal comprehension and outline future research directions and enhancements. This includes planning for scalability and deeper integration of successful methodologies.

This comprehensive methodology aims not only to enhance ChatUQ's functionality but also to contribute to the broader academic and practical applications of LLM technologies, particularly in their ability to interact more naturally and effectively with human users in time-sensitive contexts.

3.3 Milestones and Timeline

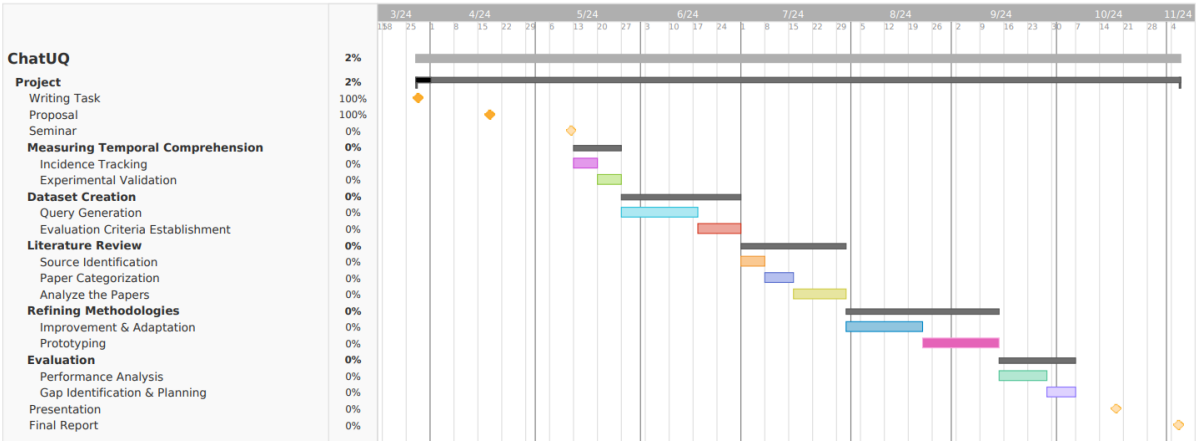

Figure 3 is a Gantt Chart representing how each milestone and its subtasks will be achieved. Please note that some activities may be carried out in parallel and dependencies exist, that is, most subtasks/milestones are dependent on the completion of other subtasks/milestones.

3.4 Risk Assessment

- The experiments that will be needed to be run in this project could be computationally expensive and I might not have enough computational power.

  Mitigation: Initial experiments will be scaled down to run only necessary tests and conserve resources. If further computational resources would be needed they would be requested to the EECS school or a budget for cloud computing would be agreed upon with the research group.

- This project would require dealing with a lot of files including research papers, codes, notes etc., so there could be a risk of loss of files.

  Mitigation: Every file related to the project would be pushed to a GitHub repository. Regular checks would be done to assure copies of all files have been uploaded on the GitHub repository.

- As this project will be moving forward, there is a risk that the research could become outdated when the project is completed.

  Mitigation: Regular literature reviews will be conducted to ensure that new technology/research methods are incorporated in my own project. This will help maintain the project's relevance and effectiveness.

3.5 Ethics Assessment

For this project, I will be exclusively utilizing publicly available data sourced from open datasets. No private or personally identifiable data from users will be involved in any phase of the research. Hence, there is no need to go over the traditional ethical assessment.

References

Ai, Q., Bai, T., Cao, Z., Chang, Y., Chen, J., Chen, Z., Cheng, Z., Dong, S., Dou, Z., Feng, F., et al. (2023). Information retrieval meets large language models: A strategic report from chinese ir community. *AI Open*, *4*, 80–90. https://doi.org/10.1016/j.aiopen.2023.08.001

Campos, R., Dias, G., Jorge, A. M., & Jatowt, A. (2014). Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, *47*(2), 1–41. https://doi.org/10.1145/2619088

Chen, W., Wang, X., & Wang, W. Y. (2021). A dataset for answering time-sensitive questions. *arXiv preprint arXiv:2108.06314*. https://doi.org/10.48550/arXiv.2108.06314

Chen, Z., Li, D., Zhao, X., Hu, B., & Zhang, M. (2023). Temporal knowledge question answering via abstract reasoning induction. *arXiv preprint arXiv:2311.09149*. https://doi.org/10.48550/arXiv.2311.09149

Dhingra, B., Cole, J. R., Eisenschlos, J. M., Gillick, D., Eisenstein, J., & Cohen, W. W. (2022). Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, *10*, 257–273. https://doi.org/10.1162/tacl_a_00459

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*. https://doi.org/10.48550/arXiv.2312.10997

Kynoch, B., & Latapie, H. (2023). Recallm: An architecture for temporal context understanding and question answering. *arXiv preprint arXiv:2307.02738*. https://doi.org/10.48550/arXiv.2307.02738

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Moulahi, B., Tamine, L., & Yahia, S. B. (2016). When time meets information retrieval: Past proposals, current plans and future trends. *Journal of Information Science*, *42*(6), 725–747. https://doi.org/10.1177/0165551515607277

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*. https://doi.org/10.48550/arXiv.2307.06435

Whiting, S. (2016). Temporal dynamics in information retrieval. *ACM SIGIR Forum*, *50*(1), 97–98. https://doi.org/10.1145/2964797.2964818

Zhang, X., Zang, L., Liu, Q., Wei, S., & Hu, S. (2024). Event temporal relation extraction based on retrieval-augmented on llms. *arXiv preprint arXiv:2403.15273*. https://doi.org/10.48550/arXiv.2403.15273

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Dou, Z., & Wen, J.-R. (2023). Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*. https://doi.org/10.48550/arXiv.2308.07107