RAG with Knowledge Structures

Tolga Tel[1,*,†]

[1]Goethe University, Robert-Mayer-Straße 10, Frankfurt, 60325, Germany

Abstract

This project explores methods to enhance the accuracy and reliability of Large Language Models (LLMs), which are prone to factual inaccuracies and hallucinations. The focus lies on improving accuracy by integrating Case-Based Reasoning and to address operational efficiency with pre- and post-processing techniques. Additionally new techniques for retrieving, adapting and retaining information will be researched.

Keywords

Process-oriented CBR, Knowledge Structures, Guided Generation, Retrieval Augmented Generation

1. Introduction

Language serves as a cornerstone for human communication and self-expression, and it also plays a crucial role in how humans interact with machines. The increasing need for machines to effectively handle complex language tasks, such as translation, summarization, information retrieval, and conversational interactions, has driven the demand for generalized language models. Recent advances in language modeling, primarily driven by the development of transformer architectures, increased computational power, and the availability of massive training datasets, have led to significant breakthroughs. These advancements have revolutionized the field by enabling the creation of Large Language Models (LLMs) that can achieve near-human performance on a wide range of tasks. LLMs have emerged as state-of-the-art AI systems capable of processing and generating human-like text, demonstrating impressive abilities in coherent communication and generalization across various tasks [1].

This PhD research focuses on enhancing the performance and efficiency of LLMs through improved information retrieval and processing techniques. While LLMs have shown remarkable capabilities in various language tasks, they often struggle with knowledge-intensive tasks and can generate incorrect information (hallucinations) [2, 3]. Retrieval-Augmented Generation (RAG) addresses this by integrating external knowledge bases. RAG retrieves relevant document chunks to help LLMs generate more accurate responses, reducing factual errors [4].

A basic RAG approach involves: 1) Indexing: extracting, segmenting, embedding, and storing data chunks in a vector database ("Retrieve-Read" [5]). 2) Retrieval: embedding the user query and finding similar chunks in the database. 3) Generation: combining the query and retrieved context into a prompt for the LLM.

This naive approach faces challenges like imprecise retrieval and the potential for hallucination, irrelevant, toxic, or biased outputs. Effectively integrating retrieved information is also difficult [4].

Advanced RAG systems focus on enhancing retrieval quality beyond the naive approach. They employ a multifaceted strategy that encompasses pre-retrieval and post-retrieval techniques [4]. Pre-retrieval focuses on optimizing the indexing structure and the original query. This involves enhancing the quality of the indexed content by refining data granularity, optimizing index structures, adding metadata, and aligning data for optimal retrieval. Simultaneously, the original user query is refined to make it clearer and more suitable for the retrieval task. In [6] a framework for query rewriting was implemented and in [7] abstraction is used as a pre-retrieval method.

This work is grounded in the principles of Case-Based Reasoning, a powerful problem-solving paradigm

ICCBR DC'25: Doctoral Consortium at ICCBR-2025, July, 2025, Biarritz, France

*Corresponding author.

tel@em.uni-frankfurt.de (T. Tel)

0009-0005-5012-8250 (T. Tel)

that leverages past experiences to address new challenges. Case-based reasoning has been formalized as a four-step process [8]:
Retrieval involves finding the most similar case in the case base to the new problem. This is like searching for information, where the new problem guides the search within the case base. Reuse involves proposing a solution for the new problem based on the solution of the retrieved case. If the cases are identical, reuse is straightforward. However, if they differ, adaptation is necessary. Revise evaluates the proposed solution. This can involve real-world testing or simulation. Revision aims to confirm the solution's applicability and may involve addressing unforeseen issues. Retain updates the case base by adding the new (learned) case, enabling future problem-solving. This step involves deciding whether to retain all solutions or only actual cases.
Knowledge Graphs (KGs) represent knowledge as nodes and edges [9], offering concise expert information for RAG [10]. While KG embeddings face sparsity [9], textual graphs add context. GraphRAG uses both. KG Question-Answering (KGQA) excels at factual questions but struggles with open-ended ones [9]. RAG can provide context to KGQA but lacks broader semantic understanding. Graph RAG aims to address this by searching KGs, though subgraph identification is hard [11]. "G-Retriever" [12] uses LLMs and GNNs for natural language interaction with graphs. Unifying LLMs and KGs [13] aims for more robust AI. RAG in LLMs has limitations [14], requiring careful implementation.
The consortium OMG has established the standard BPMN [15] for modeling languages. BPMN offers a standardized graphical notation for visualizing business processes. Its primary goal is to bridge the gap between business and technical users, providing a notation that is both understandable to business analysts and precise enough for technical developers.
This flowchart-like notation empowers stakeholders to design, manage, and realize business processes efficiently. Its independence from specific implementation environments ensures flexibility and adaptability.
Previous work with BPMN-models is found in [16], where the use of Case-Based Reasoning (CBR) within a Retrieval Augmented Generation (RAG) system to generate accessible explanations of business process models is investigated. This research introduces a novel application of CBR for process-oriented tasks.
In [17] the potential impact of Generative AI on Business Process Management (BPM) is discussed. Generative AI can significantly impact BPM by automating routine tasks, improving customer and employee satisfaction, and uncovering new opportunities for process improvement and redesign.
Another work regarding BPMN can be found in [18], where natural language process description are used to generate code. The "Tasks-Model-Extractor" is given a textual-process description as an input and uses an LLM to extract the tasks and the control flow of the process and generates the model representation.

2. Research Plan

The first approaches in this thesis can be assigned to the process-oriented CBR research. Here the idea is to work with Business Process Models and improve RAG systems with efficient pre-processing.

2.1. Research Objectives

This research primarily focuses on improving the performance of Retrieval Augmented Generation (RAG) systems, particularly in the context of Business Process Models. A core objective is to improve the retrieval component of RAG by minimizing redundant information, accurately predicting necessary information, and optimizing data structures for efficient retrieval. Additionally, this work aims to advance the automated generation and evaluation of natural language descriptions from BPMN models, including developing robust quality assessment mechanisms and error correction techniques.
Furthermore, this research explores several related objectives to improve the overall efficiency and reliability of language models and knowledge integration. This includes investigating methods for automatically extracting and structuring knowledge from text into knowledge graphs, and evaluating

the effectiveness of integrating these knowledge graphs into the RAG retrieval process to mitigate hallucinations. Finally, it examines the feasibility of fully automated evaluation metrics for LLM outputs and how CBR principles can be leveraged to enhance RAG performance by learning from past retrieval and generation experiences.

2.2. Approach / Methodology

Latest Work: A paper on using CBR generated explanations for business process models by using clever pre-processing to improve the accuracy while increasing the scalability of the input size, which has been accepted to ICCBR 2025.
LLMs Used: Different iterations of Llama 3 and Mistral were used in experiments leading to the paper mentioned above. The focus lied on smaller models, which can be run locally on a conventional computer to ensure accessibility and easy reproducibility of any result generated.
Prompt Engineering: Prompt phrasing was refined through pre-testing, with the best-performing versions used in the final experiment. These prompts are likely to be reused or adapted in future studies, with domain-specific modifications.
Construction of a Case-Base: For the experiments in the paper, the case-base was written by hand using parts of business process models and writing a corresponding descriptive text.
Retrieval of cases: In pre-testing, different retrieval strategies were used. While embedding-based similarities did not satisfy the needs for our experiments, because cases were too similar and only a small selection of keywords decided the usefulness of a case, another metric was used in the paper. A taxonomy-based metric considering keywords to ensure retrieving a relevant case for the RAG system.

3. Progress Summary

ICCBR 2025: In my first and only submission, I presented a novel approach to addressing the challenges associated with the description of large business process models. Recognizing the inherent complexity and scale of these models, which often hinder efficient analysis, I developed a methodology that helps RAG techniques. A core component of this methodology is the implementation of efficient pre-processing strategies. These strategies are crucial for transforming the raw business process model data into a format that is both useful for effective retrieval and the generation of coherent and contextually relevant descriptions.

4. Conclusion and Future Work

Building upon the foundational work presented in my submitted paper, my research trajectory focuses on a significant refinement and enhancement of the proposed methodology. Specifically, a key objective is to elevate the precision and efficacy of the pre-processing stage. To achieve this, I intend to integrate more sophisticated algorithmic approaches, moving beyond the initial strategies. This includes working with a wider range of business process models, that are not well-modeled.
Furthermore, a significant advancement will be the implementation of an automated evaluation framework for the generated textual descriptions. This framework will move beyond subjective human assessments and introduce quantifiable metrics to rigorously assess the quality of the generated outputs.
I plan to utilize Knowledge Graphs (KGs) as an information repository to develop novel retrieval methods. This approach aims to significantly enhance the relevance of retrieved information. Furthermore, I will explore the integration of the retain phase, dynamically expanding the KG with new knowledge, to continuously improve retrieval accuracy.
Possible research questions regarding my future work are:

1. How can we optimize the workload distribution between pre-processing algorithms and LLMs for maximum efficiency and accuracy?

2. How can the principles of Case-based Reasoning be effectively leveraged to improve RAG performance?
3. Can fully automated evaluation metrics reliably assess the quality of LLM outputs without the need for human intervention?
4. What are the most effective methods for automatically extracting knowledge from textual documents and representing it in a structured knowledge graph format?
5. Can the integration of knowledge graphs or textual graphs into the retrieval process of LLMs effectively mitigate the risk of hallucinations?

Question 1 and 2 were already partly explored in [19] but will reoccur multiple times in different settings throughout this PhD research. Question 3 investigates the feasibility and reliability of fully automated evaluation metrics for assessing the quality of outputs generated by LLMs. While human evaluation remains the gold standard for assessing the quality of natural language generation, it is time-consuming, expensive, and prone to subjectivity. This research aims to determine whether automated metrics can provide reliable and objective assessments of LLM performance, reducing the reliance on human evaluation.
Question 4 focuses on the challenges of information extraction, including named entity recognition, relation extraction, and event extraction, which are crucial for identifying and classifying key components and relationships within the text. Effective knowledge representation plays an important role in ensuring the accuracy, completeness, and consistency of the constructed knowledge graph.
Question 5 investigates the potential of integrating knowledge graphs or textual graphs into the retrieval process of Large Language Models (LLMs) to effectively mitigate the risk of hallucinations. Hallucinations, the generation of factually incorrect or nonsensical information by LLMs, pose a significant challenge to their reliability and trustworthiness.
By incorporating structured knowledge from knowledge graphs or the contextual relationships captured in textual graphs, the retrieval process can be enhanced. This can involve using graph-based algorithms to retrieve relevant information or improve context understanding.

Declaration on Generative AI

During the preparation of this work, the author(s) used Google Gemini, Grammarly for the purpose of: Grammar, spelling check, minor Paraphrase and translations. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Mian, A comprehensive overview of large language models, arXiv preprint arXiv:2307.06435 (2023).
[2] N. Kandpal, H. Deng, A. Roberts, E. Wallace, C. Raffel, Large language models struggle to learn long-tail knowledge, in: International Conference on Machine Learning, PMLR, 2023, pp. 15696–15707.
[3] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al., Siren's song in the ai ocean: a survey on hallucination in large language models, arXiv preprint arXiv:2309.01219 (2023).
[4] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 (2023).
[5] X. Ma, Y. Gong, P. He, H. Zhao, N. Duan, Query rewriting in retrieval-augmented large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 5303–5315. URL: https://aclanthology.org/2023.emnlp-main.322/. doi:10.18653/v1/2023.emnlp-main.322.

[6] W. Peng, G. Li, Y. Jiang, Z. Wang, D. Ou, X. Zeng, D. Xu, T. Xu, E. Chen, Large language model based long-tail query rewriting in taobao search, in: Companion Proceedings of the ACM on Web Conference 2024, 2024, pp. 20–28.
[7] H. S. Zheng, S. Mishra, X. Chen, H.-T. Cheng, E. H. Chi, Q. V. Le, D. Zhou, Take a step back: Evoking reasoning via abstraction in large language models, arXiv preprint arXiv:2310.06117 (2023).
[8] M. M. Richter, R. O. Weber, Case-based reasoning, Springer, 2016.
[9] T. T. Procko, O. Ochoa, Graph retrieval-augmented generation for large language models: A survey, in: 2024 Conference on AI, Science, Engineering, and Technology (AIxSET), IEEE, 2024, pp. 166–169.
[10] H. Zhang, M. O. Shafiq, Triple-aware reasoning: A retrieval-augmented generation approach for enhancing question-answering tasks with knowledge graphs and large language models, in: The 37th Canadian Conference on Artificial Intelligence, 2024.
[11] Y. Hu, Z. Lei, Z. Zhang, B. Pan, C. Ling, L. Zhao, Grag: Graph retrieval-augmented generation, arXiv preprint arXiv:2405.16506 (2024).
[12] X. He, Y. Tian, Y. Sun, N. V. Chawla, T. Laurent, Y. LeCun, X. Bresson, B. Hooi, G-retriever: Retrieval-augmented generation for textual graph understanding and question answering, arXiv preprint arXiv:2402.07630 (2024).
[13] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering (2024).
[14] J. Chen, H. Lin, X. Han, L. Sun, Benchmarking large language models in retrieval-augmented generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 17754–17762.
[15] O. M. Group, Business process model and notation, 2014. URL: https://www.omg.org/spec/BPMN/2.0.2#document-metadata.
[16] M. Minor, E. Kaucher, Retrieval augmented generation with llms for explaining business process models, in: International Conference on Case-Based Reasoning, Springer, 2024, pp. 175–190.
[17] S. Feuerriegel, J. Hartmann, C. Janiesch, P. Zschech, Generative ai, Business & Information Systems Engineering 66 (2024) 111–126.
[18] F. Monti, F. Leotta, J. Mangler, M. Mecella, S. Rinderle-Ma, NL2ProcessOps: Towards LLM-Guided Code Generation for Process Execution, in: A. Marrella, M. Resinas, M. Jans, M. Rosemann (Eds.), Business Process Management Forum, volume 526, Springer Nature Switzerland, Cham, 2024, pp. 127–143. URL: https://link.springer.com/10.1007/978-3-031-70418-5_8. doi:10.1007/978-3-031-70418-5_8, series Title: Lecture Notes in Business Information Processing.
[19] T. Tel, M. Minor, Utilizing the structure of process models for guided generation of explanatory texts, in: International Conference on Case-Based Reasoning, Springer, 2025.