Leveraging Generative AI Tasks using CBR Methods with Structural Knowledge

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 (CBR) and to address operational efficiency with pre- and post-processing techniques. Additionally new techniques for retrieving and adapting information will be researched. These approaches are limited to process-oriented CBR with Business Process models as the main application domain. The methods used can also be applied to other process-oriented domains.

Keywords

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

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 consists of a vector database that contains embedded data chunks. This database is accessed when the user writes a query. A semantically similar chunk of information is retrieved to ground an LLMs answer with information. While this approach already improves LLM outputs, it faces challenges like imprecise retrieval and the potential for hallucination, irrelevant, toxic, or biased outputs. Effectively integrating retrieved information is also difficult [4].

This work is grounded in the principles of Case-Based Reasoning, a powerful problem-solving paradigm that leverages past experiences to address new challenges. Case-based reasoning has been formalized as a four-step process [5]: Retrieve, Reuse, Revise and Retain. This work focuses on the steps Retrieve and Reuse.

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.

ICCBR DC'26: Doctoral Consortium at ICCBR2026, August 16, 2026, Bremen, Germany

*Corresponding author.

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

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

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. There are many techniques for adaptation.

Recently, we saw many more advanced RAG and retrieval approaches using CBR methods [6, 7].

In [6] CBR-RAG, a framework that integrates CBR into RAG to improve LLM performance in knowledge intensive domains like legal-question answering is introduced. By leveraging CBR's retrieval cycle, indexing, and similarity metrics to provide richer, contextually relevant evidence, CBR-RAG enhances query prompts and produces significantly higher-quality, evidence-backed outputs.

Numerical reasoning over hybrid data often fails when models struggle to extract critical facts. In [7] a Case-Based Driven Retriever-generator model is proposed to retrieve and distinguish these critical facts. In the retrieval stage a golden explanation is used to help the retriever construct explicit templates for inferring critical facts. In the generator stage, the retrieval algorithm enhances the representation learning ability of the encoder.

The consortium OMG has established the standard BPMN [8] 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. The research field of process-oriented case-based reasoning (PO-CBR) [9] addresses those business needs. PO-CBR has achieved considerable success in assisting business process professionals with case-based methods.

Previous work with BPMN-models is found in [10], 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. The approaches in this thesis further explore an refine this idea with advanced case representations and retrieval methods.

In [11] 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 [12], 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.

This thesis focuses on further improving the quality of descriptive texts. By using sophisticated retrieval methods with smaller, better fitting cases in the case base, the retrieved in-context examples have a greater influence on the generating LLM. A current research gap lies in the explicit adaptation. While generator-evaluator pipelines [13] are already used in many domains, this thesis focuses on process-oriented domains, where a ground truth exists within the input to verify the correctness of the output. This allows an automatic evaluation and enables the adaptation process.

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

- **RO1:** Improve the retrieval component of RAG by minimizing redundant information, accurately predicting necessary information, and optimizing data structures for efficient retrieval.
- **RO2:** Advance the automated generation and evaluation of natural language descriptions for BPMN models using CBR retrieval and adaptation methods.

- **RO3:** Developing robust quality assessment mechanisms and error correction techniques for generated descriptive texts.
- **RO4:** Realizing CBR-Adaptation using LLMs with targeted Adaptation instructions to correct mistakes in descriptive texts in an agentic AI environment.

2.2. Approach / Methodology

For the generation and adaptation of texts, following methods are used: *LLMs Used:* Different iterations of Llama 3 and gpt-oss were used in experiments leading to a working generator-evaluator pipeline. The focus lies 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 text generation, 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.

Adaptation: To correct detected mistakes, CBR-Adaptation is implemented using LLMs. While the current approach uses fixed adaptation instructions with placeholders for element names, that are added to the prompt of the LLM, a more dynamic approach is possible. Using an LLM to generate an appropriate adaptation instruction when given a list of mistakes and also pinpointing the exact location of the mistake could improve the result of the adaptation.

Agentic AI: For larger problems that can be divided into smaller tasks, agentic approaches are used. Here the agents collaborate to solve problems by using results another agent produced to solve their own task. The use of agentic AI is merely a method for evaluating and adapting descriptive texts.

3. Progress Summary

ICCBR 2025: In my first paper [14], 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.

AIxDKE 2026: This paper [15] started as an extension of the ICCBR paper, where the algorithmic approach for finding a useful ordering of elements in the BPMN model was not optimal. To derive a more sophisticated explanation order, we explored augmenting business process models, by reconstructing a block-orientation.

The constructed block-orientation, which is needed for the explanation order, can be used for other structured explanation approaches. While the ICCBR 2025 results were promising regarding precision and recall, the hierarchical block-orientation supports more coherent texts while utilizing the structure of the process model for consistent results.

Another current approach is the automated evaluation and adaptation of generated texts. Here the main goal is the targeted adaptation of the descriptive texts, by identifying mistakes and only changing the parts that are wrong, while keeping the correct parts. The adaptation is currently accomplished by using adaptation rules to add missing elements, remove hallucinated elements and modify elements that are close to the correct solution.

4. Conclusion and Future Work

Building upon the foundational work presented in my submitted papers, 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.

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.

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. Is Text-to-Model in combination with graph similarity metrics a useful method to evaluate generated texts?
5. How can adaptation of generated results be further improved?

Question 1 and 2 were already partly explored in [14] 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 graph similarity as an evaluation metric. Here we transform a given descriptive text into a BPMN model to use similarity methods like graph edit-distance to compare the transformed BPMN model to the original model. This is used to specify the missing pieces in the given text.

Question 5 focuses on these missing pieces to find more sophisticated and efficient methods to incorporate the adaptation of mistakes in the text.

Declaration on Generative AI

The author declares that no generative artificial intelligence (AI) tools were employed in the creation, writing, or editing of this manuscript.

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] M. M. Richter, R. O. Weber, Case-based reasoning, Springer, 2016.

[6] N. Wiratunga, R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, B. Fleisch, CBR-RAG: case-based reasoning for retrieval augmented generation in llms for legal question answering, in: International Conference on Case-Based Reasoning, Springer, 2024, pp. 445–460.

[7] B. Feng, H. Gao, P. Zhang, J. Zhang, CBR-REN: A case-based reasoning driven retriever-generator model for hybrid long-form numerical reasoning, in: International Conference on Case-Based Reasoning, Springer, 2024, pp. 111–126.

[8] O. M. Group, Business process model and notation, 2014. URL: https://www.omg.org/spec/BPMN/2.0.2#document-metadata.

[9] M. Minor, S. Montani, J. A. Recio-García, Editorial: Process-oriented Case-based Reasoning, Inf. Syst. 40 (2014) 103–105.

[10] 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.

[11] S. Feuerriegel, J. Hartmann, C. Janiesch, P. Zschech, Generative AI, Business & Information Systems Engineering 66 (2024) 111–126.

[12] 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.

[13] A. Singh, A. Ehtesham, S. Kumar, T. T. Khoei, Agentic retrieval-augmented generation: A survey on agentic rag, 2025. URL: https://arxiv.org/abs/2501.09136. arXiv:2501.09136.

[14] 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.

[15] T. Tel, H. V. Nguyen, M. Minor, Augmenting structural knowledge on process models to improve explainability, in: 2026 International Conference on AI x Data and Knowledge Engineering (AIxDKE), IEEE, 2026, pp. 12–17.