# Semantic Verification in Large Language Model-based Retrieval Augmented Generation

**Andreas Martin**[1], **Hans Friedrich Witschel**[1], **Maximilian Mandl**[2] **and Mona Stockhecke**[2]
[1]FHNW University of Applied Sciences and Arts Northwestern Switzerland,
School of Business, Riggenbachstrasse 16, 4600, Olten, Switzerland
andreas.martin@fhnw.ch, hansfriedrich.witschel@fhnw.ch
[2]Nagra, Hardstrasse 73, 5430 Wettingen, Switzerland
maximilian.mandl@nagra.ch, mona.stockhecke@nagra.ch

## Abstract

This position paper presents a novel approach of semantic verification in Large Language Model-based Retrieval Augmented Generation (LLM-RAG) systems, focusing on the critical need for factually accurate information dissemination during public debates, especially prior to plebiscites e.g. in direct democracies, particularly in the context of Switzerland. Recognizing the unique challenges posed by the current generation of Large Language Models (LLMs) in maintaining factual integrity, this research proposes an innovative solution that integrates retrieval mechanisms with enhanced semantic verification processes. The paper outlines a comprehensive methodology following a Design Science Research approach, which includes defining user personas, designing conversational interfaces, and iteratively developing a hybrid dialogue system. Central to this system is a robust semantic verification framework that leverages a knowledge graph for fact-checking and validation, ensuring the correctness and consistency of information generated by LLMs. The paper discusses the significance of this research in the context of Swiss direct democracy, where informed decision-making is pivotal. By improving the accuracy and reliability of information provided to the public, the proposed system aims to support the democratic process, enabling citizens to make well-informed decisions on complex issues. The research contributes to advancing the field of natural language processing and information retrieval, demonstrating the potential of AI and LLMs in enhancing civic engagement and democratic participation.

## Introduction

In the rapidly evolving landscape of artificial intelligence and natural language processing, the integration of complex information retrieval with advanced language models presents both unprecedented opportunities and challenges. This paper aims to explore and address these challenges, particularly focusing on the aspect of semantic verification within the context of large language models (LLMs) augmented by retrieval mechanisms.

### Context and Background

The advent of large language models has revolutionized the way we interact with information, offering transformative capabilities in generating, summarizing, and interpreting vast amounts of data. However, the effective utilization of these models in practical applications, especially those involving intricate technical domains, requires not only the generation of coherent and contextually relevant responses – which sometimes also need to be simplified to be understandable for broad target audiences – but also a high degree of semantic accuracy and reliability.



### Problem Identification

Despite their sophistication, current large language models often struggle with maintaining semantic integrity, leading to issues like information misrepresentation or ”hallucination,” where the model generates plausible but factually incorrect information. This problem is amplified in scenarios requiring the integration and summarization of complex, multi-source documents, where the fidelity of information is paramount. It is also amplified in cases where the content from highly complex (e.g. very technical) documents needs to be simplified in order to be understandable – simplification always introduces a risk to reduce factual correctness.

### Purpose and Scope

The purpose of this paper is to delve into the development of a framework that enhances the semantic verification process in large language models, particularly in environments where they are supplemented by retrieval systems. We aim to examine the methodologies, technologies, and strategies that can be employed to ensure the semantic integrity of the information generated by these models. The scope of this research encompasses the study of existing models, identification of their limitations, and the proposal of innovative solutions to enhance their accuracy and reliability in diverse applications.

This exploration is critical not just for the advancement of language models but also for their practical deployment in fields where the accuracy and dependability of information are crucial. The findings and solutions presented in this paper seek to contribute significantly to the field of natural language processing and information retrieval, paving the way for more reliable and trustworthy AI-driven decision-making systems.

## Related Work

This section examines the existing literature and research pertaining to Large Language Models (LLMs) and their integration with information retrieval systems, particularly focusing on the challenges and advancements in semantic verification.

Recent advancements in LLMs have opened new avenues in natural language processing, enabling machines to understand and generate text comparable with the level of a well-versed human. These models have been increasingly used for "Long-Form Question Answering," where they synthesize information from various sources to answer open-ended questions (Fan et al. 2019; Krishna, Roy, and Iyyer 2021; El-Kassas et al. 2021). However, the challenge lies in ensuring the semantic correctness of these answers, as highlighted in Ji et al. (2023), Rawte, Sheth, and Das (2023), and Li et al. (2021). The research in this domain has largely been divided into extractive and abstractive methods (Ji et al. 2023). Extractive methods, while largely maintaining the original text's integrity by selecting the most relevant passages from it, often fall short in terms of readability and user-specific information tailoring. On the other hand, abstractive methods, despite offering more fluid and coherent summaries, are prone to factual inaccuracies and hallucinations.

One significant gap in the current landscape, as identified in Ji et al. (2023) and Rawte, Sheth, and Das (2023), is the potential for "hallucination" or semantic drift in LLMs, where the model generates information that is not rooted in the source material. This issue is critical, especially in scenarios demanding high factual accuracy. Furthermore, the traditional approach of relying on human verification (aided by the linkage of LLM statements with passages in source documents), as discussed in Huang and Chang (2023) and Zhang et al. (2023), poses scalability challenges, especially when dealing with large volumes of data.

The use of Knowledge Graphs for semantic verification, as explored in Balepur et al. (2023) and Lenat and Marcus (2023), presents a promising approach, yet it is not without its limitations. The process of creating and maintaining these graphs, especially for specialized domains, can be labor-intensive and prone to errors, as noted in Perez-Beltrachini and Gardent (2017). Additionally, the integration of Knowledge Graphs with LLMs for fact verification, as indicated in Gawade and Bhattacharyya (2023), Dhingra et al. (2019), and Nie et al. (2019), is still an area requiring more research and development.

Recent works have explored various methodologies to mitigate these challenges. The concept of "Prompt Engineering" and "In-Context Learning," as discussed in Rubin, Herzig, and Berant (2022) and Zhou et al. (2023), shows potential in improving the relevance and accuracy of LLM-generated content. However, ensuring the semantic integrity of simplified texts remains a formidable challenge Lu et al. (2023), Li et al. (2022). Moreover, the role of fine-tuning LLMs to specific domains is crucial in enhancing their effectiveness but requires further exploration to avoid the pitfalls of factual inaccuracies and ethical concerns.

In summary, while there have been significant strides in the development of LLMs and their integration with retrieval systems, there remains a crucial need for more robust semantic verification mechanisms. This paper aims to build upon these findings, addressing the gaps and exploring innovative solutions to enhance the reliability and applicability of LLMs in complex, information-rich environments.

## Methodology

This work adopts a Design Science Research (DSR) methodology, a robust framework suitable for developing and investigating technology-based solutions. The DSR approach encompasses several iterative stages: "Awareness and Requirements," "Suggestion and Design," "Development," "Evaluation," and "Deployment and Communication." Each stage plays a critical role in ensuring the successful development and application of the proposed solution.

### Awareness and Requirements

In the initial phase, we focus on constructing personas that represent the diverse range of users, their existing knowledge in the given field and their specific information needs. This step is crucial for understanding the varying contexts in which the Retrieval Augmented Generation (RAG) agent will operate. By defining these personas, we can better tailor the solution to meet the distinct requirements of different user groups, thus ensuring relevance and applicability.

### Suggestion and Design

The next phase involves the suggestion and design of the RAG agent, incorporating multiple Large Language Models (LLMs). This stage is iterative, starting with the conceptualization of the RAG agent, followed by refining the design based on feedback and insights gained from the initial personas.

Subsequently, the terminology database is transformed into a comprehensive knowledge graph. This transformation is critical for later stages of semantic verification, as it provides a structured and accessible repository of information that can be used to validate the outputs of the LLMs.

### Development

Following the design phase, the project enters the development stage, where the following three research questions (RQs) are iteratively addressed:

1. Prompt Engineering/In-Context Learning: This involves tailoring the interaction of the RAG agent with the user, considering the specific context and requirements of different personas. In the context of a public debate in a direct democracy, such personas could be e.g. citizens, journalists, domain experts or politicians, all with different prior knowledge and questions towards the system.
2. Semantic Verification: This step focuses on ensuring the accuracy and factual correctness of the information generated by the RAG agent, utilizing the knowledge graph as a key resource.
3. Target-Group Appropriate Fine-Tuning: The final research question addresses the customization of responses

to suit the specific needs and understanding levels of various user personas. There, the knowledge graph could be utilized to generate data for fine-tuning the LLM.

### Evaluation

The evaluation stage assesses the effectiveness, efficiency, and applicability of the developed solution. This process involves rigorous testing and validation of the understandability and factual correctness of the system's answers across different scenarios and user groups.

### Deployment and Communication

In the final phase, the research findings and the developed solution are disseminated to the broader community. This stage not only involves the deployment of the solution but also the communication of the research outcomes, insights, and potential implications to stakeholders, practitioners, and the academic community. This dissemination is crucial for fostering further research, collaboration, and innovation in the field.

## Approach

The approach of this research aligns with the Conversational AI lifecycle, systematically developing and refining components crucial for a robust Retrieval Augmented Generation (RAG) system with Large Language Models (LLMs) and semantic verification. This lifecycle encompasses defining personas, designing conversations, establishing the RAG-LLM-semantic-verification architecture, and iterative development and training.

### Personas

The first step involves defining personas representing various user groups, each with unique information needs and interaction patterns. These personas guide the conversational design, ensuring that the system's responses are tailored to different user contexts and requirements.

### Conversational Design

Conversational design focuses on creating interaction flows that are intuitive, efficient, and user-friendly. It involves structuring dialogues in a way that facilitates clear and effective communication between the user and the AI system.

### Architecture

As depicted in Figure 1, the basic architecture of our hybrid dialogue system integrates an RAG mechanism with LLMs. This system starts with interpreting and enriching user queries in the Dialog Manager, considering the query type and associated persona (see RQ1 above). The enriched query, or "Prompt," is then processed by the Language Model Agent.

The agent's responsibilities include finding relevant text passages from the document base (Text Retrieval), summarizing, and possibly simplifying these passages using an

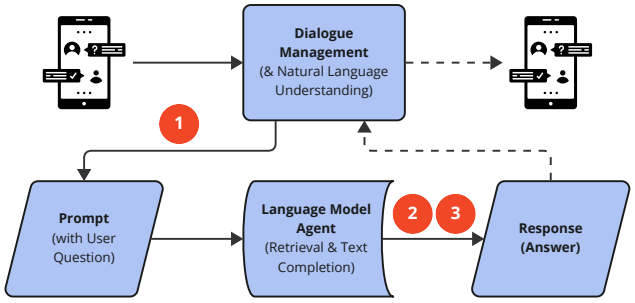


Figure 1: General architecture of the hybrid dialogue system. Red circles refer to our research questions.

LLM (see RQ3). The final step before presenting the response to the user is the crucial phase of semantic verification, ensuring the correctness and consistency of the information (RQ2).

### Iterative Development and Training

The system is iteratively developed and trained, focusing on the three core research areas: Prompt Engineering/In-Context Learning, Semantic Verification, and Target-Group Appropriate Fine-Tuning. Each area is refined through cycles of development, testing, and feedback.

### Semantic Verification

Semantic verification, illustrated in Figure 2, is at the heart of our approach. The foundation for verifying statements lies in a well-curated Knowledge Graph, which includes concepts, relationships, and rules for verification derived from the terminology database. The Knowledge Graph is used to translate the summaries of retrieval results into graph structures (semantic triples) using semantic parsing techniques (Lenat and Marcus 2023; Martin 2023). This graph-based structure allows for efficient reasoning processes to identify inconsistencies or inaccuracies.

A critical aspect of semantic verification is the implementation of a fact-driven re-ranking system, as described by Xie et al. (2023). This system re-evaluates the generated responses based on factual accuracy, ensuring that the most reliable and accurate information is prioritized in the final response.

The innovation lies in the combination of these elements – the Knowledge Graph, semantic parsing, and fact-driven re-ranking – which work together to minimize the risks of hallucination and semantic drift inherent in LLMs (Ji et al. 2023; Rawte, Sheth, and Das 2023). The research challenge and novelty involve developing scalable, efficient verification rules and integrating them seamlessly with the LLM-generated content. By addressing these challenges, the project aims to significantly enhance the reliability and trustworthiness of AI-driven information systems, particularly in contexts where factual accuracy is paramount.

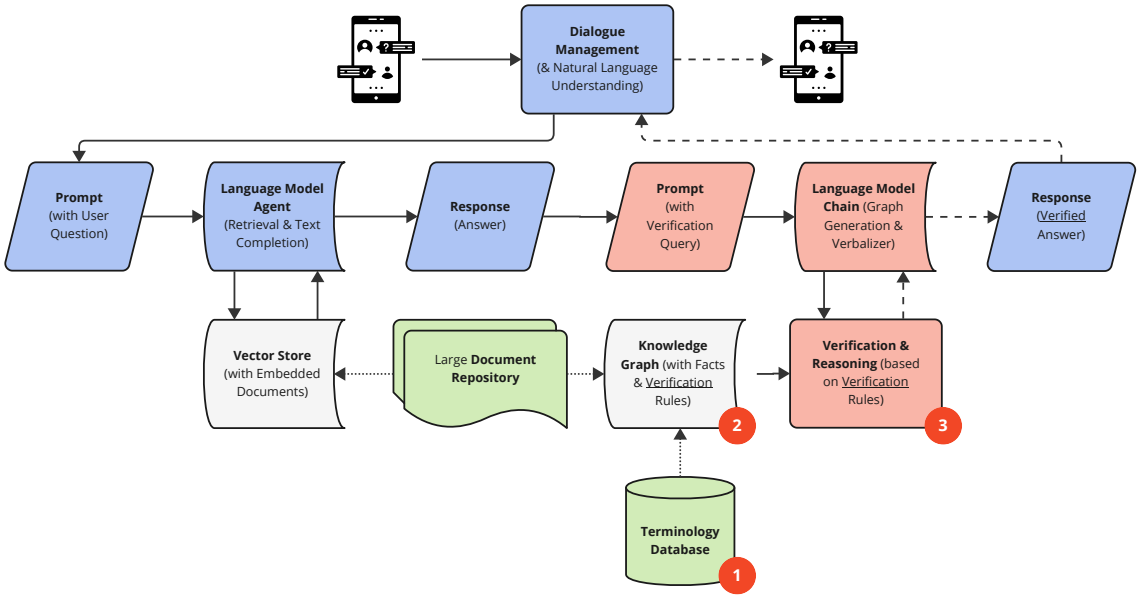


Figure 2: General architecture of the hybrid dialogue system

## Discussion and Conclusion

The application of large language models (LLMs) in generating and verifying information plays a pivotal role in contemporary information dissemination, especially in contexts like Switzerland, where direct democracy forms the bedrock of societal decision-making. This paper's focus on enhancing semantic verification in retrieval-augmented LLM systems has profound implications in ensuring the accuracy and reliability of information, a necessity for informed decision-making in democratic processes.

### Discussion

In Switzerland's direct democracy, the importance of providing factually correct information to the public cannot be overstated. The populace's ability to make informed decisions on complex issues, ranging from environmental policies to technological advancements, hinges on their access to accurate, comprehensible, and verified information. The intricate nature of such topics often requires distilling complex technical data (e.g. long and highly specialised reports written by engineers) into understandable formats without losing factual integrity. This is where the advancement in semantic verification within LLMs, as proposed in this research, becomes crucial.

The development and iterative refinement of a retrieval-augmented LLM system, underpinned by robust semantic verification, marks a significant step towards addressing this need. By ensuring that the generated information is not only contextually relevant but also factually accurate, the system aligns with the foundational requirements of democracy. It aids in mitigating the risks associated with misinformation and "hallucination" tendencies of current LLMs, thus fostering a more informed and engaged citizenry.

However, the implementation of such systems is not without its challenges. The need for continuous development and adaptation of the knowledge graph, the potential complexity of managing diverse personas, and the balance between simplifying technical information while maintaining its accuracy are areas that require ongoing attention and innovation.

### Conclusion

The research presented in this paper contributes to the field of natural language processing and information retrieval, particularly in the realm of semantic verification. The proposed approach, centered around a retrieval-augmented LLM with enhanced semantic verification capabilities, is poised to play a crucial role in contexts demanding high factual accuracy, such as in Swiss direct democracy.

As the system evolves, it will not only facilitate the dissemination of reliable information to a diverse audience but also set a precedent for the application of AI in supporting democratic processes. This advancement underscores the potential of AI and LLMs in serving society beyond commercial and technological domains, extending into the realm of enhancing democratic engagement and decision-making.

Ultimately, the success of this endeavor will be measured not just by the sophistication of its technology but by its contribution to the informed participation of citizens in the democratic process. As the system continues to develop, it holds the promise of becoming an indispensable tool in the arsenal of a society striving for informed decision-making and active civic engagement.

## Acknowledgments

This work has been funded by the Innosuisse – Swiss Innovation Agency of the Swiss Confederation under grant 109.093 IP-ICT.

## References


Balepur, N.; Huang, J.; Moorjani, S.; Sundaram, H.; and Chang, K. C.-C. 2023. Mastering the ABCDs of Complex Questions: Answer-Based Claim Decomposition for Fine-grained Self-Evaluation. *arXiv preprint arXiv:2305.14750*.

Dhingra, B.; Faruqui, M.; Parikh, A.; Chang, M.-W.; Das, D.; and Cohen, W. W. 2019. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*.

El-Kassas, W. S.; Salama, C. R.; Rafea, A. A.; and Mohamed, H. K. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165: 113679.

Fan, A.; Jernite, Y.; Perez, E.; Grangier, D.; Weston, J.; and Auli, M. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3558–3567. Florence, Italy: Association for Computational Linguistics.

Gawade, S.; and Bhattacharyya, P. 2023. Knowledge Graph and Deep Learning Assisted Question Answering and Ontology Construction: A Survey.

Huang, J.; and Chang, K. C.-C. 2023. A key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Krishna, K.; Roy, A.; and Iyyer, M. 2021. Hurdles to Progress in Long-form Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4940–4957. Online: Association for Computational Linguistics.

Lenat, D.; and Marcus, G. 2023. Getting from Generative AI to Trustworthy AI: What LLMs might learn from Cyc. *arXiv preprint arXiv:2308.04445*.

Li, C.; Bi, B.; Yan, M.; Wang, W.; and Huang, S. 2021. Addressing semantic drift in generative question answering with auxiliary extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 942–947.

Li, J.; Li, Z.; Ge, T.; King, I.; and Lyu, M. R. 2022. Text revision by on-the-fly representation optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10956–10964.

Lu, J.; Li, J.; Wallace, B. C.; He, Y.; and Pergola, G. 2023. Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. *arXiv preprint arXiv:2302.05574*.

Martin, A. 2023. AAAI-MAKE 2023: Challenges requiring the combination of machine learning and knowledge engineering. *AI magazine*, 44: 204–205.

Nie, F.; Yao, J.-G.; Wang, J.; Pan, R.; and Lin, C.-Y. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2673–2679.

Perez-Beltrachini, L.; and Gardent, C. 2017. Analysing Data-To-Text Generation Benchmarks. In *Proceedings of the 10th International Conference on Natural Language Generation*, 238–242. Santiago de Compostela, Spain: Association for Computational Linguistics.

Rawte, V.; Sheth, A.; and Das, A. 2023. A Survey of Hallucination in Large Foundation Models. *arXiv preprint arXiv:2309.05922*.

Rubin, O.; Herzig, J.; and Berant, J. 2022. Learning To Retrieve Prompts for In-Context Learning. ArXiv:2112.08633 [cs].

Xie, Q.; Hu, J.; Zhou, J.; Peng, Y.; and Wang, F. 2023. Factreranker: Fact-guided reranker for faithful radiology report summarization. *arXiv preprint arXiv:2303.08335*.

Zhang, S.; Pan, L.; Zhao, J.; and Wang, W. Y. 2023. Mitigating Language Model Hallucination with Interactive Question-Knowledge Alignment. *arXiv preprint arXiv:2305.13669*.

Zhou, H.; Wan, X.; Proleev, L.; Mincu, D.; Chen, J.; Heller, K.; and Roy, S. 2023. Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering. ArXiv:2309.17249 [cs].