This paper presents our studies with Large Language Models (LLM) and Retrieval-augmented-Generation (RAG). The goals of these architectures are to reduce or eliminate hallucinations in LLMs, overcome the limitations of context window length, e.g., GPT4 128k, and allow LLMs to generate up-to-date content for a query without retraining or fine-tuning the LLM.

Various proposed solutions for RAGs are to find the best matching context for a query using similarity search and extend the context of the query in the LLM to generate a response. Some RAG systems provide canned text as potential answers for queries, others organize text fragments and indexes (including links to sources) in Knowledge Graphs (KG). To the best of our knowledge, formal semantic knowledge representations are not utilized in these RAG systems, leading to problematic responses in the documented evaluations as measured using user judgements of the response relevance.

Our approach involves KG generation from text sources using Description Logic (DL), ontological classification, and axiom environments to control the extraction semantically. This way we enable DL-based logical operations and reasoning over sparse concept and relation extraction from text sources. We can control KG generation using a domain-specific ontology. We picked specific medical domains for our experiments.

In our experiments, we used PubMed articles related to mental health, Parkinson’s, and Alzheimer's. For each subdomain, we created a corpus of 5,000 articles containing the abstracts, results, keywords, and author information.

We compare the extraction of KG from these abstracts using three different techniques: a. LLM-based entity and relation extraction, b. classical NLP-pipeline parsing and entity-relation extraction, and c. language-model-based methods using entity and relation extraction. In our presentation, we discuss the results and problems of these three different frameworks for triple extraction and ontology learning.