Improving LLM Reasoning Through Ontology-driven Knowledge Graphs

R. Shrivastava S.L. Anusha Chebolu M. Kodandapani Naidu T. Jayaprakash J. Decatur T. Sun
A. Bajpai R. Wang D. Cavar NLP-Lab

Objectives

How to enhance the explainability and factual grounding of LLM-based Retrieval-Augmented Generation (RAG) [1, 2] systems through semantic knowledge (graphs) in the medical domain.

- Construct a classical RAG pipeline using PubMed documents and design evaluation questions collaboratively with human experts and LLMs for use as the baseline environment.
- Evaluate and compare triple extraction methods from biomedical text using (a) LLMs, (b) classical NLP pipelines, and (c) hybrid approaches as the knowledge backend for RAGs.
- Apply these extraction techniques to domain-specific corpora.

Introduction

Retrieval-Augmented Generation (RAG) systems improve Large Language Models (LLMs):

- by grounding responses in external sources
- reducing hallucinations
- extending context with up-to-date information on outdated LLMs

Classical RAG pipelines:

- retrieve unstructured text based on similarity (embeddings)
- lack semantic control and interpretability
- lack reasoning capabilities

Our experiments:

- a graph-based RAG approach that uses ontologies and Description Logic
- generate and query Knowledge Graphs (KGs) generated from biomedical texts
- enable structured reasoning and improve the relevance and explainability of LLM outputs

Evaluation:

- using PubMed articles from domains: mental health, COVID, Obesity, Parkinson's, and Alzheimer's

Dataset Description

Constructed five biomedical corpora by querying PubMed using domain-specific search terms and publication date filters (2024–2025)

- **Mental health**: 6,057 articles
- **COVID-19**: 6,759 articles
- **Parkinson's disease**: 5,026 articles
- **Obesity**: 4,898 articles
- **Alzheimer's disease and dementia**: 7,186 articles

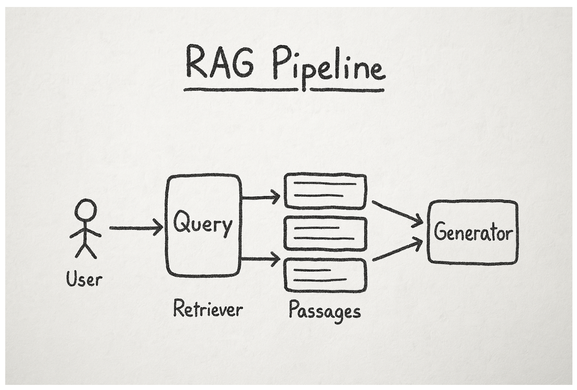
Each article is available as an individual JSON file in MEDLINE format.

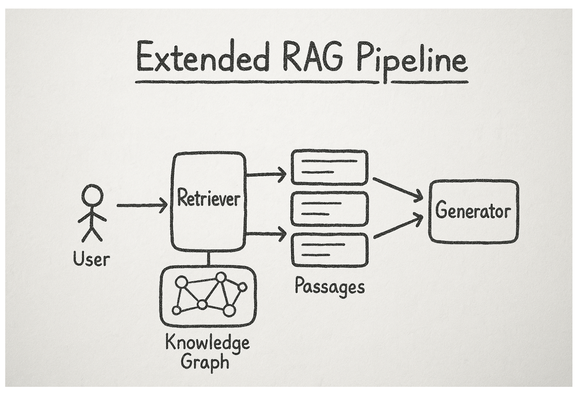
Corpus Purpose

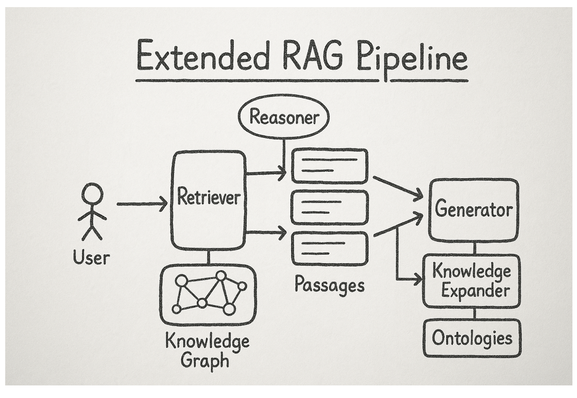
- retrieval-augmented generation (RAG) data and index
- knowledge graph construction and concept/relation extraction

Questions for Text Chunks

- LLM-generated questions: 100 text chunks from each domain with maximally 5 questions generated by LLMs (GPT-4o, Claude 3 Opus April 2025)[3, 4]
- Human-generated questions: 50 text chunks from each domain with at least one question







Methods

RAG Classic

- Corpus articles chunked and stored in Vector DB
- Vectorization model Sentence Transformer *all-MiniLM-L6-v2*
- Similarity query using vectorized query and stored text chunks
- Context augmentation using matching text chunks

Knowledge Graph RAG

- Corpus articles mapped to concept-relation triples and stored in Graph DB (Neo4j)
- Query mapped to:
  - Cypher query on Graph DB generated by LLM
  - Concept-based query to Graph DB using entities mentioned in query
- Graph to text or triple set as context augmentation

Semantic RAG

- Ontology and Knowledge Graph from corpus articles
- Reasoner-extended Knowledge Graph
- Query using graph DB and SPARQL
- Context generation from resulting subgraph

Results

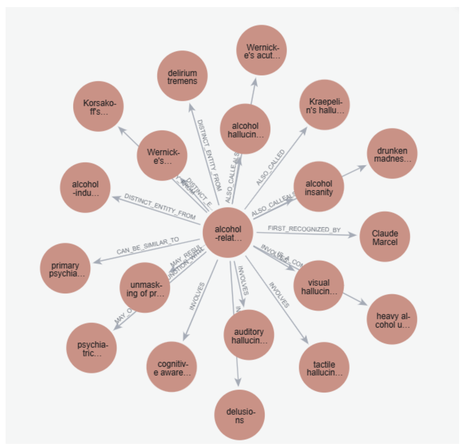


Figure 1:Subgraph related with "alcohol-related psychosis"

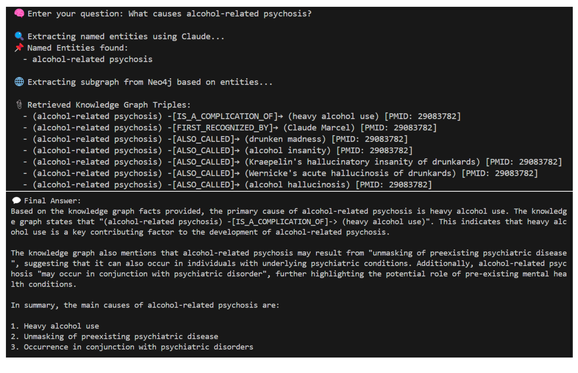


Figure 2:Sample response using KG-RAG

Conclusion

- Retrieval results:
  - Classical RAG approaches yield moderate precision and recall (MRR 0.45, Success@3 0.57), which is insufficient for high-stakes medical applications where higher accuracy is essential.
  - The substantial performance gap between Top-1 and Top-3 retrieval indicates classical vector similarity alone cannot consistently identify the most relevant information.
- Limitations:
  - We cannot capture the intentions and goals of users using extended information.
  - Learning ontologies from text sources is challenging and, at best, limited to a few concepts and relation types.
- Ongoing experiments:
  - Semantic graphs with reasoning capabilities.
  - True end-to-end evaluation by human subjects.

References

[1] P. Lewis et al.
Retrieval-augmented generation for knowledge-intensive nlp tasks.
Advances in neural information processing systems, 33:9459–9474, 2020.

[2] D. Edge et al.
From local to global: A graph rag approach to query-focused summarization.
arXiv preprint arXiv:2404.16130, 2024.

[3] A. Hurst et al.
Gpt-4o system card.
arXiv preprint arXiv:2410.21276, 2024.

[4] D. Kevian et al.
Capabilities of large language models in control engineering: A benchmark study on gpt-4, claude 3 opus, and gemini 1.0 ultra.
arXiv preprint arXiv:2404.03647, 2024.

[5] C.A. Stewart et al.
Indiana university pervasive technology institute, 2017.

Acknowledgements

The authors acknowledge the Indiana University Pervasive Technology Institute (`https://pti.iu.edu/`) for providing supercomputing and storage resources that have contributed to the research results reported within [5].

Contact Information

- Web: http://nlp-lab.org
- Email: dcavar@iu.edu