CS 598 JH ASSIGNMENT REPORT: ENHANCING KG-RAG FOR BIOMEDICAL QUESTION ANSWERING

Xinxi Lyu **NetID: seanlyu2** **School ID: 669479137** **Email: seanlyu2@illinois.edu**

1 INTRODUCTION

This report is for the course assignment for CSE 598 JH: Enhancing KG-RAG for Biomedical Question Answering. In this work, we explore the framework of Knowledge Graph (KG) enhanced Retrieval Augmented Generation (RAG). Given a query, KG-RAG retrieves external knowledge from knowledge graphs and appends to the context of a Large Language Model (LLM) to facilitate it to generate the answer.

In this assignment, we use SPOKE (Morris et al., 2023), a massive biomedical KG as the source, and evaluate KG-RAG on the multiple-choice subset of BiomixQA (Soman et al., 2023), a biomedical question-answering datasets.

The assignment includes reproducing a baseline of using the simplest form of KG-RAG and implementing three distinct strategies to enhance the performance on top of the baseline.

We aim to gain a better understanding of the effect of using external knowledge beyond the implicit knowledge that LLM learned from pretraining. We also learn the power of structuring external knowledge and how human experience can enhance the model's performance.

2 METHOD

2.1 CONTEXT RETRIEVAL

We denote the given KG as $\mathcal{G}$ and the query as q. We treat the generation LLM as a function that maps an input text to an output response, detonated $\mathbf{M}$, the node embedding model and context embedding model as function that maps an input text to an embedding vector space, denoted as $\mathbf{En}$ and $\mathbf{Ec}$. The retrieval pipeline is as follows:

1. We prompt the LLM to extract disease entities from the query to obtain $\mathbf{M}(q|\text{entity extraction prompt}) = \{e_1, e_2, \cdots, e_n\}$ for an arbitrary n.
2. In G, we search for the most similar node to the each disease entity: $n^\star = \arg\max_{n \in \mathcal{G}} \cos\big(\mathbf{En}(n), \mathbf{En}(e_i)\big)$ for each e_i, and combine them into a single list of nodes.
3. For each node n_i, we retrieve the stored list of contexts $C_i = \{c_{i1}, c_{i2}, \cdots, c_{im}\}$. We then only keep the top K (predefined) contexts whose similarity against the query q is below a threshold τ predefined: $\text{TopK}\big\{c \in C_i \big| \cos\big(\mathbf{Ec}(c), \mathbf{Ec}(q)\big) \geq \tau\big\}$.
4. We concatenate all qualified contexts into a single text string as the final context $\mathcal{C}$.

2.2 ARGUMENTATION AND GENERATION

Original (Baseline). The original KG-RAG method that appends the retrieved context from the KG directly to the input query to obtain $\mathbf{M}(\mathcal{C} \oplus q)$.

We experiment with three argumentation method to improve on the original KG-RAG:

Jsonlization. We structure the retrieved context from the KG into the form of JSON so that the model can understand the relations within the retrieved context. Specifically, we prompt the LLM itself to perform the jsonlization. The whole process is formally: $\mathbf{M}\big(\mathbf{M}(\mathcal{C}|\text{jsonlization prompt}) \oplus q\big)$.

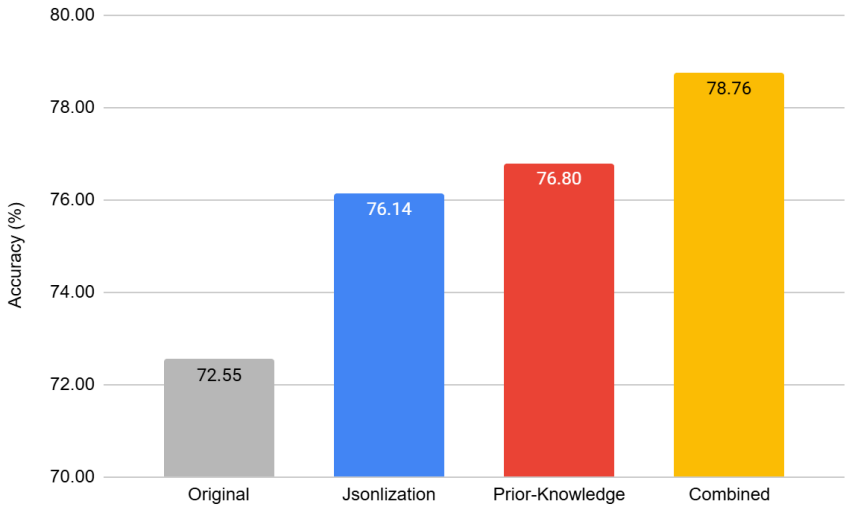

Figure 1: The performance of four strategies of KG-RAG on BiomixQA.

Prior-Knowledge. To help the LLM to utilize the given context and make the decision better, we obtain prior knowledge about the biomedical domain and express them as a list of short knowledge points, denoted as p. We concatenate these knowledge information with the context and feed them into the LLM, formally: $\mathbf{M}(\mathcal{C} \oplus p \oplus q)$.

Combined. We apply both Jsonlization and Prior-Knowledge together, formally: $\mathbf{M}\big(\mathbf{M}(\mathcal{C}|\text{jsonlization prompt}) \oplus p \oplus q\big)$.

3 Implementation

We use gemini-2.0-flash (Comanici et al., 2025) for entity extraction, jsonlization, and final answer generation; we use *sentence-transformers/all-MiniLM-L6-v2* as the node embedding model and *pritamdeka/S-PubMedBert-MS-MARCO* as the context embedding model; we use SPOKE (Morris et al., 2023) as the KG.

Our prompt for performing jsonlization is: *You are an expert in organizing a knowledge graph triple list and constructing a JSON structure from it. Please only report your constructed JSON.*

To obtain the prior knowledge about the biomedical domain, we use the response from ChatGPT as a proxy for human knowledge with the following prompt: *Suggests a few concrete short knowledge points that would help answer biomedical questions that asks about the associated Gene / Variant of certain diseases. Some examples: - Provenance & Symptoms information is useless. - Similar diseases tend to have similar gene associations.*

Our evaluation is conducted on the multiple-choice subset of BiomixQA (Soman et al., 2023) using accuracy (Exact Match) as the metric.

4 Experimental Results

Our experimental results are shown in Figure 1. Jsonlization improves the original KG-RAG by 3.6%, suggesting that structuring the retrieved knowledge helps LLM understanding. Prior-Knowledge improves the original KG-RAG by 4.4%, suggesting that general human experience on the domain is meaningful for LLM's success. Finally, combining these methods together raise performance of KG-RAG from 72.55% to 78.76%, achiving the best performance across the four strategies we experiment with.

REFERENCES

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

John H Morris, Karthik Soman, Rabia E Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C Meng, Conrad C Huang, Gabriel Cerono, Gundolf Schenk, Angela Rizk-Jackson, et al. The scalable precision medicine open knowledge engine (spoke): a massive knowledge graph of biomedical information. *Bioinformatics*, 39(2):btad080, 2023.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, et al. Biomedical knowledge graph-enhanced prompt generation for large language models. *arXiv preprint arXiv:2311.17330*, 2023.