Bioinformatics, 2024, **40(9)**, btae560
https://doi.org/10.1093/bioinformatics/btae560
Advance Access Publication Date: 17 September 2024
Original Paper

Databases and ontologies

Biomedical knowledge graph-optimized prompt generation for large language models

Karthik Soman[1], Peter W. Rose[2], John H. Morris [3], Rabia E. Akbas[1], Brett Smith[4], Braian Peetoom[1], Catalina Villouta-Reyes[1], Gabriel Cerono[1], Yongmei Shi[5], Angela Rizk-Jackson[5], Sharat Israni[5], Charlotte A. Nelson[6], Sui Huang [4], Sergio E. Baranzini [1,*]

[1]Department of Neurology, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94158, United States
[2]San Diego Supercomputer Center, University of California, San Diego, CA 92093, United States
[3]Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, CA 94158, United States
[4]Institute for Systems Biology, Seattle, WA 98109, United States
[5]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA 94158, United States
[6]Mate Bioservices, Inc. Swallowtail Ct., Brisbane, CA 94005, United States

*Corresponding author. Department of Neurology, Weill Institute for Neurosciences, University of California San Francisco, 675 Nelson Rising Ln., San Francisco, CA 94158, United States. E-mail: sergio.baranzini@ucsf.edu (S.E.B.)

Associate Editor: Peter Robinson

Abstract

Motivation: Large language models (LLMs) are being adopted at an unprecedented rate, yet still face challenges in knowledge-intensive domains such as biomedicine. Solutions such as pretraining and domain-specific fine-tuning add substantial computational overhead, requiring further domain-expertise. Here, we introduce a token-optimized and robust Knowledge Graph-based Retrieval Augmented Generation (KG-RAG) framework by leveraging a massive biomedical KG (SPOKE) with LLMs such as Llama-2-13b, GPT-3.5-Turbo, and GPT-4, to generate meaningful biomedical text rooted in established knowledge.

Results: Compared to the existing RAG technique for Knowledge Graphs, the proposed method utilizes minimal graph schema for context extraction and uses embedding methods for context pruning. This optimization in context extraction results in more than 50% reduction in token consumption without compromising the accuracy, making a cost-effective and robust RAG implementation on proprietary LLMs. KG-RAG consistently enhanced the performance of LLMs across diverse biomedical prompts by generating responses rooted in established knowledge, accompanied by accurate provenance and statistical evidence (if available) to substantiate the claims. Further benchmarking on human curated datasets, such as biomedical true/false and multiple-choice questions (MCQ), showed a remarkable 71% boost in the performance of the Llama-2 model on the challenging MCQ dataset, demonstrating the framework's capacity to empower open-source models with fewer parameters for domain-specific questions. Furthermore, KG-RAG enhanced the performance of proprietary GPT models, such as GPT-3.5 and GPT-4. In summary, the proposed framework combines explicit and implicit knowledge of KG and LLM in a token optimized fashion, thus enhancing the adaptability of general-purpose LLMs to tackle domain-specific questions in a cost-effective fashion.

Availability and implementation: SPOKE KG can be accessed at https://spoke.rbvi.ucsf.edu/neighborhood.html. It can also be accessed using REST-API (https://spoke.rbvi.ucsf.edu/swagger/). KG-RAG code is made available at https://github.com/BaranziniLab/KG_RAG. Biomedical benchmark datasets used in this study are made available to the research community in the same GitHub repository.

1 Introduction

Large language models (LLM) have shown impressive performance in solving complex tasks across various domains that involve language modeling and processing (Zhao *et al.* 2023). LLMs are pre-trained on a large corpora of text data in a self-supervised learning framework which can be either masked language modeling (e.g. BERT-like models) (Kenton and Toutanova, 2019, Liu, 2019) or auto-regressive framework (GPT-like models) (http://arxiv.org/abs/2005.14165) (Lu *et al.* 2022). This pretraining encodes knowledge about the language in the model parameters. Similar to the transfer learning approach commonly used in deep neural networks, this implicit knowledge can be refined through supervised training to excel in a range of domain-specific tasks (Luo *et al.* 2023). Nevertheless, the "implicit representation" of knowledge in LLM has also been shown to generate non-factual information despite linguistically coherent answers (i.e. "hallucination") as a response to the input prompt (Maynez *et al.* 2020, Raunak *et al.* 2021, Ji *et al.* 2023). This issue poses a significant challenge for the adoption of LLM in domains with stringent requirements for accuracy, such as biomedicine.

Various strategies have been introduced to address hallucinations in LLMs. One such solution involves the utilization of domain-specific data for pretraining the LLM, rather than

Received: 15 May 2024; **Revised:** 29 August 2024; **Editorial Decision:** 4 September 2024; **Accepted:** 15 September 2024

relying on generic text corpora. This approach has led to the creation of models such as PubMedBERT (Lee *et al.* 2020, Gu *et al.* 2021) BlueBERT (Peng *et al.* 2019), SciBERT (Beltagy *et al.* 2019), ClinicalBERT (Huang *et al.* 2019), BioGPT (Luo *et al.* 2022), Med-PaLM (Singhal *et al.* 2022), and BioMedGPT (Luo *et al.* 2023). However, pretraining an LLM from scratch imposes a significant computational cost and time overhead to attain the desired human-like performance. An alternative approach, known as prompt tuning, was recently proposed as a means to enhance LLM performance, for instance through the use of zero-shot (Kojima *et al.* 2022), few-shot (http://arxiv.org/abs/2005.14165), and Chain-of-Thought (Wei *et al.* 2022b) prompting strategies.

Although prompt tuning methods have proven to be effective, their performance is restricted on knowledge-intensive tasks that require providing provenance and up-to-date knowledge about the world to address the user prompt. To address such knowledge-intensive tasks, an alternative approach that integrates knowledge graphs (KG) with language models was recently introduced (Lin *et al.* 2019, Yang *et al.* 2019, Wang *et al.* 2019, Feng *et al.* 2020, Lv *et al.* 2020, Yasunaga *et al.* 2021, Yasunaga *et al.* 2022). This approach was primarily implemented in question-answering tasks, where the structured information contained within the KG was used to provide context for predicting the answer to the question. While such multimodal integrative approach showed promise, its downstream supervised training was tailored to a specific task, limiting its versatility and broad applicability, thereby constraining its ability to fully harness the "emergent capabilities" of LLMs (Wei *et al.* 2022). Retrieval-augmented generation (RAG) involves enhancing a parametric pretrained LLM with the ability to access a nonparametric memory containing updated knowledge about the world (e.g. Wikipedia or SPOKE) (Lewis *et al.* 2020).

In this article, we propose a robust and token-optimized framework called Knowledge Graph-based Retrieval Augmented Generation (KG-RAG) that integrates a KG with a pretrained LLM within an RAG framework, thus capturing the best of both worlds. To achieve this, we make use of the biomedical KG called SPOKE (Morris *et al.* 2023) that integrates more than 40 publicly available biomedical knowledge sources of separate domains where each source is centered around a biomedical concept, such as genes, proteins, drugs, compounds, diseases, and one or more of their known relationships. Because these concepts are recurrent entities forming defined sets (e.g. all named human genes, all FDA approved drugs, etc.) the integration of these concepts into a single graph exposes novel multi-hop factual relationships that connect the knowledge sources and provides the biological and ontological context for each concept. Unlike other RAG approaches, the proposed framework performs an optimized retrieval, specifically obtaining only the essential biomedical context from SPOKE, referred to as "prompt-aware context," which is adequate enough to address the user prompt with accurate provenance and statistical evidence. This enriched prompt is further used as input for the LLM in the RAG framework for meaningful biomedical text generation. We evaluated this approach using various pretrained LLMs including Llama-2-13b, GPT-3.5-Turbo, and GPT-4.

2 Materials and methods

2.1 KG-RAG framework

The schema of the proposed KG-RAG framework is shown in Fig. 1. The following sections explain each component of this framework.

2.2 Disease entity recognition

This is the first step in KG-RAG. The objective of this step is to extract the disease concept (an entity) from the input text prompt and then find the corresponding matching disease node in the KG (a concept in SPOKE graph). This was implemented as a two-step process: (i) entity extraction from prompt and (ii) entity matching to SPOKE. Entity extraction identifies and extracts disease entities mentioned in the input text prompt, otherwise called as "Prompt Disease extraction" (Fig. 2). To achieve this, zero-shot prompting (Kojima *et al.*

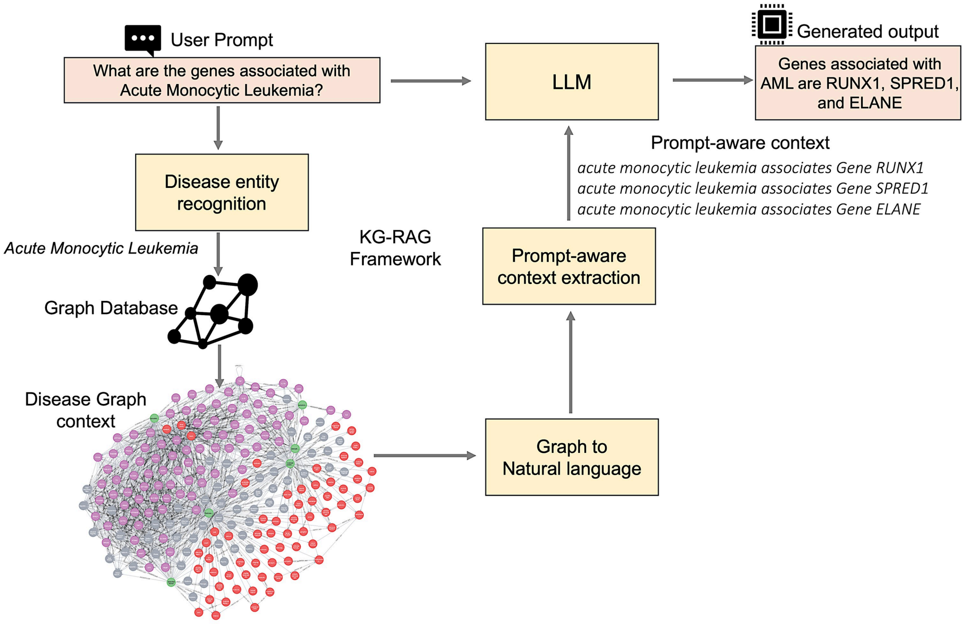

User Prompt
What are the genes associated with Acute Monocytic Leukemia?
LLM
Generated output
Genes associated with AML are RUNX1, SPRED1, and ELANE
Disease entity recognition
Acute Monocytic Leukemia
KG-RAG Framework
Prompt-aware context
acute monocytic leukemia associates Gene RUNX1
acute monocytic leukemia associates Gene SPRED1
acute monocytic leukemia associates Gene ELANE
Prompt-aware context extraction
Graph Database
Disease Graph context
Graph to Natural language

Figure 1. Schema for the Knowledge Graph based Retrieval-Augmented Generation (KG-RAG) Framework. The direction of the arrows indicates the flow of the pipeline in this framework.

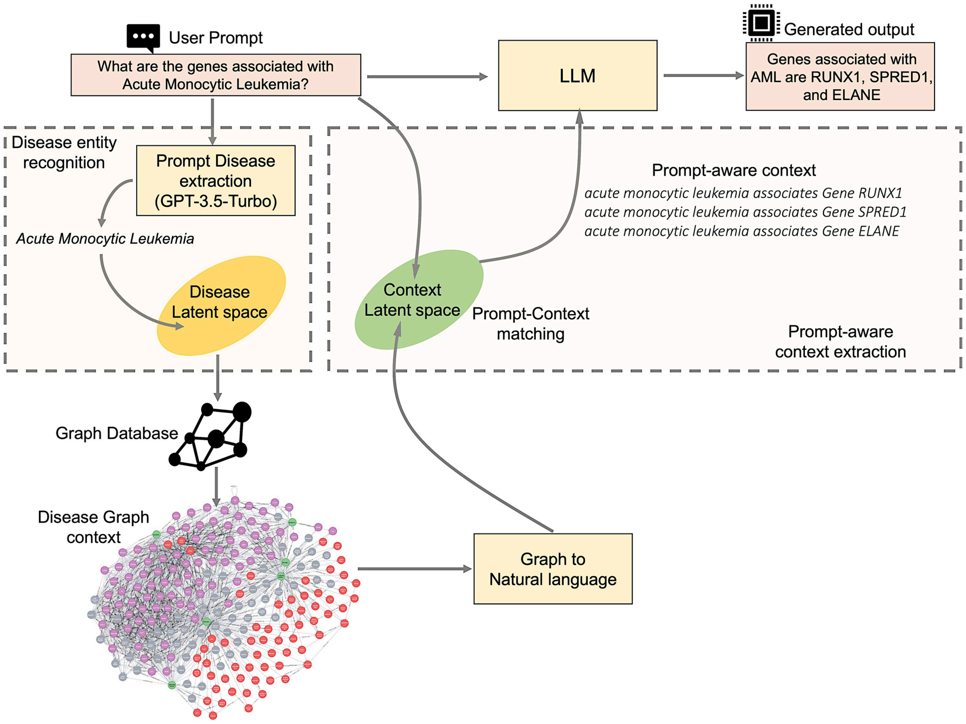

User Prompt
What are the genes associated with Acute Monocytic Leukemia?
LLM
Generated output
Genes associated with AML are RUNX1, SPRED1, and ELANE
Disease entity recognition
Prompt Disease extraction (GPT-3.5-Turbo)
Acute Monocytic Leukemia
Disease Latent space
Prompt-aware context
acute monocytic leukemia associates Gene RUNX1
acute monocytic leukemia associates Gene SPRED1
acute monocytic leukemia associates Gene ELANE
Context Latent space
Prompt-Context matching
Prompt-aware context extraction
Graph Database
Disease Graph context
Graph to Natural language

Figure 2. Detailed schema of KG-RAG. Dashed boxes show the details of "Disease entity recognition" and "Prompt-aware context extraction" from the Knowledge Graph.

2022) was used on GPT-3.5-Turbo model (Version: 0613) (Fig. 2). Specifically, a system prompt was designed to extract disease entities from the input text and return them in JSON format (Supplementary S1 Text). We compared the GPT-based entity extraction method with more traditional approaches that use models specifically trained for this task. However, we chose the zero-shot GPT option due to its superior performance in terms of precision and recall (Supplementary Fig. S1 and Table A) and its flexibility in handling various entity types (Supplementary S1 Text).

Next, entity matching was used to obtain the concept name of the disease as it is represented within the KG. For this, the embeddings for all disease concepts (i.e. nodes) in SPOKE were precomputed using the "all-MiniLM-L6-v2" sentence transformer model (i.e. disease latent space in Fig. 2) (Reimers, 2019). This procedure translates names of the disease concepts into a 384-dimensional dense vector space, making it suitable for semantic search. We chose "MiniLM" for two main reasons: firstly, when combined with entity extraction, it successfully retrieved the disease nodes from the graph with an accuracy of 99.7% (Supplementary S1 Text). Second, it produced lightweight embeddings (384 dimensions) compared to other sentence transformers like the PubmedBERT model (Gu *et al.* 2021) (768 dimensions), making it more memory efficient. Next, these newly created disease concept embeddings were stored in the "Chroma" vector database (https://medium.com/@kbdhunga/an-overview-of-chromadb-the-vector-database-206437541bdd). Disease concepts with the highest vector similarity to the extracted entity are selected for subsequent context retrieval (Fig. 2). If the zero-shot approach fails to identify a disease entity in the prompt, five disease concepts from the vector database with the most significant semantic similarity to the entire input text prompt are selected instead.

2.3 Disease context retrieval from SPOKE

The SPOKE KG connects millions of biomedical concepts through semantically meaningful relationships (Morris *et al.* 2023). The KG consists of 42 million nodes of 28 different types and 160 million edges of 91 types, implemented as a property graph and is assembled by downloading and integrating information from 41 different biomedical databases. Notably, the vast majority of SPOKE is composed of curated information determined by systematic experimental measurements, not text mining from the literature. In this study, SPOKE was used as the source of biomedical context for the diseases mentioned in the input prompt.

The process of converting the graph data to natural language involves the following steps:

a) *Context extraction*: We utilize SPOKE's REST-API service, specifically the "/api/v1/neighborhood/" endpoint, to fetch neighbors associated with a disease node (Supplementary S1 Text). To ensure high-quality context, we apply filters such as restricting treatment associations to those with clinical phase ≥ 3, using only SwissProt-curated protein nodes, and excluding text-mining-based gene-disease associations.
b) *Predicate to natural language transformation*: SPOKE uses a specific schema for predicates: UpperCase(predicateName)_<upperCase(firstLetter(subjectType)), lowerCase(firstLetter(predicateName)), and upperCase(firstLetter(objectType))>. For example, "ASSOCIATES_DaG" is a predicate that connects disease nodes with gene nodes.

The context triples (Subject, Predicate, and Object) associated with a disease node from SPOKE KG (Morris *et al.* 2023) follow a specific schema for its predicates (Supplementary S1 Text): UpperCase(predicateName) _<upperCase(firstLetter(subjectType)), lowerCase(first Letter(predicateName)), and upperCase(firstLetter (objectType))>. For example, "ASSOCIATES_DaG" is a predicate that connects Disease nodes with Gene nodes with the predicate name ASSOCIATES. This schema allowed for the direct conversion of the extracted triples into English language using the following rule (Supplementary S1 Text):

(S, P, O) → Subject lowerCase(predicateName) Object

For example:

(disease hypertension, ASSOCIATES_DaG, and gene VHL) → "disease hypertension associates Gene VHL"

This makes it compatible for inputting into the LLM (Fig. 1; Supplementary S1 Text).

c) *Provenance and evidence incorporation*: In addition to extracting the connectivity between the disease and its neighbors, we also extracted the provenance information associated with those edges which is given as the edge attribute. Additionally, we have implemented an option ("-e" flag in our KG-RAG script) to include supporting evidence like *P*-values or *z*-scores in the output (Supplementary S1 Text).

2.4 Context pruning

Next, the extracted disease context was pruned by selecting the most semantically pertinent context that could be used to answer the given prompt. First, the input prompt and all the extracted contextual associations were embedded to the same vector space (Context Latent space in Fig. 2) using a sentence transformer model (model selection was done using hyperparameter tuning). Next, for context selection, prompt-context cosine similarity should satisfy two conditions: (i) greater than 75th percentile of the similarity distribution encompassing all the context related to the chosen disease node and (ii) having a minimum similarity value of 0.5. This makes the retrieved context more fine-grained and contextually relevant (i.e. "prompt-aware context").

2.5 Large language model

The input prompt, when combined with the prompt-aware context, resulted in an enriched prompt that was used as input to the LLM for text generation. For that purpose, three pretrained chat models were used: Llama-2-13b (Touvron *et al.* 2023), GPT-3.5-Turbo (Version: 0613), and GPT-4 (http://arxiv.org/abs/2005.14165). A Llama model with 13 billion parameters and with a token size of 4096 was downloaded and deployed in the Amazon Elastic Compute Cloud (EC2) GPU P3 instance. GPT models were accessed using the OpenAI API. Since GPT models featured a higher parameter count in comparison to Llama, this gave us the opportunity to compare the performance of KG-RAG as a function of the size of the LLM in terms of its parameter count. In this study, the "temperature" parameter (http://arxiv.org/abs/2005.14165), governing the level of randomness in the LLM output, was set to 0 for all LLMs.

2.6 Hyperparameter analysis and validation

The performance of KG-RAG was evaluated across two sets of hyperparameters such as "Context volume" and "Context embedding model." Context volume defines the upper limit on the number of graph connections permitted to flow from the KG to the LLM. This hyperparameter introduced a balance between context enhancement and input token space utilization of the LLM (Supplementary S1 Text). Next hyperparameter called "context embedding model" determined which model exhibited greater proficiency in retrieving the accurate biomedical context from the KG to respond to the input prompt. Since "MiniLM" (Reimers, 2019) was used in the disease entity recognition stage, we considered that as a candidate for "context embedding model." Acknowledging that biomedical contexts often utilize vocabulary that could differ from general domain scenarios, we next considered "PubMedBert" (Gu *et al.* 2021, Deka *et al.* 2022) as another candidate for this hyperparameter since it was pretrained on biomedical text. Additionally, two sets of validation data (total 165 questions) were created using prompts mentioning a single disease (75 questions), while the other set involved prompts mentioning two diseases (90 questions; Supplementary S1 Text). The second set of "two disease prompts" were created to see if KG-RAG had the ability to retrieve contexts related to multiple diseases. These prompts were executed using the GPT-4 model in the KG-RAG framework, utilizing specific system prompts to return the results in JSON format (Supplementary S1 Text). Jaccard similarity was computed by parsing these JSON responses and comparing them with the ground truth. Through these analyses, an empirical selection of hyperparameters for the downstream tasks was made.

2.7 Test dataset

Three types of test datasets were used to quantitatively analyze the performance of the proposed framework (Supplementary Table D): (i) True/False dataset; (ii) Multiple Choice Questions (MCQ) dataset; and (iii) RAG comparison dataset. Datasets i and ii were consolidated from external knowledge bases and were further thoroughly reviewed by domain experts to remove any false positives.

A True/False dataset was created from three external data sources, such as DisGeNET (Piñero *et al.* 2016), MONDO (Vasilevsky *et al.* 2022), and SemMedDB (Kilicoglu *et al.* 2012). DisGeNET consolidates data about genes and genetic variants linked to human diseases from curated repositories, the GWAS catalog, animal models, and the scientific literature (Piñero *et al.* 2016) (Supplementary S1 Text). MONDO provides information about the ontological classification of disease entities in the open biomedical ontologies (OBO) format (Vasilevsky *et al.* 2022) (Supplementary S1 Text). SemMedDB contains semantic predications extracted from PubMed citations (Kilicoglu *et al.* 2012), and we used this resource to formulate True/False questions about drugs and diseases (Supplementary S1 Text). MCQ, comprising five choices with a single correct answer for each question, were created using data from the Monarch Initiative (Mungall *et al.* 2017) and ROBOKOP (Reasoning Over Biomedical Objects linked in Knowledge-Oriented Pathways; Bizon *et al.* 2019) (Supplementary S1 Text). To assess the graph context retrieval capabilities of various RAG frameworks, we extracted disease–gene associations from the SPOKE graph and designed questions based on these associations. These questions were then used on KG-RAG, Cypher-RAG, and

Full-Text Indexing frameworks to compare how well they retrieve the associated context from SPOKE graph.

Thus, 311 True/False, 306 MCQ, and 100 RAG comparison biomedical question datasets were created for a systematic quantitative analysis of the proposed framework. To assess the performance of LLMs on True/False and MCQ datasets, 150 questions were randomly sampled with replacement 1000 times (using bootstrapping). The accuracy metric was then calculated for each sampling iteration, resulting in a performance distribution.

2.8 Cypher-RAG

Cypher-RAG is a technique utilized for retrieving context associated with a node in a Neo4j graph database (https://medium.com/neo4j/using-a-knowledge-graph-to-implement-a-devops-rag-application-b6ba24831b16). This context can then be leveraged to generate information about the node in natural language using an LLM. The method involves explicitly embedding the schema of the graph into the input prompt, directing the LLM to generate a structured Cypher query based on this schema. The resulting Cypher query is used to make a call to the Neo4j database, and the returned information is utilized as context to respond to the user's prompt. This methodology is integrated into the LangChain python library as GraphCypherQAChain class (https://medium.com/neo4j/using-a-knowledge-graph-to-implement-a-devops-rag-application-b6ba24831b16). An advantage of this approach is that it allows for the creation of Cypher queries directly from natural language, eliminating the need for users to have the knowledge of Cypher query syntax. However, our analysis revealed certain limitations of this approach. We found that the explicit embedding of the graph schema restricts the input token space and increases token usage for this method. As the complexity of the graph schema increases, users may need to utilize LLMs with longer context window sizes for optimal performance. Additionally, we demonstrated that this method can be sensitive to how the prompt is formulated. Even slight perturbations to the prompt can lead to incorrect Cypher queries and subsequently impact downstream generative processes.

2.9 Full-text index

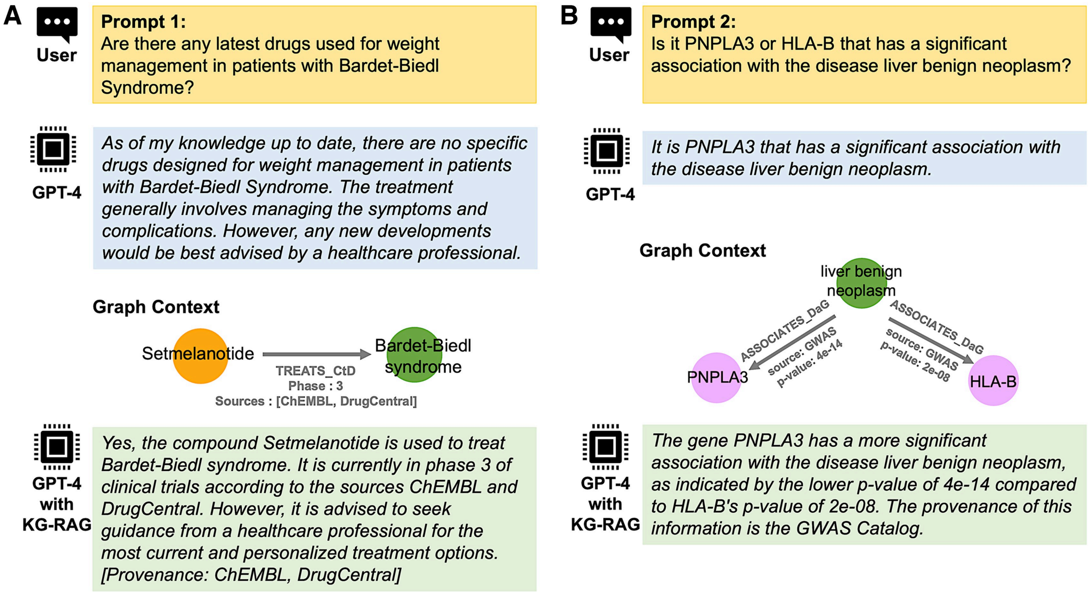
Neo4j offers full-text indexing that is powered by the Apache Lucene indexing and search library. In the context of the SPOKE graph, all nodes were full-text indexed on their "name" and "identifier" properties. This indexing method stores individual words within these string properties and allows semantic interpretation of string data beyond simple exact or substring matching. Hence, this enables content-based matching. When processing user queries, the full-text index procedure in Neo4j was used to compare the query against the indexed disease names. This comparison yields a proximity score, indicating the semantic closeness between the query and stored values. The disease node with the highest score was selected as the mapped node for subsequent context extraction. Compared to traditional exact or pattern-based matching, this method provides a more nuanced and flexible approach to node matching, potentially improving the information retrieval process.

3 Results

We developed KG-RAG, a framework that integrates LLMs with the SPOKE knowledge graph. This integration enables accurate biomedical context retrieval and reliable text generation in an optimized and cost-effective manner. This framework involves multiple steps, namely: i) entity recognition from user prompt, ii) extraction of biomedical concepts from SPOKE, iii) concept embedding, iv) prompt-aware context generation, v) conversion into natural language, vi) prompt assembly, and vii) answer generation. The performance of this approach was extensively tested using different scenarios.

3.1 Prompting KG-RAG framework

Figure 3 shows two biomedical prompts (yellow box) given as input to the GPT-4 model using two approaches:



Prompt 1:
Are there any latest drugs used for weight management in patients with Bardet-Biedl Syndrome?

As of my knowledge up to date, there are no specific drugs designed for weight management in patients with Bardet-Biedl Syndrome. The treatment generally involves managing the symptoms and complications. However, any new developments would be best advised by a healthcare professional.

Yes, the compound Setmelanotide is used to treat Bardet-Biedl syndrome. It is currently in phase 3 of clinical trials according to the sources ChEMBL and DrugCentral. However, it is advised to seek guidance from a healthcare professional for the most current and personalized treatment options. [Provenance: ChEMBL, DrugCentral]

Prompt 2:
Is it PNPLA3 or HLA-B that has a significant association with the disease liver benign neoplasm?

It is PNPLA3 that has a significant association with the disease liver benign neoplasm.

The gene PNPLA3 has a more significant association with the disease liver benign neoplasm, as indicated by the lower p-value of 4e-14 compared to HLA-B's p-value of 2e-08. The provenance of this information is the GWAS Catalog.

Figure 3. Prompting examples. Example prompts (top boxes) and prompt-based (middle boxes)/KG-RAG-based (bottom boxes) responses using GPT-4. (A) A biomedical prompt to search the drugs used for weight management in patients with "Bardet-Biedl syndrome" and (B) Another prompt to compare the statistical associations of PNPLA3 and HLA-B genes with the disease liver benign neoplasm. Both (A) and (B) show the associated "prompt aware graph" context.

(i) prompt based, i.e. without KG-RAG (blue box) and (ii) with KG-RAG (green box). We observed that only KG-RAG was able to provide an accurate answer for both prompts, accompanied by supporting evidence and provenance information. (For more prompting examples refer Supplementary Table C. We have also provided these prompts in a Jupyter notebook which can be accessed at https://github.com/BaranziniLab/KG_RAG/blob/main/notebooks/kg_rag_based_gpt_prompts.ipynb.)

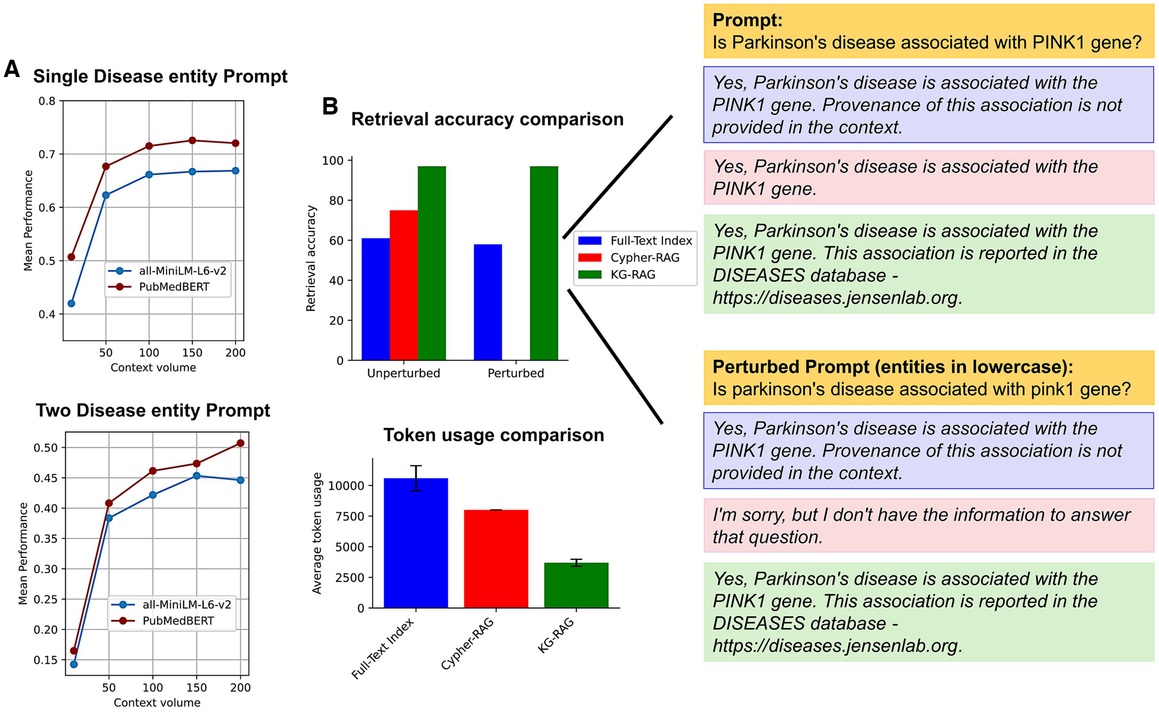
3.2 Hyperparameter analysis

KG-RAG has two hyperparameters ("context volume" and "context embedding model"), which enable it to conduct optimized context retrieval from a KG. Context volume defines the upper limit on the number of graph connections permitted to flow from the KG to the LLM (see Materials and methods, Supplementary S1 Text). Context embedding model extracts graph context that shows semantic similarity to the user prompt, facilitating the refinement of extracted context to those that are contextually relevant (see Section 2, Supplementary S1 Text). To optimize these hyperparameters, we used two context embedding models (MiniLM and PubMedBert) with increasing sizes of context volume (Fig. 4A). For prompts with single disease entity, the PubMedBert based model exhibited a mean performance (Jaccard similarity) approximately 10% higher than that of the MiniLM model across all context volume settings (mean performance of PubMedBert = 0.67, mean performance of MiniLM = 0.61). For prompts with two disease entities, PubMedBert achieved a performance that was 8.1% higher than the MiniLM across all context volume settings (mean performance of PubMedBert model = 0.4, performance of MiniLM = 0.37).

Figure 4A shows that the performance curve reaches a plateau for prompts with single disease entity and follows a similar trend for prompts with two disease entities (for both models). Based on these findings, we selected PubMedBert-based model as the context embedding model and set the context volume to a value between 100 and 200 (for most downstream tasks, we opted for a context volume of 150, and for True/False questions, we chose a context volume of 100).

3.3 RAG comparative analysis

Figure 4B shows the comparative analysis between the proposed KG-RAG and the other two methods, such as Cypher-RAG and full-text index approach for context retrieval from a KG (see Materials and methods). We compared these three frameworks based on their retrieval accuracy, retrieval robustness, and token usage. For a test dataset with 100 biomedical questions (Supplementary S1 Text), full-text index, Cypher-RAG, and KG-RAG showed 61%, 75%. and 97% retrieval accuracy, respectively (Fig. 4B, top). To test the robustness in context retrieval, we introduced a slight perturbation to the test dataset by converting the entity names to lowercase (Fig. 4B, insight). We observed a significant decrease in the retrieval accuracy of Cypher-RAG to 0% (indicating failure to retrieve any context from the graph). This mainly occurs because Cypher-RAG utilizes precise matching of the entity keywords provided in the user prompt to formulate the Cypher query for extracting graph context.



Prompt:
Is Parkinson's disease associated with PINK1 gene?

Yes, Parkinson's disease is associated with the PINK1 gene. Provenance of this association is not provided in the context.

Yes, Parkinson's disease is associated with the PINK1 gene.

Yes, Parkinson's disease is associated with the PINK1 gene. This association is reported in the DISEASES database - https://diseases.jensenlab.org.

Perturbed Prompt (entities in lowercase):
Is parkinson's disease associated with pink1 gene?

Yes, Parkinson's disease is associated with the PINK1 gene. Provenance of this association is not provided in the context.

I'm sorry, but I don't have the information to answer that question.

Yes, Parkinson's disease is associated with the PINK1 gene. This association is reported in the DISEASES database - https://diseases.jensenlab.org.

Figure 4. Hyperparameter analysis and RAG comparison. (A) Hyperparameter analysis performance curves using prompts with single (top) and two (below) disease entities mentioned in it. The x-axis denotes the "Context volume" (number of associations from KG) and the y-axis denotes the mean performance (Jaccard similarity) across the prompts. The red curve denotes "PubMedBert" and the blue curve denotes "MiniLM" transformer models. (B) The comparative analysis between KG-RAG, Cypher-RAG, and Full-Text Index using Apache Lucene in terms of retrieval accuracy (top) and token usage (bottom). Insight shows an example where Cypher-RAG fails to retrieve context from the KG when the input prompt is slightly perturbed, but KG-RAG and Full-Text Index remain robust in context retrieval. However, only KG-RAG could extract provenance information linked to the assertion from the context, whereas Full-Text Index could not. It is evident that KG-RAG has significantly lesser token usage when compared to Cypher-RAG and Full-Text Index method (bottom). Error bar in the token utilization bar plot (bottom) represents standard error of the mean (SEM).

Compared to Cypher-RAG, full-text index, and KG-RAG showed robustness against perturbation. Full-text index showed a slight decrease in the accuracy after query perturbation (dropped to 58% from initial 61%), but KG-RAG was able to maintain its retrieval accuracy at 97% which indicates its higher robustness against input perturbation (Fig. 4B, top). This is because KG-RAG employs a semantic embedding approach to extract graph context, which enhances its ability to effectively handle various representations of entities within user prompts. Next, we analyzed the total token usage of each framework for generating the response for the same test dataset (Fig. 4B, bottom). We found that full-text index had an average token usage of 10 590 tokens, while Cypher-RAG averaged 8006 tokens. However, KG-RAG had an average token usage of only 3693 tokens (Fig. 4B, bottom). This represents a 53.9% reduction in the token usage compared to Cypher-RAG and 65.1% reduction compared to full-text index. This highlights the significant cost-effective retrieval ability of KG-RAG.

3.4 Performance on True/False and MCQ datasets

Figure 5 shows bootstrap distributions of performance (accuracy) of the three LLMs using prompt-based and KG-RAG framework on True/False (Fig. 3A) and MCQ (Fig. 5B) datasets. Table 1 summarizes the performance of the three LLMs across these datasets. We observed a consistent performance enhancement for the LLM models under KG-RAG framework on both True/False and MCQ datasets (Table 1). KG-RAG significantly elevated the performance of Llama-2 by approximately 71% from its initial level (0.31 ± 0.03 to 0.53 ± 0.03) on the more challenging MCQ dataset (Table 1). Intriguingly, we also observed a small but significant drop in the performance of GPT-4 model (0.74 ± 0.03) compared to GPT-3.5-Turbo model (0.79 ± 0.02) on MCQ dataset using KG-RAG framework (T-test, P-value < 0.0001, t-statistic $= -47.7$, $N = 1000$) but not in the prompt-based approach. We have also conducted an additional comparative analysis between Cypher-RAG and KG-RAG using the same benchmark datasets, where KG-RAG showed significantly higher performance on both datasets (Supplementary S1 Text, Supplementary Fig. S2 and Table B).

In this work, we introduce a simple but highly effective framework that combines a biomedical knowledge graph with LLM chat models in a token optimized fashion. This integration resulted in a domain-specific generative system whose responses were firmly grounded in well-established biomedical knowledge. We compared the proposed

Table 1. LLM performance (accuracy: mean ± std) on True/False and MCQ dataset.

Model	True/False dataset		MCQ dataset	
	Prompt based	KG-RAG	Prompt based	KG-RAG
Llama-2-13b	0.89 ± 0.02	0.94 ± 0.01	0.31 ± 0.03	0.53 ± 0.03
GPT-3.5-Turbo	0.87 ± 0.02	0.95 ± 0.01	0.63 ± 0.03	0.79 ± 0.02
GPT-4	0.9 ± 0.02	0.95 ± 0.01	0.68 ± 0.03	0.74 ± 0.03

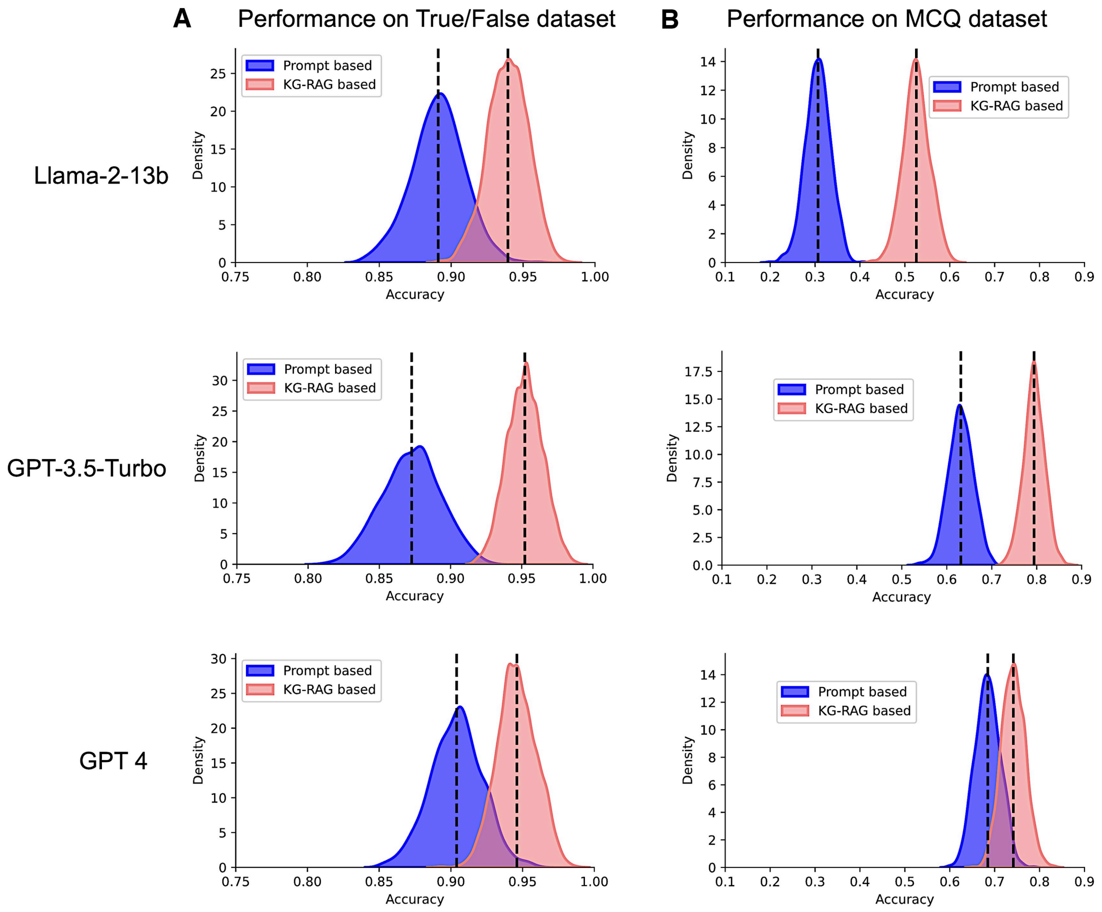


Figure 5. LLM performance on True/False and MCQ datasets. Performance (Accuracy) distributions of LLMs on (A) True/False and (B) MCQ datasets. Figure panel shows the distributions corresponding to both prompt-based and KG-RAG-based approaches. Vertical-dashed line indicates the mean value of the distribution. The higher the value, the better the performance.

framework with other approaches that utilize Cypher query and full-text indexing and showed that KG-RAG was more robust to prompt perturbation and more efficient in token utilization. In addition, KG-RAG consistently demonstrated superior performance compared to the prompt-based baseline LLM model on all human-curated benchmark datasets. We hypothesize that this performance improvement arises from the fusion of the explicit knowledge from the KG and the implicit knowledge from the LLM. This shows the value of providing domain-specific ground truth at a fine-grained resolution as context at the prompt level.

A heterogeneous knowledge graph with diverse concepts (the biomedical concepts in this case) interconnected at a massive scale has the potential to generate new knowledge as an "emergent property" (Baranzini *et al.* 2022, Morris *et al.* 2023). In fact, as LLMs scale up in various dimensions like model parameters, training data, and training compute, they have been thought to exhibit reasoning or "emerging capabilities" (Wei *et al.* 2022) although this observation could also be explained by "in-context learning" or other aspects of the examples (http://arxiv.org/abs/2005.14165; Min *et al.* 2022, Lu *et al.* 2023). In any case, KG-RAG capitalized this capability and generated biomedical text with rich annotations such as provenance and statistical evidence (if available) thereby resulting in more reliable and knowledge-grounded responses. Additionally, the optimized and fine-grained context retrieval capability of KG-RAG ensured a budget friendly RAG system to apply on proprietary LLMs. This way, KG-RAG democratizes SPOKE's vast biomedical knowledge by enabling researchers, clinicians, and other professionals to leverage its comprehensive information through natural language interactions. This approach facilitates access to complex biomedical data without requiring specialized graph query language skills.

Previous studies have utilized KG in conjunction with LLM for knowledge intensive tasks such as question-answering (Yasunaga *et al.* 2022), multi-hop relational reasoning (Feng *et al.* 2020), commonsense reasoning (Lin *et al.* 2019, Lv *et al.* 2020), and model pretraining (Moiseev *et al.* 2022, Yasunaga *et al.* 2022). Furthermore, enhancing prompts by incorporating structured knowledge has been described and studied (Lewis *et al.* 2020, Pan *et al.* 2023). Naturally, these approaches have bolstered the positive reinforcement between KG and LLM. Nevertheless, it is worth noting that these approaches are often task specific and, in some cases, the knowledge infusion could grow exponentially with the inclusion of higher-order relations (https://doi.org/10.18653/v1/d18-1454) (Lin *et al.* 2019). Such approaches could compromise the limited token space of the LLM. Alternative methods used knowledge infusion through the direct use of query languages such as SPARQL (https://doi.org/10.5445/IR/1000151291). However, this could render the system constraint to the schema of the underlying KG, potentially affecting the flexibility and adaptability of prompts. Moreover, as the KG expands and its schema grows, it could potentially occupy a significant portion of the LLM input token space. This explains why we noticed a greater token usage with the Cypher-RAG method (average usage of 8006 tokens), as it incorporates the entire graph schema into the input prompt for converting natural language into structured Cypher queries. In contrast, KG-RAG requires minimal graph schema, thus eliminating the need to include it in the prompt and resulting in substantial token savings, with a reduction of over 50% in token utilization compared to Cypher-RAG. This finding suggests that Cypher-RAG, when dealing with a graph as large as SPOKE (which contains over 40 million nodes), requires LLMs that enable a larger context window. This limitation was also observed with traditional information retrieval frameworks such as full-text indexing using Apache Lucene. In contrast, KG-RAG is capable of managing this with LLMs that require a relatively smaller window size.

To conduct robust benchmarking, we curated datasets that underwent review by domain experts. Given the swift progress in LLM research, we believe that such rigorously vetted datasets could serve as valuable resources not only for evaluating KG-RAG but also for assessing other ongoing LLM endeavors in biomedicine. In our benchmarking analysis, we found an enhancement in LLM performance as a function of the model size in terms of the number of parameters. Intriguingly, with the KG-RAG framework the performance of GPT-4 on the MCQ dataset, despite its model size, dropped significantly compared to that of the GPT-3.5-Turbo on the MCQ dataset. In fact, the performance of GPT-3.5-Turbo under KG-RAG framework was on par with that of the GPT-4 model on True/False datasets. These results suggest that at present, GPT-3.5 may be a better context listener than GPT-4. In fact, a recent study compared the March 2023 version of GPT-4 with the June 2023 version, shedding light on the drift in the LLM performance over time (Chen *et al.* 2023). The study revealed that, as time progressed, GPT-4 exhibited a reduced tendency to adhere to user instructions. In contrast, there was no consistent alteration observed in the instruction-following behavior of GPT-3.5 over time. Studies have also shown that GPT-4's larger model size and more diverse training data enable it to excel in a multitude of complex reasoning tasks (http://arxiv.org/abs/2305.03195). GPT-3.5-Turbo has shown to excel in few-shot learning by context adherence (http://arxiv.org/abs/2005.14165). We believe that the context adherence nature of GPT-3.5, in contrast to the general application of GPT-4, might make it more suitable for frameworks like KG-RAG. Therefore, these factors could have contributed to GPT-3.5's superior performance over GPT-4 in the KG-RAG framework.

When the proprietary GPT models were compared to the open-source Llama-2-13b model, they showed a narrow margin in performance on the biomedical True/False dataset. However, on the more challenging MCQ dataset, Llama-2 initially demonstrated lower performance compared to GPT models. Interestingly, the KG-RAG framework provided a substantial performance boost to Llama-2, improving its performance by ~71% from the baseline. Despite this boost that narrowed the performance gap, the performance of Llama-2 remained lower than that of the GPT models. This suggests that the KG-RAG framework has the potential to capitalize the intrinsic context comprehension capabilities of open-source pretrained models like Llama-2, making them more competitive with proprietary models like GPT. The findings underscore the importance of context enrichment techniques for improving the performance of language models on complex tasks in specialized domains.

While the proposed framework has successfully addressed numerous challenges, we recognize there are opportunities for improvement. Currently, this approach is limited to handling biomedical questions centered around diseases, as our focus has been on embedding disease concepts from the SPOKE KG for recognizing disease entities in the prompt. However, we have implemented a proof of concept that extends KG-RAG to non-disease-centric questions by embedding more than 87 000 nodes from SPOKE, including genes, proteins, enzymes, symptoms, compounds, reactions, anatomical structures, and side effects (Supplementary S1 Text). Future work may expand this scope by including all biomedical concepts (nodes) in SPOKE and other KG. Since SPOKE contains more than 40 million biomedical nodes, this expansion will enable the KG-RAG framework to address a broader range of biomedical questions and thereby enhance its versatility. Currently, we have implemented the KG-RAG framework exclusively on the SPOKE biomedical knowledge graph, which is maintained as a Neo4j property graph database. While we utilize embeddings to enhance node and context retrieval, all operations are performed within this Neo4j graph structure. As future work, we plan to extend KG-RAG to other biomedical KGs and domains, demonstrating its versatility and broader applicability. Finally, the quality of the retrieved context relies on the information stored in the underlying graph. In our case, SPOKE utilizes meticulously curated knowledge bases to construct its nodes and edges; however, we do not assert that it is entirely error free or ready for clinical use. Thus, while SPOKE's reliability has been demonstrated through its successful application in various biomedical contexts (Supplementary S1 Text; Himmelstein and Baranzini, 2015, Himmelstein *et al.* 2017, Nelson *et al.* 2019, Nelson *et al.* 2021a,b, Baranzini *et al.* 2022, Morris *et al.* 2023, Soman *et al.* 2023a,b, Tang *et al.* 2024), it is important to note that this work primarily focused on the development of a framework, rather than conducting a rigorous and formal evaluation of the KG itself.

In summary, the KG-RAG framework retrieves semantically meaningful context from a knowledge graph using minimal tokens, then combines this explicit knowledge with the parameterized implicit knowledge of an LLM. This knowledge integration results in the generation of domain specific, reliable and up-to-date meaningful biomedical responses with rich annotations.

Acknowledgements

S.E.B. holds the Heidrich Family and Friends Endowed Chair of Neurology at UCSF. S.E.B. holds the Distinguished Professorship in Neurology I at UCSF. We also acknowledge the Versa team at UCSF for providing the infrastructure to access OpenAI API.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

S.E.B. is a co-founder of Mate Bioservices. CAN is CEO of Mate Bioservices.

Funding

The development of SPOKE and its applications are being funded by grants from the National Science Foundation [NSF_2033569], NIH/NCATS [NIH_NOA_1OT2TR003450].

References

Baranzini SE, Börner K, Morris J *et al.* A biomedical open knowledge network harnesses the power of AI to understand deep human biology. *AI Mag* 2022;**43**:46–58.

Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. arXiv, arXiv:1903.10676, 2019, preprint: not peer reviewed.

Bizon C, Cox S, Balhoff J *et al.* ROBOKOP KG and KGB: integrated knowledge graphs from federated sources. *J Chem Inf Model* 2019; **59**:4968–73.

Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? arXiv, arXiv:2307.09009, 2023, preprint: not peer reviewed.

Deka P, Jurek-Loughrey A, P D. Improved methods to aid unsupervised evidence-based fact checking for online health news. *JDI* 2022; **3**:474–504.

Feng Y, Chen X, Lin BY *et al.* Scalable multi-hop relational reasoning for knowledge-aware question answering. arXiv, arXiv:2005.00646, 2020, preprint: not peer reviewed.

Gu Y, Tinn R, Cheng H *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM TransComputHealthcare* 2021;**3**:1–23.

Himmelstein DS, Baranzini SE. Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. *PLoS Comput Biol* 2015;**11**:e1004259.

Himmelstein DS, Lizee A, Hessler C *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* 2017; **6**:1–35.

Huang K, Altosaar J, Ranganath R. Clinicalbert: modeling clinical notes and predicting hospital readmission. arXiv, arXiv:1904.05342, 2019, preprint: not peer reviewed.

Ji Z, Lee N, Frieske R *et al.* Survey of hallucination in natural language generation. *ACM Comput Surv* 2023;**55**:1–38.

Kenton J-WC, Toutanova LK. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of naacL-HLT*. 2019, 2.

Kilicoglu H, Shin D, Fiszman M *et al.* SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics* 2012; **28**:3158–60.

Kojima T, Gu SS, Reid M *et al.* Large language models are zero-shot reasoners. *Adv Neural Inform Process Syst* 2022;**35**:22199–213.

Lee J, Yoon W, Kim S *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40.

Lewis P, Perez E, Piktus A *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv Neural Inform Process Syst* 2020;**33**:9459–74.

Lin BY, Chen X, Chen J *et al.* Kagnet: knowledge-aware graph networks for commonsense reasoning. arXiv, arXiv:1909.02151, 2019, preprint: not peer reviewed.

Liu Y. Roberta: a robustly optimized BERT pretraining approach. arXiv, arXiv:1907.11692, 2019, preprint: not peer reviewed.

Lu S, Bigoulaeva I, Sachdeva R *et al.* Are Emergent abilities in large language models just in-context learning? arXiv, arXiv:2309.01809, 2023, preprint: not peer reviewed.

Luo R, Sun L, Xia Y *et al.* BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform* 2022; **23**:bbac409.

Luo Y, Zhang J, Fan S *et al.* Biomedgpt: open multimodal generative pre-trained transformer for biomedicine. arXiv, arXiv:2308.09442, 2023, preprint: not peer reviewed.

Lv S, Guo D, Xu J *et al.* Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. *AAAI* 2020;**34**:8449–56.

Maynez J, Narayan S, Bohnet B *et al.* On faithfulness and factuality in abstractive summarization. arXiv, arXiv:2005.00661, 2020, preprint: not peer reviewed.

Min S, Lyu X, Holtzman A *et al.* Rethinking the role of demonstrations: What makes in-context learning work? arXiv, arXiv:2202.12837, 2022, preprint: not peer reviewed.

Moiseev F, Dong Z, Alfonseca E *et al.* SKILL: structured knowledge infusion for large language models. arXiv, arXiv:2205.08184, 2022, preprint: not peer reviewed.

Morris JH, Soman K, Akbas RE *et al.* The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics* 2023;**39**:1–7.

Mungall CJ, McMurry JA, Köhler S *et al.* The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res* 2017;**45**:D712–22.

Nelson CA, Acuna AU, Paul AM *et al.* Knowledge network embedding of transcriptomic data from spaceflown mice uncovers signs and symptoms associated with terrestrial diseases. *Life (Basel)* 2021a;**11**:42.

Nelson CA, Bove R, Butte AJ *et al.* Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. *JAm Med Inform Assoc* 2021b; **29**:424–34.

Nelson CA, Butte AJ, Baranzini SE. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. *Nat Commun* 2019; **10**:3045.

Pan JZ, Razniewski S, Kalo JC *et al.* Large language models and knowledge graphs: opportunities and challenges. arXiv, preprint arXiv:2308.06374, 2023, preprint: not peer reviewed.

Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv, arXiv:1906.05474, 2019, preprint: not peer reviewed.

Piñero J, Bravo À, Queralt-Rosinach N *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2016;**45**:D833–9.

Raunak V, Menezes A, Junczys-Dowmunt M. The curious case of hallucinations in neural machine translation. arXiv, arXiv:2104.06683, 2021, preprint: not peer reviewed.

Reimers N. Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv, arXiv:1908.10084, 2019, preprint: not peer reviewed.

Singhal K *et al.* Large language models encode clinical knowledge. arXiv, arXiv:2212.13138, 2022, preprint: not peer reviewed.

Soman K, Nelson CA, Cerono G *et al.* Time-aware embeddings of clinical data using a knowledge graph. *Pac Symp Biocomput* 2023a; **28**:97–108.

Soman K, Nelson CA, Cerono G *et al.* Early detection of Parkinson's disease through enriching the electronic health record using a biomedical knowledge graph. *Front Med (Lausanne)* 2023b; **10**:1081087.

Tang AS, Rankin KP, Cerono G *et al.* Leveraging electronic health records and knowledge networks for Alzheimer's disease prediction and sex-specific biological insights. *Nat Aging*2024;**4**:379–95.

Touvron H, Martin L, Stone K *et al.* Llama 2: open foundation and fine-tuned chat models. arXiv, arXiv:2307.09288, 2023, preprint: not peer reviewed.

Vasilevsky NA, Matentzoglu NA, Toro S *et al.* Mondo: unifying diseases for the world, by the world. *MedRxiv* 2022; 2022.2004.2013.22273750, preprint: not peer reviewed.

Wang X, Kapanipathi P, Musa R *et al.* Improving natural language inference using external knowledge in the science questions domain. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019, Vol. **33**, 7208–15.

Wei J, Tay Y, Bommasani R *et al.* Emergent abilities of large language models. arXiv, arXiv:2206.07682, 2022a, preprint: not peer reviewed.

Wei J, Wang X, Schuurmans D *et al.* Chain-of-thought prompting elicits reasoning in large language models. *AdvNeural Inform Process Syst* 2022b;**35**:24824–37.

Yang A, Wang Q, Liu J *et al.* Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, 2346–57.

Yasunaga M, Bosselut A, Ren H *et al.* Deep bidirectional language-knowledge graph pretraining. *AdvNeural Inform Process Syst* 2022;**35**:37309–23.

Yasunaga M, Ren H, Antoine B *et al.* QA-GNN: reasoning with language models and knowledge graphs for question answering. arXiv, arXiv:2104.06378, 2021, preprint: not peer reviewed.

Zhao WX, Zhou K, Li J *et al.* A survey of large language models. arXiv, arXiv:2303.18223, 2023, preprint: not peer reviewed.

Bioinformatics, 2024, 40, 1–10
https://doi.org/10.1093/bioinformatics/btae560
Original Paper