Advanced Medical Knowledge Graphs through LLMs Integration

Koushik Reddy Parukola
Luddy School of Informatics, Computing and Engineering
Indiana University

May 4, 2024

Abstract

In addressing the significant challenges posed by hallucinatory outputs from Large Language Models (LLMs) within the medical and healthcare sectors, this paper examines the efficacy of a tailored domain-specific AI application designed to mitigate such inaccuracies. The implementation of this application encompasses numerous obstacles, including the extraction of Named Entity Recognition (NER), Entity-Relation, and the creation of Knowledge Graphs. This study delineates the essential methodologies required to develop and refine such an application, proposing a structured approach to enhance accuracy and reliability in medical data processing through AI.

1 Introduction

The emergence of large language models (LLMs) has been transformative in AI, especially in processing and generating text with human-like text. However, their application in sensitive sectors like healthcare raises concerns about reliability and the risk of hallucinations—where LLMs produce incorrect or misleading information. Such inaccuracies are critical in healthcare, potentially leading to misdiagnoses or inappropriate treatments. This underscores the urgent need for medical-specific LLMs that are finely tuned to medical data, reducing the likelihood of hallucinations and ensuring the accuracy essential for patient safety.

This paper explores the development of a domain-specific artificial intelligence application designed to address the problem of hallucinations in Large Language Models (LLMs) used within the medical and healthcare sectors. It provides an in-depth analysis of the methodologies involved in extracting named entities, establishing entity relations, and creating comprehensive knowledge graphs. Furthermore, the study evaluates the effectiveness of these approaches in improving the accuracy and reliability of LLM outputs in clinical settings.

2 Related Work

Large language models such as GPT, which have been intricately fine-tuned, are currently being tested in various healthcare environments to potentially enhance the efficiency and effectiveness of clinical practices, medical education, and research. These models have already shown promising outcomes in several biomedical applications. However, their deployment in the healthcare sector generates both enthusiasm and concern. Clinicians must make well-informed decisions regarding the use of these advanced AI tools to maximize benefits for both patients and healthcare practitioners[1].

Another paper formalizes the issue of hallucinations in large language models (LLMs), demonstrating through learning theory that LLMs cannot learn all computable functions, making hallucinations inevitable. It further outlines the types of tasks prone to hallucinations and discusses the efficacy of existing mitigation strategies, underscoring their importance for the safe deployment of LLMs in complex real-world scenarios[2].

Also, some research papers review strategies to mitigate hallucinations in large language models (LLMs), focusing on the use of knowledge graphs as a source of external information to fill knowledge gaps and enhance reasoning accuracy. It categorizes and evaluates these knowledge-graph-based augmentation techniques, discussing current trends, challenges, and future research directions in this area[3].

3 Experimental Setup, Data and Procedure

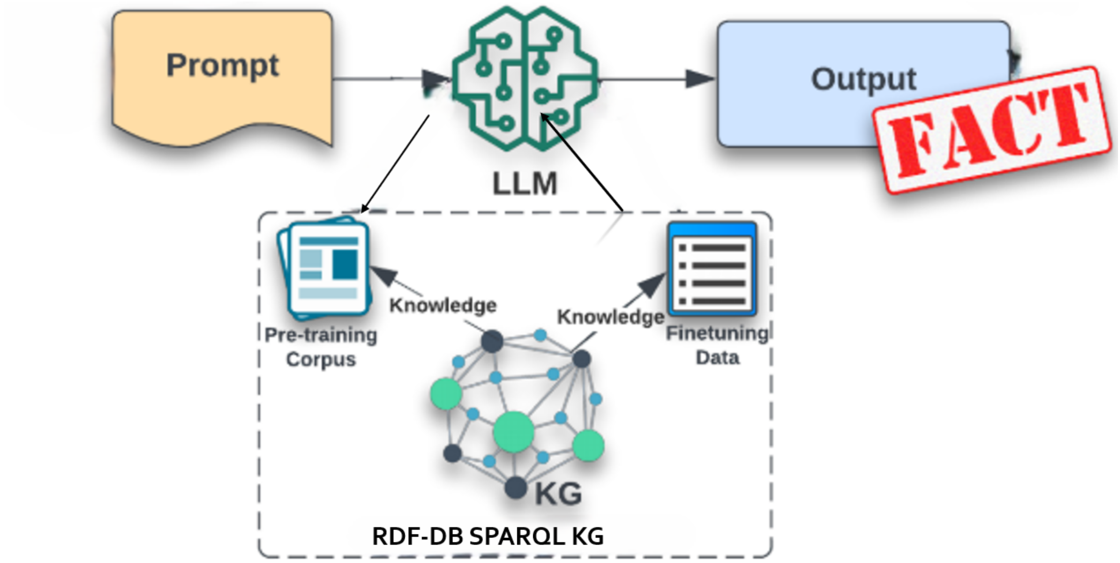

Figure 1: Workflow Diagram

Fig. 1 shows the experimental setup and the data workflow in the application. We can divide the working and architecture of this application into 3 parts.

Named Entities and Relations Extraction: To efficiently extract entities and their relations, we used advanced models such as BioBERT[4], GPT-4[5], Claude-3 Sonet, and Claude-3 Opus. These models are pivotal in organizing unstructured data, thereby enhancing the retrieval, and analysis of information.

Knowledge Graph Representation (RDF database Integration): RDF databases, when queried by SPARQL, can effectively serve as both storage and retrieval mechanisms for knowledge graphs and are used to create much more Semantically enriched applications. However, it's important to understand that simply having a SPARQL-capable RDF database does not automatically create a knowledge graph. A true knowledge graph emerges from integrating a rich schema and extensive metadata that provide deep context and understanding of the data.

LLM call to generate query and text: Upon receiving the prompt from the user, we employ a large language model (LLM) to translate the prompt or the question to create a SPARQL query, aimed at retrieving entity relationships from the knowledge graph. Subsequently, when we extract these results from the knowledge graphs, with the help of an LLM we transform the results back into general text as the result in a human-readable format.

Datasets used:

MACCROBAT: It is a comprehensive dataset containing detailed medical reports of patients with meticulously annotated information. Uncover valuable insights to advance healthcare analytics, providing a foundation for improved patient care and medical research[6].

NCBI Disease corpus: This is a medical data resource that contains 793 PubMed abstracts and lists 6892 disease mentions, which are linked to 790 unique concepts.[7]

4 Experiments Conducted and Results:

Prompted a subset of our dataset to various language models conducting a manual analysis of extracting named entities and relations among them. Talking about them one by one.

Firstly, we used a closed-source language model known as "biomedical-ner-all" from the Hugging Face Transformers library for Named Entity Recognition (NER) analysis. This model proved to be one of the most effective in accurately identifying a wide range of entities from the text, with labels including disease, symptoms, diagnosis, treatments, and more. Despite its efficacy in entity recognition, the model presents challenges in post-processing text and extracting relations, which are critical steps for the successful creation of comprehensive knowledge graphs.

The next model we explored for Named Entity Recognition (NER) was 'BERT'.

While the standard BERT model excels at identifying conventional entities such as person, place, age, and sex, its performance significantly diminishes with medical entities. However, a specialized variant known as "BioBERT," which has been trained specifically on domains like drugs/chemicals, diseases, genes/proteins, and species, shows marked improvement. BioBERT effectively discriminates entities under labels such as 'disease' and 'not a disease.' Nonetheless, constructing a knowledge graph solely with these details would be insufficient, as it overlooks the nuanced interrelations and contextual dependencies characteristic of comprehensive knowledge graphs.

The conversational capabilities of Large Language Models such as GPT, CLAUDE, and COPILOT are utilized to extract entities and relations from text. Notably, GPT-4 and CLAUDE-3 Sonnet have demonstrated proficiency in extracting meaningful relations for knowledge graph creation. However, these models have limitations, particularly in identifying complex entities. For instance, they struggle to accurately extract entities like "occupying lesion," a term that refers to damage in brain tissue caused by illness or injury. This highlights a gap in the current capabilities of LLMs in recognizing and processing sophisticated medical entities.

Finally, for the extraction of entities and relations, we bought the CLAUDE-3 OPUS API from Anthropic. We processed all the textual medical documents through this API and saved the outcomes in text files as subject-predicate-object triplets.

The prompt that is given to GPT, CLAUDE, and COPILOT models: "Here is a text. Extract all the NERs as triplets of subject-predicate-object from this text and format them as follows:(subject, predicate, object),(subject, predicate, object), and so on."

Here are the results ie., each model's accuracy and efficiency in extracting named entities, relations.

	GPT 4 - chat	Claude-3 sonnet	Claude-3 opus	Copilot
Medical datasets	85.8%	83.1%	88.2%	73%

Table 1: Accuracy Comparison of Various Language Models for Named Entity Recognition and Relation Extraction. The accuracies were determined through manual annotation of documents and subsequent comparison of named entities and relations identified by each model. This meticulous process ensures a robust evaluation of each model's capabilities in accurately parsing and understanding textual data.

	BERT	BioBert	biomedical-ner-all
Medical datasets	7%	62%	89.1%

Table 2: Comparative Analysis of Accuracy Across Different Language Models only for Named Entity Recognition (NER). Accuracies were assessed through the manual annotation of documents, followed by a comparison of named entities identified by each model.

Secondly, regarding knowledge graphs: establishing a graph database that offers

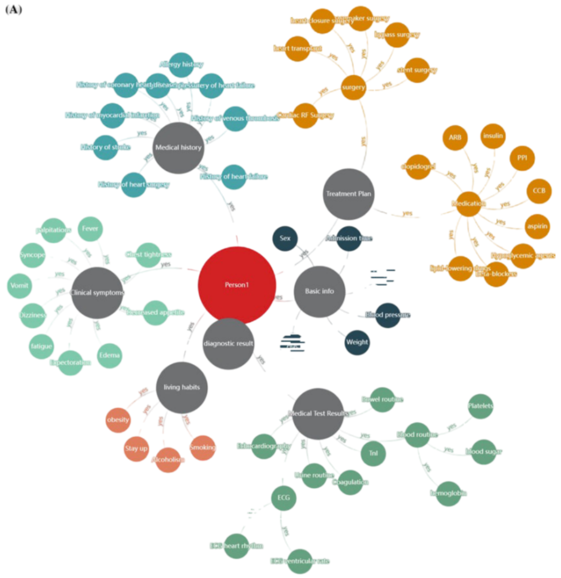

Figure 2: Sample Knowledge graph with details of a patient, doctor, patient medical condition, symptoms, etc.

meaningful insights inherently constitutes a knowledge graph (KG). Throughout this paper, we will refer to knowledge graphs as ”KGs.”

In constructing a knowledge graph (KG), we employ the triplets comprising subject, predicate, and object that are derived from language models to form the nodes and edges of the graph. While Large Language Models (LLMs) such as GPT and Claude exhibit robust Named Entity Recognition (NER) capabilities, they are not entirely suited for direct KG generation. Consider, for instance, the triplet ”The anterior tricuspid valve leaflet - was - elongated.” In this case, the verb ”was” serves as the edge linking the entities ”The anterior tricuspid valve leaflet” and ”elongated.” However, utilizing the verb ”was” alone does not aptly characterize the relationship between these entities. This example underscores the necessity for additional processing of the relations extracted by LLMs to ensure their appropriateness for inclusion in a KG. Numerous instances like this highlight the challenges in directly leveraging LLM outputs for knowledge graph construction without further refinement.

Now let’s talk about the efficiency in extracting named entities, relations, and capability of query generation. Both GPT 4 and Claude-3 were exceptionally good at generating queries on the RDF database I created. Also, the DB we created wasn’t much complex, so a much more complex database might stress LLMs to generate incorrect queries. For example here are some prompts given to GPT-4 and Claude-3 Sonnet: ’Generate a SPARQL Query to see how many doctors are there in total.’ and ’Generate a SPARQL Query to see how many patients have diabetes. The results they gave varied slightly but they were correct as per the knowledge graph.

5 Future Work

In my analysis, the recognition of Named Entity Recognition (NER) and relations from text has significantly progressed. However, fully automating this process without human intervention may not yet be efficient, particularly in pre-processing the named relation triplets. Consider, for example, a triplet from our research: (”operation”, ”noted”, ”submucosal tumor about 2.3 cm in size adherent to the calcified plate”). There is a need for modifications to refine the object to ’submucosal tumor’, and the additional details should be restructured into another relation, such as (’submucosal tumor’ - ’size’ - ’2.3 cm’). Such post-processing adjustments are essential for enhancing the quality of relation extraction, which in turn facilitates the construction of a meaningful knowledge graph. Future work will need to focus on developing more sophisticated methods for automating these adjustments to ensure the utility and accuracy of the resulting knowledge graphs. Future research should therefore focus not only on enhancing the precision of these models but also on developing robust mechanisms for human-machine collaboration to improve the strengths of both in creating more reliable and insightful medical knowledge graphs.

6 Conclusion

In conclusion, despite significant advancements in the recognition of Named Entity Recognition (NER) through models such as ’biomedical-ner-all’, and relation extraction through Large Language Models like GPT and Claude, the necessity of human in the loop remains critical due to the inconsistencies in automatic relation extraction. This paper has focused on extracting Named Entity Relations (NERs) and the relationships among them, with a particular emphasis on medical texts.

While various models have demonstrated competence in extracting NERs and relations from general texts, the complexity of medical texts poses unique challenges. These texts demand extensive datasets, rigorous training, and meticulous fine-tuning to extract the specialized terminologies of the medical field accurately. In our experiments, the ’Biomedical-ner-all’ model exhibited a commendable performance, achieving approximately 89% accuracy in identifying entities within medical texts. This was followed by Bio-BERT, which attained 62% accuracy in similar tasks within our datasets.

Furthermore, in the domain of relation extraction from texts, models like GPT and Claude have notably outperformed competitors such as Copilot, achieving nearly 88% accuracy. Despite these promising results, the complexity and critical nature of medical data necessitate a cautious approach. Automated systems must be closely monitored and frequently adjusted by human experts to ensure the reliability and relevance of the extracted information.

References

[1] Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K. et al, "*Large language models in medicine*", Nat Med . **29**, 1930–1940 (2023).

[2] Ziwei Xu, Sanjay Jain, Mohan Kankanhalli, "*Hallucination is Inevitable: An Innate Limitation of Large Language Models.*", arXiv:2401.11817 (2024).

[3] Agrawal, Garima, Kumarage, Tharindu, Alghamdi, Zeyad, Liu, Huan, "*Can Knowledge Graphs Reduce Hallucinations in LLMs?: A Survey*", arXiv.2311.07914 (2023).

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, "*BioBERT: a pre-trained biomedical language representation model for biomedical text mining*", Oxford Academic - Bioinformatics, Volume 36, Issue 4, February (2020).

[5] Achiam, Josh and Adler, Steven and Agarwal, Sandhini and Ahmad, Lama and Akkaya, Ilge and Aleman, Florencia Leoni and Almeida, Diogo and Altenschmidt, Janko and Altman, Sam and Anadkat, Shyamal and others, "*Gpt-4 technical report*", arXiv preprint arXiv:2303.08774 (2023).

[6] Caufield, J. Harry, "*MACCROBAT*", doi:10.6084/m9.figshare.9764942.v2 (2019).

[7] Rezarta Islamaj Doğan, Robert Leaman, Zhiyong Lu, "*NCBI disease corpus: A resource for disease name recognition and concept normalization,* ",Journal of Biomedical Informatics, Volume 47, 1-10 (2014).