Semantic grounding of LLMs using knowledge graphs for query reformulation in medical information retrieval

Antonela Tommasel
ISISTAN, CONICET-UNCPBA, Argentina
Email: antonela.tommasel@isistan.unicen.edu.ar

Ira Assent
Aarhus University, Department of Computer Science
DIGIT Aarhus University Centre for
Digitalisation, Big Data and Data Analytics, Denmark
Email: ira@cs.au.dk

***Abstract*—The widespread adoption of electronic health records has generated a vast amount of patient-related data, mostly presented in the form of unstructured text, which could be used for document retrieval. However, querying these texts in full could present challenges due to their unstructured and lengthy nature, as they may contain noise or irrelevant terms that can interfere with the retrieval process. Recently, large language models (LLMs) have revolutionized natural language processing tasks. However, despite their promising capabilities, their use in the medical domain has raised concerns due to their lack of understanding, hallucinations, and reliance on outdated knowledge. To address these concerns, we evaluate a Retrieval Augmented Generation (RAG) approach that integrates medical knowledge graphs with LLMs to support query refinement in medical document retrieval tasks. Our initial findings from experiments using two benchmark TREC datasets demonstrate that knowledge graphs can effectively ground LLMs in the medical domain.**

***Index Terms*—Information retrieval, medical documents, query expansion, large language models, knowledge graphs**

I. Introduction

The rapid growth of published medical literature poses challenges for clinicians when attempting to retrieve relevant information through conventional systems [1]. Furthermore, the availability of electronic records has increased the volume of patient-related data, highlighting the need for tailored Clinical Decision Support systems for document retrieval and recommendation [2], aimed at assisting clinicians in patient care and decision-making processes.

Patient-related data, including demographics, family medical history, and current health issues, is typically conveyed as unstructured text. While this format eases documentation, it also poses challenges for document retrieval [3]. Issues such as medical jargon, negated content, and vocabulary mismatches between queries and documents can lead retrievers to prioritize documents closely matching lengthy queries rather than focusing on matching related information [4]. This can hinder retrieval performance, leading to the inclusion of irrelevant or noisy terms. The main challenge is to understand the complex information needs of medical applications and develop methods to enhance the discovery of relevant documents [5]. In this context, accurate query reformulation can help clinicians to better express their information needs, quickly access relevant medical knowledge, and make informed decisions, ultimately improving patient care.

Various language processing methods have been developed for query reformulation by extracting insights from clinical narratives [6]. These include techniques for information extraction, synonym resolution, and disambiguation. Recently, large language models (LLMs) have revolutionized natural language processing tasks. While LLMs excel in dialogue generation tasks, they struggle with factual information due to a lack of knowledge integration [7, 8]. This limitation hinders their effectiveness in knowledge-intensive tasks, often requiring domain knowledge [9]. In this regard, concerns about their suitability for the medical domain have been raised, including issues with hallucinations, reliance on outdated knowledge, and lack of interpretability [10, 11]. For example, *ChatGPT* can generate plausible but incorrect answers, contradicting clinical and societal values and potentially leading to the spread of medical misinformation and incorporating biases that could worsen health disparities [12, 13].

One effective solution for addressing these issues is exercising a Retrieval Augmented Generation (RAG) approach in which an external knowledge source is used to retrieve documents aligned with an initial task prompt, which are then passed to the LLM to produce the final output. Knowledge Graphs (KGs) represent interconnected entities that contextualize and link heterogeneous data resources, enabling the integration and analysis of information in an explicit and structured format [14, 15]. KGs are known for their symbolic reasoning capabilities and can evolve by incorporating new knowledge [11], allowing to adapt to dynamic domains. In the medical domain, KGs synthesize unstructured information by connecting patients, drugs, and diagnoses [14], and are used in medical knowledge retrieval, and drug analysis and discovery, among other tasks. In a RAG context, KGs play an instrumental role for LLMs by grounding them in factual knowledge, enhancing their knowledge awareness, and providing more control over generated answers, as the accessed knowledge can be inspected and interpreted [7].

We contribute a study to *assess whether medical KGs*

We gratefully acknowledge funding of this research by the Innovation Fund Denmark (IFD) through the Grand Solution project Hospital@Night.

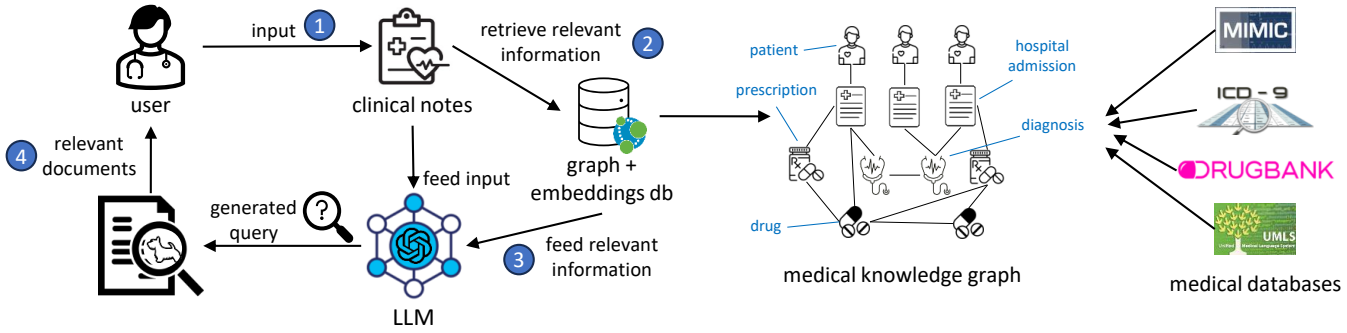

Fig. 1. Overview of the proposed pipeline

can guide LLMs in information extraction tasks and enhance document retrieval performance. Our approach integrates a medical KG and LLMs for query reformulation, focusing on case-based information retrieval, which involves retrieving the most relevant documents based on a patient's clinical note. Given a query represented by a patient's clinical notes, the goal is to reformulate it into a set of keywords for improved retrieval. We conducted an experimental evaluation over two benchmark TREC datasets. Results showed improved performance compared to existing approaches and, in some cases, comparable performance to human-generated queries. These preliminary findings suggest that KGs can effectively ground LLMs in the medical domain, avoiding fine-tuning costs, and contribute to developing techniques that increase the reliability and transparency of retrieval systems, thereby enhancing support for clinical decision-making [16]. Furthermore, this work shows how integrating structured knowledge with advanced language models can address challenges in managing and retrieving insights from large, unstructured datasets, a common issue in big data applications.

II. Related works

Query reformulation is essential for retrieval systems, aiding users in navigating the information space by refining queries, switching to alternative search tasks, or broadening the retrieved documents [17]. Query expansion involves extracting relevant concepts and entities (e.g., symptoms, drugs, events, medical procedures or treatments) from queries and augmenting them using various methods such as thesauri, query logs, and semantic relationships. For example, Zhu and Carterette [18] and Liu et al. [19] enriched queries with keywords from external medical resources, observing mixed impacts on performance. Huang et al. [20] used dictionaries, medical embeddings, and LDA for expansion, finding the best results with topic-based approaches. Peikos et al. [3] investigated expansion techniques for medical document retrieval and patient allocation in clinical trials. While some methods improved patient allocation, they did not significantly enhance document retrieval.

Different works have leveraged LLMs for query expansion in general domain tasks [21, 22]. Wang et al. [21] prompted the LLM to generate a paragraph answering the query (in a zero- or few-shot scenario), which was then concatenated to the query for retrieval. Results showed that the generated paragraph achieved worse retrieval performance than the original query, but combining both improved performance. Jagerman et al. [22] proposed several prompt alternatives (zero-shot, few-shot, RAG, chain-of-thought) to obtain either a set of keywords or a paragraph, which were concatenated to the original query for retrieval. For the RAG approach, context was defined as the top-3 documents retrieved for the original query. Results showed that the best performance was obtained when expanding the query with the explanation paragraph obtained with the chain-of-though approach. Results also showed that the RAG approach was more robust when combined with small LLMs. Observed differences were small, and, in most cases, not statistically significant.

Despite concerns about hallucinations (i.e., answers that are incorrect or non-sense), outdated knowledge, and inconsistent answers [23], LLMs have been applied in the medical domain [24] for tasks like, text anonymization [25], information extraction [26], and clinical note summarization [27]. However, outcomes have been inconclusive. General-purpose LLMs like GPT often fail to consistently outperform specialized models like BioBERT, particularly in tasks reliant on semantic similarity. Furthermore, performance has been found to be highly dependent on prompt formulation. Related to this study, Wang et al. [28] and Peikos et al. [29] leveraged ChatGPT for query expansion. Wang et al. [28] focused on generating and refining queries for systematic literature reviews. Results showed that ChatGPT included nonexistent medical terms in queries, although no significant correlation was found between the ratio of incorrect terms and query effectiveness. Peikos et al. [29] evaluated three expansion alternatives for clinical trial retrieval, varying the LLM's role and task description comprehensiveness. Results showed that some expansion alternatives were able to outperform baselines and, in some cases, human-generated queries. Nonetheless, several shortcomings like query drift, correct abbreviation resolution, and semantic changes were encountered.

III. Method

Retrieval-Augmented Generation (RAG) enhances the capabilities of LLMs by integrating external knowledge sources

during the generation process, allowing the model to access relevant and real-time information. While the LLM's knowledge remains fixed, it is grounded by up-to-date or domain-specific knowledge without requiring fine-tuning. Thus, RAG improves the reliability and consistency of responses, which helps to mitigate hallucinations [7]. In this context, we focus on *whether medical KGs are adequate knowledge sources to guide LLMs in query reformation tasks to enhance document retrieval performance*.

Most RAG approaches in the literature rely on textual documents as the external knowledge source. These documents are split into chunks and queried via similarity search to be appended to the task prompt. While effective for general tasks, this method may struggle when the task requires linking diverse pieces of information through shared attributes [30], or a holistic understanding of the semantic concepts across large datasets. Additionally, achieving the necessary document granularity to answer complex queries within similar chunks remains a challenge [30, 31]. In contrast, KGs not only allow to retrieve relevant nodes but also focused, context-rich subgraphs, enabling more accurate and informative responses [30]. KGs also provide transparent reasoning paths, aiding in the potential explanation of the generated responses.

To address the limitations of LLMs in the medical domain, we exercise a RAG model by leveraging an extended medical KG and LLMs to generate keyword-based query representations. Figure 1 provides an overview of our query expansion methodology. In particular, we study the combination of a KG and LLMs to generate the new query q'. Starting with a patient's clinical note representing a query q (1), the KG retrieves the context relevant to q (represented by graph nodes and their relations) (2). This context is then fed into the LLM (3) to generate a new query q' (e.g., by extracting relevant medical keywords from the context). Finally, the generated query q' is used for medical document retrieval (4).

Knowledge graph extraction

In the clinical domain, KGs usually include nodes representing entities like drugs, diseases, and proteins, along with their relations [14]. However, these KGs often only cover basic medical facts, lacking details on clinical outcomes, disease-drug prescriptions, and diagnostic interplays. For example, there might be an "adverse interaction" between drugs, but no information on how it affects simultaneous treatments [14]. Clinical data can bridge this gap by enriching traditional KGs with practical clinical outcomes [14]. Wang et al. [14] combined biomedical KGs with medical entities extracted from MIMIC-III (Medical Information Mart for Intensive Care III)[1], where nodes represent patients, admissions, prescriptions, drugs, and ICD (International Classification of Diseases) diagnoses, with hospital admissions linking patients to their medical histories and drug prescriptions.

Although the KG generated by Wang et al. [14] includes a comprehensive set of nodes and relations derived from patients' hospital admissions, it lacks of textual data describing the different nodes. In this context, we enrich the initial graph from [14][2] by adding new relations between nodes and different textual features that later will be used to find the relevant context for a query. For the drug nodes, we added their name, description, indication and actuation extracted from *DrugBank*[3]. We also add relations between drugs based on drug interaction information, also from *DrugBank*. For the diagnosis nodes, add their name from ICD-9[4]. We also add relations between them based on the ICD-9 hierarchy and semantic relations from UMLS (Unified Medical Language System)[5]. Finally, we link the hospital admission nodes to their corresponding entry in MIMIC-III and add the "chief complaint" and "history of recent illness" attributes. Using BioBERT [32], we process the "chief complaint" and "history of recent illness" to extract symptoms. These symptoms are added to the KG as a new node type to allow linking hospital admissions sharing common symptoms.

The resulting combination of structured (nodes and edges) and unstructured text enhances transparency and interpretability. The structure enables navigation and multi-hop reasoning for complex tasks [11] involving multiple heterogeneous entities, and allowing the discovery of indirect relations between them. For example, given a clinical note describing symptoms, it is possible to identify hospital admissions with similar symptoms, the drugs prescribed for those admissions, and other drugs targeting the same symptoms. Similarly, starting from a drug node, it is possible to find all treated diagnoses, and the hospital admissions related to them.

Retrieval of relevant context

We use a combined retrieval approach, first retrieving the graph nodes that are most similar to the query, and then expanding the node selection to include their induced subgraph based on their relations.

To find the most similar nodes (i.e., documents) to a given clinical note in the medical KG, we search the KG based on the embedding similarity of the newly added textual attributes, including diagnoses (name), drugs (name, description, action), and hospital admissions (chief complaint and history of recent illness). Both the clinical note and these textual attributes are embedded using a pre-trained *SentenceBERT* model[6]. The top-n most similar nodes to the clinical note are retrieved based on Cosine Similarity. Then, starting from these retrieved nodes, we capture their multi-hop context by defining a sub-graph including their neighbours at distance k. While all nodes can help identify interactions and paths, their textual content may not always align with the query's original intent. Therefore, restrictions can be applied to the types of nodes retrieved,

[1]https://mimic.mit.edu/

[2]A depiction of the graph schema can be found in the companion repository: https://github.com/tommantonela/query-rag-medical-retrieval

[3]https://go.drugbank.com/releases/5

[4]https://www.cdc.gov/nchs/icd/icd9cm.htm

[5]https://uts.nlm.nih.gov/uts/umls/home

[6]https://www.sbert.net/

such as limiting the search to nodes representing hospital admissions, or to how paths are navigated to find neighbours.

Retrieval Augmented Generation

In this RAG model, the external source is the medical KG, and the documents are the nodes in the retrieved sub-graph. The textual content of these nodes is concatenated (including connectors related to their semantic relation in the KG) with the original input prompt and fed to the LLM to produce the final output.

LLMs behaviour has been shown to be greatly influenced by the role assigned to them and the phrasing of the prompt [29, 33], affecting the generated responses, the vocabulary used, and the level of expertise reflected. We assigned the LLM the role of a "medical assistant" [29], resembling the duties of physician assistants, which include both clinical tasks and administrative responsibilities[7]. We define different prompts (shown in Table I), responding to diverse task definitions [29]:

i) `KG+LLM_Extract`. Extracts the most relevant keywords from the clinical note based on the provided context. Context should only be used as a ground for selecting the keywords in the clinical note without adding new keywords. Figure 2 shows an example for this prompt, including the query, the top-2 retrieved documents (no neighbouring documents are included), and the reformulated query.

ii) `KG+LLM_Expand`. Expands the clinical note with the most similar and relevant keywords from the context.

iii) `KG+LLM_Combined`. First, extracts the most relevant concepts from the clinical note. Then, expands them by adding keywords based on the provided context.

As shown in Table I, prompts also guide the LLM in the expected answer format, a comma-separated list of keywords with no extra information. The model is instructed to extract information only, which could also help to reduce hallucinations.

IV. Experimental setup

Data collections

Evaluation is based on the clinical notes and documents from the TREC 2014 and TREC 2015 Clinical Decision Support Track [34][8]. The goal is to retrieve relevant documents for answering generic questions. The collections comprise 60 clinical notes, each including a textual description (the base for all queries) and its summary. Documents belong to a subset of PubMed Central[9], and are judged as "non-relevant", "possibly relevant" and "definitely relevant". For each document, title, abstract and body text are indexed.

[7]https://www.aama-ntl.org/medical-assisting/what-is-a-medical-assistant

[8]Although there are newer data collections and TREC shared tasks, in addition to the TREC 2014 and TREC 2015 collections still being used [35], we chose them due to the availability of human curated keyword-based queries [36].

[9]http://www.ncbi.nlm.nih.gov/pmc/

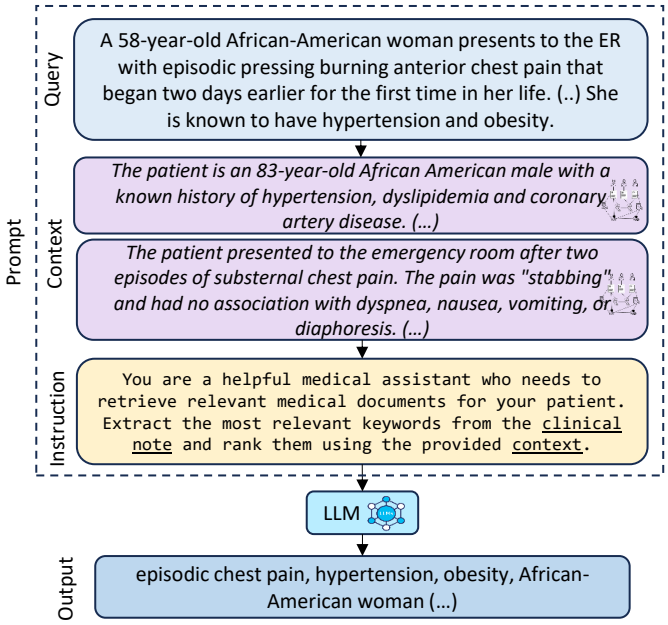

Fig. 2. A query reformulation example

Analyzed LLMs

We evaluate five general-domain LLMs, including commercial and open-source: `GPT-3.5`, `Llama3-7b`, `Mistral7b`, `Gemma2b-instruct`, and `Phi3:mini`. All models are accessed from *Ollama* or *HuggingFace* through *LangChain*. In all cases, temperature is set to 0, to reduce the responses' randomness, making it more focused and deterministic[10]. Each clinical note is processed in a separate interaction with the API. These models differ in scale, purpose, and adaptability. `GPT-3.5` is the largest, suitable for general-purpose tasks like content generation and reasoning. `Llama3-7b` and `Mistral7b` are smaller but offer high efficiency and strong performance in content generation. `Gemma2b-instruct` specializes in instruction-following tasks but can be less effective for broader applications, while `Phi3:mini` is a small model designed for resource-constrained environments.

Selected baselines

The RAG-generated queries are compared with[11]:

i) Full query description and summary.

ii) Keyword-based queries by extracting entities with pre-trained *biomedical models* (e.g., *BioBERT* and *SciSpacy*).

iii) Query expansion based on *RM3* [37].

iv) Query variants proposed by Peikos et al. [3], including detection of family history, negated content, historical information and entity expansion. These queries are chosen as they summarize multiple enrichment techniques in the literature.

v) Keyword-based queries by LLMs without KG grounding [29] (defined in Table I as `LLM w/o KG`).

[10]https://platform.openai.com/docs/api-reference/chat

[11]As the goal is not to outperform the results of the shared-task, but show the feasibility of grounding LLMs with KGs, we only compare the RAG generated queries with other keyword-based query reformulation approaches.

TABLE I
PROMPTS USED IN THE PERFORMED EVALUATION

System template	You are a helpful medical assistant that needs to retrieve relevant medical documents from the PMD medical document repository for your medical patient. Your task is to write a list of keywords that can be used in a search engine to search relevant medical documents.
ANSWER_FORMAT	Follow this format closely. Answers' format: • Respond with a comma-separated list of keywords that will be used for searching documents. • Do not include the same keywords twice. • Do not use bullets. • Do not use abbreviations. • Do not explain or elaborate. • Do not add any introductory or explaining sentence. • Return the keywords and only the keywords.
`LLM w/o KG`	From the *MEDICAL NOTE*, identify the patient's medical condition and current treatments, including any alternative names, abbreviations, or synonyms for these terms, as well as any additional criteria to create a keyword-based query that can be used to retrieve medical documents of interest. MEDICAL NOTE: {*medical_note*} *{ANSWER_FORMAT}*
`KG+LLM_Extract`	Extract the most relevant keywords from the *MEDICAL NOTE*, including medical concepts, symptoms, diseases, synonyms, or other relevant information. Use the CONTEXT to rank the extracted keywords. MEDICAL NOTE: *medical_note* CONTEXT: *context* {*ANSWER_FORMAT*}
`KG+LLM_Expand`	Using the provided CONTEXT, expand the MEDICAL NOTE with keywords referring to medical concepts, symptoms, diseases, synonyms or other relevant information. MEDICAL NOTE: *medical_note* CONTEXT: *context* ANSWER_FORMAT
`KG+LLM_Combined`	1. Using the provided CONTEXT, extract the most relevant keywords from the MEDICAL NOTE that fully describe its content. *ANSWER_FORMAT* 2. Use the provided *CONTEXT* and only the provided *CONTEXT*, expand the extracted list of keywords with 10 additional relevant keywords. *ANSWER_FORMAT* MEDICAL NOTE: *medical_note* CONTEXT: *context*

vi) Concatenation of keyword-based queries written by four human *clinician* assessors [36]. We summarized the overall medical assessor's knowledge by concatenating all unique keywords suggested by each assessor.

vii) Three LLMs tailored for the medical domain (`BioLLMs`): `MMedLM2`, `PMC-Llama`, and `Biomistral`. Since these models are already based on medical knowledge, they were only considered for the `LLM w/o KG` scenario [12].

Document retrieval

Retrieval is based on Pyterrier's [38] BM25. We focus on an sparse retriever, where query reformulation (and expansion) plays a crucial role in bridging the vocabulary gap or mismatch [22]. In contrast, dense retrieval systems, such as dual encoders, are less affected by vocabulary gaps and thus benefit less from query expansion techniques [22]. Their architecture inherently bridges semantic gaps without requiring external interventions, making them less sensitive to mismatches between query and document terms, a common issue in sparse retrieval systems. However, sparse retrieval systems are often preferred for their interpretability, efficiency, and scalability [39]. They provide transparent relevance scoring based on exact term matches, which is critical in domains like legal and medical applications. Additionally, sparse systems are generally more robust to domain shifts since they rely on exact matches rather than on embeddings learned from training data, which can limit the effectiveness of dense retrievers on out-of-domain data.

Documents are indexed using the default Pyterrier stemming and stopword removal parameters. Performance is evaluated following TREC guidelines, considering precision, nDCG (normalized Discount Cummulative Gain) and binary preference (bpref) [40]. Paired t-tests, implemented by PyTerrier, are used to evaluate the significance of differences.

We also evaluate the quality of the LLMs' responses and the provided KG-extracted context. This evaluation uses three metrics from *DeepEval*[13]: i) *Answer Relevancy* (measures how relevant the obtained response is compared to the provided input), ii) *Faithfulness* (measures whether the obtained response factually aligns with the RAG context), and iii) *Contextual Relevancy* (measures the relevance of the RAG context for the given input, independently from the obtained LLM output). As these metrics rely on LLMs to compute scores and provide explanations of those scores, we average the results obtained for two LLMs, `Gemma2b` and `Llama3`.

V. EXPERIMENTAL EVALUATION

Table II presents the average top-10 retrieval performance for the evaluated clinical notes[14], and Table III presents the

[12]We evaluated the inclusion of three additional models, but they were discarded due to the quality of the generated responses, reported performance lower than GPT-3.5 or availability. More details regarding the medical LLMs can be found in the supplementary material in the repository.

[13]https://docs.confident-ai.com/

[14]Detailed results for other top-k can be found in the repository.

TABLE II
AVERAGE TOP-10 RETRIEVAL PERFORMANCE FOR THE GENERATED QUERIES
BEST RESULTS IN **BOLD**, SECOND BEST ARE UNDERLINED.

	TREC 2014			TREC 2015		
	prec	nDCG	bpref	prec	nDCG	bpref
description	0.083	0.066	0.036	0.143	0.123	0.059
summary	0.077	0.068	0.035	0.143	0.11	0.076
biomedical models	0.12	0.122	0.113	0.143	0.114	0.096
description + RM3	0.126	0.095	0.135	0.156	0.121	0.106
Peikos et al. [3]	0.103	0.089	0.035	0.153	0.113	0.075
BioLLM-pmc-Llama	0.13	0.12	0.107	0.12	0.1	0.081
BioLLM-BioMistral	0.06	0.059	0.044	0.043	0.045	0.033
BioLLM-MMedLM2	0.077	0.059	0.102	0.11	0.103	0.094
human-generated [36]	**0.3**	**0.279**	**0.26**	**0.34**	**0.275**	**0.223**
GTP-3.5 w/o KG	0.11	0.102	0.084	0.15	0.115	0.054
Mistral w/o KG	0.09	0.089	0.102	0.127	0.106	0.043
Phi3 w/o KG	0.13	0.124	0.118	0.123	0.096	0.082
Llama3 w/o KG	0.13	0.134	0.111	0.127	0.112	0.126
Gemma2b w/o KG	0.113	0.093	0.105	0.113	0.105	0.101
KG+LLM_Ext-GPT3.5	0.103	0.098	0.06	0.15	0.123	0.086
KG+LLM_Exp-GPT3.5	0.11	0.11	0.094	0.11	0.084	0.108
KG+LLM_Comb-GPT3.5	0.103	0.097	0.052	0.023	0.022	0.041
KG+LLM_Ext-Mistral	0.13	0.117	0.073	0.167	0.151	0.107
KG+LLM_Exp-Mistral	0.127	0.145	0.083	0.147	0.128	0.098
KG+LLM_Comb-Mistral	0.09	0.073	0.06	0.147	0.122	0.097
KG+LLM_Ext-Phi3	0.097	0.092	0.112	0.093	0.088	0.078
KG+LLM_Exp-Phi3	0.123	0.107	0.099	0.087	0.071	0.071
KG+LLM_Comb-Phi3	0.063	0.053	0.07	0.13	0.102	0.067
KG+LLM_Ext-Llama3	0.083	0.083	0.05	0.06	0.05	0.048
KG+LLM_Exp-Llama3	0.083	0.09	0.066	0.107	0.079	0.053
KG+LLM_Comb-Llama3	0.137	0.12	0.099	0.13	0.099	0.096
KG+LLM_Ext-Gemma2b	0.09	0.07	0.072	0.07	0.057	0.073
KG+LLM_Exp-Gemma2b	0.06	0.046	0.068	0.057	0.044	0.064
KG+LLM_Comb-Gemma2b	0.06	0.045	0.081	0.083	0.066	0.078

average results for the *DeepEval* metrics across clinical notes. *Faithfulness* and *Contextual Relevancy* are only available for those alternatives exercising the RAG scenario. Retrieved documents are considered correct if they are tagged as "possibly relevant" or "definitely relevant" to the corresponding clinical note. For [3, 29], we selected the best query in the original papers.

The response quality of LLMs varied across models. GPT-3.5 and Gemma2b closely following the defined output format, while Mistral7b and Llama3 required some pre-processing to remove extra phrases and itemizations. In contrast, BioLLMs required extensive pre-processing, as their responses often did not match the expected output and included, for example, repetitions of the provided input.

Baselines

For the *clinicians* query, Table II reports the results for the concatenation of all medical assessors, which achieved the highest results. As reported by Peikos et al. [3], *clinicians* achieved the best performance, showing statistically significant differences for most clinical notes compared to *description*, *summary*, *biomedical models*, BioLLMs, [3] and LLM w/o KG. BioLLMs performed statistically significantly better than *summary*. BioLLM-PMC-Llama achieved similar results to *biomedical models*, but no statistically significant differences were observed. *Answer Relevancy* showed similar scores for the three BioLLMs.

TABLE III
AVERAGE RESULTS FOR LLM PERFORMANCE IN TERMS OF RESPONSE RELEVANCE TO INPUT AND ALIGNMENT BETWEEN THE KG CONTEXT AND THE RESPONSE
BEST RESULTS IN **BOLD**, SECOND BEST ARE UNDERLINED.

	TREC 2014		TREC 2015	
	Answer Rel.	Faithful.	Answer Rel.	Faith.
BioLLM-PMC-Llama	0.769	-	0.763	-
BioLLM-BioMistral	0.728	-	0.754	-
BioLLM-MMedLM2	0.791	-	0.797	-
GPT-3.5 w/o KG	0.746	-	0.762	-
Mistral w/o KG	0.723	-	0.735	-
Phi3 w/o KG	0.772	-	0.761	-
Llama3 w/o KG	0.782	-	0.804	-
Gemma2b w/o KG	0.718	-	0.714	-
KG+LLM_Ext-GPT-3.5	0.856	0.132	0.816	0.146
KG+LLM_Exp-GPT-3.5	0.715	0.06	0.712	0.096
KG+LLM_Comb-GPT-3.5	0.893	0.109	0.882	0.163
KG+LLM_Ext-Mistral	0.814	0.114	0.864	0.138
KG+LLM_Exp-Mistral	0.705	0.150	0.717	0.107
KG+LLM_Comb-Mistral	0.89	0.167	0.9	0.151
KG+LLM_Ext-Phi3	0.744	0.520	0.802	**0.568**
KG+LLM_Exp-Phi3	0.707	0.487	0.711	0.517
KG+LLM_Comb-Phi3	0.850	**0.515**	0.814	0.535
KG+LLM_Ext-Llama3	0.783	0.214	0.743	0.188
KG+LLM_Exp-Llama3	0.705	0.183	0.726	0.093
KG+LLM_Comb-Llama3	**0.906**	0.159	0.908	0.167
KG+LLM_Ext-Gemma2b	0.866	0.142	0.888	0.147
KG+LLM_Exp-Gemma2b	0.706	0.203	0.714	0.162
KG+LLM_Comb-Gemma2b	**0.907**	0.175	**0.919**	0.158

LLM w/o KG

Statistically significant differences favouring LLM w/o KG were observed compared with *summary* and *description*, with improvements in up to 70% of queries for bpref. No statistically significant differences were found for GPT-3.5 and Mistral7b compared with *biomedical models* and *clinicians*. On other hand, Phi3, Llama3, and Gemma2b showed statistically significant differences with *clinicians* for up to 30% of queries in terms of nDCG. Most differences for BioLLMs were not statistically significant, except for Mistral7b and BioLLM-PMC-Llama (with Mistral7b improving results for up to 50% of clinical notes) and Phi3, Llama3, and Gemma2b compared to BioLLM-MMedLM2 (with improvements for up to 60% of clinical notes).

Phi3 and Llama3 achieved similar levels of *Answer Relevancy* to BioLLMs, even improving the scores of BioLLM-PMC-Llama, although differences were not statistically significant. These results suggest that fine-tuning LLMs to a specific domain does not necessarily improve performance across all related tasks. Results also highlighted the strong performance of *biomedical models* (even compared to BioLLMs) for entity/keyword identification. As noted by Peikos et al. [29], LLMs sometimes struggled with correctly identifying negated content, which can affect concept expansion and, thus, retrieval performance.

KG+LLM – Number of retrieved documents

We evaluated retrieving the top-1, 2, 3 and 5 nodes/documents and the corresponding sub-graphs including neighbours at distance 1 from the KG. The best results for KG+LLM, reported in Table II, were achieved when retrieving the top-2 documents/nodes, for which statistically

significant differences were observed for most models[15]. These results suggest that adding more context does not necessarily lead to better performance, as a larger context can "confuse" the LLMs, affecting their ability to generate relevant responses [41]. This is supported by the observation that *Contextual Relevancy* decreased as the number of top documents increased. For example, retrieving the top-2 documents decreased the *Contextual Relevancy* by a 14% and 2% for TREC 2014 and TREC 2015 compared to retrieving the top-1, while retrieving the top-3 documents decreased it by 35% and 29% for TREC 2014 and TREC 2015 when compared to retrieving the top-2 documents. Similarly, the *Faithfulness* of responses tended to peak when retrieving the top-2 or top-3 documents. This highlights the importance of thoroughly assessing the quality of the KG and the retrieved context.

KG+LLM – Context length

Since increasing the number of documents in the context also increases the number of tokens in the final prompt, we assessed the effects of prompt compression by removing non-essential tokens. This helps accelerate model inference and reduce costs [42].

We implemented prompt compression using *LLMLingua* [42][16] with the default model parameters. After preliminary evaluations, we set the number of tokens to 1000. To assess the semantic similarity between the original and compressed prompts, we computed the Cosine Similarity over their *SentenceTransformers* embeddings[17]. The average similarity between the original and compressed context across queries was over 0.75 ($\pm$0.1), indicating a high semantic agreement between both texts.

For TREC 2015, as the number of retrieved documents/nodes increased, differences between compressing and not compressing the prompt became larger. While minimal differences were observed when retrieving the top-1 and top-2 documents, significant differences emerged when retrieving the top-3 and top-5 documents across the three defined prompts. In these cases, compressing the context improved results for at least half of the queries. For TREC 2014, statistically significant differences favouring the compressed prompts were observed when retrieving the top-2 and top-3 documents. Unlike the results observed for the uncompressed prompts when retrieving the top-2 documents, compressing prompts also increased *Contextual Relevancy* by 115% (0.247 vs 0.532) for TREC 2014 and 104% (0.246 vs 0.504) for TREC 2015 for the top-2 documents. These improvements increased as the number of selected top documents also increased. These findings suggest that prompt compression is a valuable strategy for enhancing the relevance of the context, particularly when dealing with larger ones, by effectively mitigating noise and ensuring that relevant information is prioritized, without negatively affecting retrieval performance.

[15]The result for the other top-n can be found in the companion repository.
[16]https://github.com/microsoft/LLMLingua
[17]*all-mpnet-base-v2*: https://www.sbert.net/.

KG+LLM – Baseline comparison

Although KG+LLM did not improve the best results of *clinicians* for more than 20% of queries, they improved results for at least one medical assessor for over 30% of queries. Statistically significant differences were observed when compared to *description* and *summary*, with KG+LLM improving, on average, over 50% of queries in terms of bpref for TREC 2014. For TREC 2015, differences were observed in precision and nDCG, with improvements for 15% of queries. Similar trends were observed when compared to *biomedical models*, where KG+LLM improved results for 10% of queries in terms of bpref for Mistral7b. For GPT-3.5, statistically significant differences were only observed for KG+LLM_Combine for 10% of queries.

When compared with BioLLMs, Llama3 statistically significant improved bref for 27% of queries. When compared to LLM w/o KG, statistically significant differences were observed for Mistral7b and Llama3 in both datasets, with KG+LLM improving results for at least 30% and 13% of queries, respectively. For Gemma2b, differences were only statistically significant for TREC 2015. No statistically significant differences were observed for GPT-3.5 and Phi3. The lack of statistically significant differences between KG+LLM and BioLLMs for most LLMs and prompt definitions suggests that grounding LLMs with a medical KG could be a cost-effective alternative to re-training and fine-tuning LLMs, while still providing adequate responses [7].

In general, KG+LLM statistically significantly improved the *Answer Relevancy* of LLM w/o KG for all models and at least two prompt variations, for both TREC 2014 and TREC 2015. However, for some combinations of LLM and prompt (e.g., for Gemma2b), performance decreased when adding the KG context, suggesting a limited capacity to leverage such provided context, possibly due to the quality of the prompt. These findings highlighted that KGs can provide LLMs with sufficient information for extracting and expanding medical concepts while offering control over the information source.

KG+LLM – LLM comparison

For TREC 2014, statistically significant improvements favouring Mistral7b were observed when compared to Gemma2b and Llama3 for the KG+LLM_Extract and KG+LLM_Extend prompts. Mistral7b improved 73% of queries when compared to Llama3 and 36% for Gemma2b in terms of bpref. For KG+LLM_Extract, GPT-3.5 and Phi3 improved the bpref of Llama3 for over 53% of queries. For KG+LLM_Combine, the only statistically significant differences were observed for GPT-3.5 over Gemma2b, improving results for 43% of queries. For TREC 2015, statistically significant differences in bpref were observed for KG+LLM_Extract and KG+LLM_Combine favouring Phi3, Mistral7b, and Gemma2b over Llama3, with improvements for 26% to 40% queries. Additionally, Gemma2b outperformed Phi3 for 26% of queries for KG+LLM_Combine. Phi3 achieved slightly lower average *Answer Relevancy* than the other LLMs, with an aver-

age 5% difference (with significant differences observed for KG+LLM_Combine), but achieved the highest *Faithfulness*, with an average 290% statistically significant improvement, highlighting its ability to capture information from the provided KG context. Conversely, Mistral7b, despite being one of the best-performing LLMs, achieved good *Answer Relevancy* but low *Faithfulness*. These results suggested the need for further analysis of the KG's quality and its integration with LLMs.

KG+LLM – Prompt analysis

For TREC 2014, statistically significant differences were observed between KG+LLM_Extend and KG+LLM_Extract, favouring the former for 50% of the queries in terms of bpref. For Phi3, KG+LLM_Combine outperformed KG+LLM_Extend in terms of precision and nDCG for nearly 50% of queries. Overall, KG+LLM_Combine achieved the highest *Answer Relevancy* for all LLMs, followed by KG+LLM_Extract, with significant differences observed for all LLMs. No significant differences were observed for *Faithfulness*. Differences varied based on the number of retrieved documents, emphasizing the importance of appropriately defining the context for the LLM.

To examine the effect of prompting, we introduced slight variations to KG+LLM_Extract's prompt by asking the LLM to "summarize" the clinical note with keywords instead of "extracting" keywords. For TREC 2014, contrasting results were observed for Mistral7b and Llama3. While KG+LLM_Extract achieved the best results for about 60% of queries in terms of bpref for Mistral7b, the "summarize" prompt improved results for about 30% of queries in terms of precision for Llama3. No differences were observed for GPT-3.5. These observations suggested varying levels of sensitivity to slight changes in prompt wording.

KG+LLM – Impact of neighbour expansion

We compared results with and without expanding the retrieved documents with their neighbours beyond distance 1. We observed small but significant differences favouring not further expanding the retrieved documents. This finding, similar to the effect of the number of retrieved documents, may be related to the length and order of the context, reinforcing the importance of further evaluating the quality of the KG and the retrieved context.

In summary, results showed that KGs can effectively ground LLMs in the medical domain, achieving performance comparable to baselines in the literature, *BioLLM* models and hand-crafted clinician queries, without the need for costly fine-tuning. KGs provided sufficient information for extracting and expanding medical concepts while allowing control over information sources. In particular, sensitivity to prompt wording and the length and order of context emerged as important factors of retrieval quality, emphasizing the need for further evaluation of the retrieved context quality and its integration with the LLMs.

VI. Conclusions

We studied whether KGs can enhance LLMs query generation capabilities for document retrieval. Experimental evaluation showed that grounding LLMs with external knowledge sources leads to comparable, and sometimes improved, performance. These results suggest that KGs can successfully ground LLMs in the medical domain. While focused on the medical domain, the methods and findings could be transferable to other domains involving big unstructured datasets (such as finance and legal), broadening the impact of the proposal.

The study acknowledges several limitations for future improvement. First, the study was conducted using specific benchmark TREC datasets, which may not fully represent the diversity of real-world clinical texts. Then, to ensure the generalizability of observations, the study can be extended to other related tasks, such as clinical trial allocation [43], or even more complex scenarios such as treatment suggestion. Second, while major approaches are included, our study does not cover all query reformulation techniques, thus, replication with additional techniques and advanced retrievals is recommended. Also, LLMs could be used for the full retrieval pipeline. Third, based on the observed differences, a more comprehensive examination of prompt variations is needed [29]. Additional more complex prompting techniques, like chain-of-thought, could be evaluated [22], in which the LLMs work in smaller tasks and provide the reasoning for them. For example, instead of directly extracting or suggesting keywords, LLMs could be asked to summarize the text, identify the most central terms, and then expand that set with related keywords. Fourth, the evaluation showed that each LLM had different performance under the different prompts and added context, suggesting that an ensemble of LLMs could be employed to leverage their complementary strengths and improve coverage across various medical document retrieval tasks. Finally, further investigation is necessary into the impact of document and graph embedding representation, and how to select the relevant documents, as well as context's length and order, as these factors can affect LLMs' response generation ability [41].

Ethical statement

The presented method does not aim to influence patient care decisions; hence, its outputs do not directly impact patient care. Since the KG grounds the LLM, and the task involves extracting rather than generating new content, the risk of hallucinations should not affect performance. However, using LLMs to analyze patient data raises valid privacy and security concerns. Implementing the proposed approach will require anonymizing and de-identifying data, or even using a locally hosted LLM.

References

[1] M. Agosti, G. M. D. Nunzio, S. Marchesin, and G. Silvello, "A relation extraction approach for clinical decision support," in *Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management*

(CIKM 2018), Torino, Italy, October 22, 2018, ser. CEUR Workshop Proceedings, A. Cuzzocrea, F. Bonchi, and D. Gunopulos, Eds., vol. 2482. CEUR-WS.org, 2018. [Online]. Available: https://ceur-ws.org/Vol-2482/paper4.pdf

[2] W. Hersh, *Information retrieval: a health and biomedical perspective*. Springer Science & Business Media, 2008.

[3] G. Peikos, D. Alexander, G. Pasi, and A. P. de Vries, "Investigating the impact of query representation on medical information retrieval," in *European Conference on Information Retrieval*. Springer, 2023, pp. 512–521.

[4] Y. Wang and H. Fang, "Key terms guided expansion for verbose queries in medical domain," in *Information Retrieval Technology: 14th Asia Information Retrieval Societies Conference, AIRS 2018, Taipei, Taiwan, November 28-30, 2018, Proceedings 14*. Springer, 2018, pp. 143–156.

[5] L. Tamine and L. Goeuriot, "Semantic information retrieval on medical texts: Research challenges, survey, and open issues," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1–38, 2021.

[6] M. Y. Landolsi, L. Hlaoua, and L. Ben Romdhane, "Information extraction from electronic medical documents: state of the art and future research directions," *Knowledge and Information Systems*, vol. 65, no. 2, pp. 463–516, 2023.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[8] M. Kang, J. M. Kwak, J. Baek, and S. J. Hwang, "Knowledge graph-augmented language models for knowledge-grounded dialogue generation," *arXiv preprint arXiv:2305.18846*, 2023.

[9] W. Yu, D. Iter, S. Wang, Y. Xu, M. Ju, S. Sanyal, C. Zhu, M. Zeng, and M. Jiang, "Generate rather than retrieve: Large language models are strong context generators," *arXiv preprint arXiv:2209.10063*, 2022.

[10] J. Albrecht, E. Kitanidis, and A. J. Fetterman, "Despite" super-human" performance, current llms are unsuited for decisions about ethics and safety," *arXiv preprint arXiv:2212.06295*, 2022.

[11] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[12] S. B. Patel and K. Lam, "Chatgpt: the future of discharge summaries?" *The Lancet Digital Health*, vol. 5, no. 3, pp. e107–e108, 2023.

[13] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[14] M. Wang, J. Zhang, J. Liu, W. Hu, S. Wang, X. Li, and W. Liu, "Pdd graph: Bridging electronic medical records and biomedical knowledge graphs via entity linking," in *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II 16*. Springer, 2017, pp. 219–227.

[15] X. Wu, J. Duan, Y. Pan, and M. Li, "Medical knowledge graph: Data sources, construction, reasoning, and applications," *Big Data Mining and Analytics*, vol. 6, no. 2, pp. 201–217, 2023.

[16] M. I. Hossain, G. Zamzmi, P. R. Mouton, M. S. Salekin, Y. Sun, and D. Goldgof, "Explainable ai for medical data: Current methods, limitations, and future directions," *ACM Comput. Surv.*, dec 2023, just Accepted. [Online]. Available: https://doi.org/10.1145/3637487

[17] A. Mustar, S. Lamprier, and B. Piwowarski, "On the study of transformers for query suggestion," *ACM Transactions on Information Systems (TOIS)*, vol. 40, no. 1, pp. 1–27, 2021.

[18] D. Zhu and B. Carterette, "Using multiple external collections for query expansion." in *TREC*, 2011.

[19] L. Liu, L. Liu, X. Fu, Q. Huang, X. Zhang, and Y. Zhang, "A cloud-based framework for large-scale traditional chinese medical record retrieval," *Journal of biomedical informatics*, vol. 77, pp. 21–33, 2018.

[20] E. W. Huang, S. Wang, D. J.-L. Lee, R. Zhang, B. Liu, X. Zhou, and C. Zhai, "Framing electronic medical records as polylingual documents in query expansion," in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 940.

[21] L. Wang, N. Yang, and F. Wei, "Query2doc: Query expansion with large language models," *arXiv preprint arXiv:2303.07678*, 2023.

[22] R. Jagerman, H. Zhuang, Z. Qin, X. Wang, and M. Bendersky, "Query expansion by prompting large language models," *arXiv preprint arXiv:2305.03653*, 2023.

[23] X. Yang, A. Chen, N. M. Pournejatian, H.-C. Shin, K. E. Smith, C. Parisien, C. B. Compas, C. Martin, A. B. Costa, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. P. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian, and Y. Wu, "A large language model for electronic health records," *NPJ Digital Medicine*, vol. 5, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:255175535

[24] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, "A bibliometric review of large language models research from 2017 to 2023," *arXiv preprint arXiv:2304.02020*, 2023.

[25] Z. Liu, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, W. Liu, D. Shen, Q. Li *et al.*, "Deid-gpt: Zero-shot medical text de-identification by gpt-4," *arXiv preprint arXiv:2303.11032*, 2023.

[26] B. Jimenez Gutierrez, N. McNeal, C. Washington, Y. Chen, L. Li, H. Sun, and Y. Su, "Thinking about GPT-3 in-context learning for biomedical IE? think again," in *Findings of the Association for Computational*

Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4497–4512. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.329

[27] Y.-N. Chuang, R. Tang, X. Jiang, and X. Hu, "Spec: A soft prompt-based calibration on mitigating performance variability in clinical notes summarization," *arXiv preprint arXiv:2303.13035*, 2023.

[28] S. Wang, H. Scells, B. Koopman, and G. Zuccon, "Can chatgpt write a good boolean query for systematic review literature search?" in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1426–1436. [Online]. Available: https://doi.org/10.1145/3539618.3591703

[29] G. Peikos, S. Symeonidis, P. Kasela, and G. Pasi, "Utilizing chatgpt to enhance clinical trial enrollment," *arXiv preprint arXiv:2306.02077*, 2023.

[30] D. Sanmartin, "Kg-rag: Bridging the gap between knowledge and creativity," *arXiv preprint arXiv:2405.12035*, 2024.

[31] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," *arXiv preprint arXiv:2212.10496*, 2022.

[32] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[33] A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, and K. Narasimhan, "Toxicity in chatgpt: Analyzing persona-assigned language models," *arXiv preprint arXiv:2304.05335*, 2023.

[34] M. S. Simpson, E. M. Voorhees, and W. R. Hersh, "Overview of the trec 2014 clinical decision support track." in *TREC*, 2014.

[35] Z. Zhao, Q. Jin, F. Chen, T. Peng, and S. Yu, "A large-scale dataset of patient summaries for retrieval-based clinical decision support systems," *Scientific Data*, vol. 10, no. 1, p. 909, 2023.

[36] B. Koopman and G. Zuccon, "A test collection for matching patients to clinical trials," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 669–672.

[37] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade, "Umass at trec 2004: Novelty and hard," *Computer Science Department Faculty Publication Series*, p. 189, 2004.

[38] C. Macdonald and N. Tonellotto, "Declarative experimentation in information retrieval using pyterrier," in *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, 2020, pp. 161–168.

[39] N. Arabzadeh, X. Yan, and C. L. Clarke, "Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 2862–2866.

[40] C. Van Gysel and M. de Rijke, "Pytrec_eval: An extremely fast python interface to trec_eval," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 873–876.

[41] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *arXiv preprint arXiv:2307.03172*, 2023.

[42] H. Jiang, Q. Wu, C.-Y. Lin, Y. Yang, and L. Qiu, "LLMLingua: Compressing prompts for accelerated inference of large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 358–13 376. [Online]. Available: https://aclanthology.org/2023.emnlp-main.825

[43] I. Soboroff, "Overview of trec 2021," in *30th Text REtrieval Conference. Gaithersburg, Maryland*, 2021.