UNIMIB at TREC 2023
Clinical Trials Track

Georgios Peikos

Information and Knowledge Representation, Retrieval, and Reasoning (IKR3) Lab,
Department of Informatics, Systems, and Communication (DISCo),
University of Milano-Bicocca, Milan, Italy
georgios.peikos@unimib.it

Abstract. This notebook summarizes our participation as the UNIMIB team in the TREC 2023 Clinical Trials Track. Our research evaluates the efficacy of Large Language Models (LLMs) in assessing patient eligibility for clinical trials. For this purpose, we integrated GPT-3.5 as the final stage in our retrieval pipeline. The results indicate that GPT-3.5 may enhance the performance of retrieval tasks in this context. Nonetheless, comparable results may be attained with less complex retrieval systems that utilize BM25.

1 Introduction

The 2023 TREC Clinical Trials track is a continuation of the two previous years by introducing a new document collection and simulating a different search situation. This year's document collection has been sourced from a recent snapshot of ClinicalTrials.gov (May 8, 2023) and has the same characteristics as those used in previous years. The track organizers initially created forty queries, of which relevance judgments are available for thirty-seven. As a results, the retrieval performance presented in this notebook is based on 37 queries. The current year's queries simulate a "questionnaire" completed by patients or their healthcare providers to ascertain their suitability for clinical trials. The track introduces a variety of high-level disorder-specific questionnaire templates, such as those for glaucoma and anxiety. The ultimate aim of the system is to accurately identify and retrieve clinical trials where a patient's profile satisfies the inclusion and does not meet any exclusion criteria.

This report outlines our team's (UNIMIB) submission to TREC 2023 Clinical Trials Track, that diverges from the methodologies employed in our 2021 and 2022 submissions [1]. The objective of our experiments is to explore the extent to which Large Language Models (LLMs) can determine a patient's eligibility for clinical trials and assess the impact of their deployment on the retrieval efficacy.

2 Experiments

This section outlines the essential aspects of our experimental methodology. It encloses the query processing and creation approach, the extraction of information from the clinical trial documents, and the estimation of relevance between the queries and the clinical trials.

2.1 Query Processing

Our query-processing approach is designed to process a patient questionnaire by extracting relevant information and constructing it into a search keyword-based query. Initially, it identifies and discards any negative or non-applicable responses from the questionnaire. For instance, if in the questionnaire the "¡field name=" prior cataract surgery"¿no¡/field¿," none of the "prior cataract surgery" will be included in the final query. Then, it assembles a query by affirmatively stating the patient's condition and including only pertinent, positive details about their health status. For example, "The patient has glaucoma" or "The intraocular pressure is 48 mmHg." The resulting query synthesizes the patient's present information in a way that aligns with identifying suitable clinical trials. The methodology utilized in the script selectively excludes negative responses by employing a predefined list of negated terms. However, for eligibility screening, this exclusion is problematic because the screening process demands a thorough profile that encompasses both positive and negative health attributes. Including such negative attributes is essential to assess a patient's eligibility for clinical trials fully. Nonetheless, the created queries contain sufficient information for retrieval. Also, by using them, the system may avoid retrieving clinical trials unrelated to the patient's conditions.

2.2 Information Extraction from Documents & Indexing

Clinical trials are structured documents with various fields such as title, summary, studied condition, among others. In addition, in this task, a trial's inclusion and exclusion criteria are mentioned in a semi-structured format within its eligibility section, holding great importance.

Our methodology exploits four document representations, each containing different document information. In detail, using a set of regex rules that leverage the semi-structured format of the inclusion and exclusion criteria, we extract them to create two distinct indices. Specifically, the *I_in* index contains only a trial's inclusion criteria, and the *I_ex* index only the trial's exclusion criteria. In the cases where their extraction was not feasible, the whole eligibility section has been indexed in both indices, i.e., the *I_in* and *I_ex*. In addition to those two indices, we construct a third one. Here, the title, description, studied condition, and summary sections were combined in a single text and indexed; this index will be referred to as *I_main*. Lastly, we indexed all document sections to create the *I_comb* representation used in one of our experiments. For indexing we have employed PyTerrier [2], with its default parameters.

2.3 Relevance Estimation

To estimate topical relevance in our initial retrieval stage, we employed the BM25 model along with the RM3 model for pseudo-relevance feedback. The four submitted runs represent different configurations of retrieval strategies combining BM25, the RM3 model, and Large Language Models, i.e. GPT-3.5. The following explain each run:

BM25RM3_single_run. This run utilizes the BM25 algorithm enhanced by the RM3 model for pseudo-relevance feedback. The RM3 model leveraged the top-5 documents retrieved by BM25 and expanded the query with 20 terms. Relevance has been estimated using the *I_comb* index.

BM25RM3_two_stage. Here, the BM25 model, integrated with the RM3 model, performs the initial retrieval of clinical trials. This process utilizes an index, denoted as *I_comb*, to retrieve a set of 1000 clinical trials. The output from this stage is then re-ranked using the same models, but this time it incorporates solely information from the eligibility and the title sections of the clinical trial documents. The aim here is to refine the retrieval by adding an additional relevance signal from the title, possibly enhancing the precision of the results.

BM25RM3_single_gpt3.5_strict. In this approach, the initial retrieval is performed by the BM25RM3_single_run, which leverages BM25 integrated with RM3 for pseudo-relevance feedback. The top 50 clinical trials identified through this process are then subjected to a re-ranking process, where GPT-3.5 is employed to determine the patient's eligibility. GPT3.5 was instructed to determine with a "YES" or a "NO," whether the considered patient is eligible for a given clinical trial. The clinical trials that received a positive eligibility response were reordered based on their topical relevance scores, positioning the trial with the highest combined relevance and eligibility score at the forefront. Similarly, trials with a negative eligibility response were reordered, with their ranking determined solely by their topical relevance scores. The prompt utilized in this experiment is:

> *Based solely on the given patient and trial information and ignoring [age, locations,gender] requirements, provide a simple 'YES' if the patient is eligible for the clinical trial or 'NO' if they are not eligible. Do not provide explanations or assumptions. [patient information]: [trial information]:*

The "strict" aspect refers to the prompt constraining GPT-3.5 to use only the information explicitly stated in the provided texts. That ensures that the LLM's estimations are based solely on the available data without introducing external knowledge, aiming to maintain the specificity of the relevance assessment.

BM25RM3_single_gpt3.5_run. This run is similar to the third but with a less stringent prompt provided to ChatGPT, implying that the LLM may utilize its knowledge in addition to the provided texts to estimate eligibility. That could allow for a broader interpretation of the queries and capture more relationships between patient information and trial criteria. The sole distinction in the current prompt, as compared to preceding one, lies in the initial sentence.

> *Based on the given patient and trial information, your knowledge, and ignoring [age, locations,gender] requirements, provide....*

The experiments were designed to explore the efficacy of incorporating GPT-3.5 in the evaluation of clinical trial eligibility, comparing its performance in comparison to conventional lexicon-based retrieval systems.

3 Results

Table 1 displays the results obtained by the four distinct runs submitted for evaluation. As it can be seen in Table 1, integrating LLMs such as GPT-3.5 into the retrieval process can enhance retrieval effectiveness in clinical trial retrieval. Notably, the BM25RM3_single_gpt3.5_run outperformed other runs in terms of nDCG@10 and P@10. Processing the top 50 clinical trials using the OpenAI's API for each of the forty patient cases cost approximately ten dollars and was completed within thirty minutes. The sensitivity of GPT-3.5 to variations in prompt structure is evident. The strict prompt, which constrained GPT-3.5 to use only the provided patient and trial information, achieved the lowest scores for nDCG@10 and P@10 and the higher performance in reciprocal rank. Conversely, the less restricted prompt allowing GPT-3.5 to utilize its broader knowledge base yielded the highest scores across all evaluated metrics. It achieved the highest nDCG@10 and P@10 and the lowest reciprocal rank score. That suggests that the specificity of the instructions provided in the prompt can significantly influence GPT-3.5's performance in clinical trial eligibility assessment.

Table 1. Retrieval performance achieved by the submitted runs.

	NDCG@10	P@10	Reciprocal Rank
BM25RM3_single_run	.618	.400	.552
BM25RM3_two_stage	.641	.420	.558
BM25RM3_single_gpt3.5_strict	.618	.397	**.589**
BM25RM3_single_gpt3.5_run	**.651**	**.449**	.520

Nonetheless, a simplified pipeline employing BM25RM3_two_stage to enhance the initial trial retrieval—by factoring in additional relevance signals from the patient's similarity to the trial's title and eligibility sections—can achieve similar effectiveness to the more complicated GPT-3.5-based systems.

4 Conclusions

In conclusion, utilizing LLMs, particularly the highly capable GPT-3.5, has shown its potential to facilitate the eligibility screening process for clinical trials. The findings from our experiments reveal that the retrieval performance achieved by BM25 models closely aligns with that of GPT-3.5, suggesting that while advanced LLMs introduce innovative approaches to this task, established methods remain robust and competitive. Also, a notable aspect identified in our study is the influence of prompt variations on the performance outcomes of LLMs. This sensitivity to prompt structure underscores the necessity for careful and precise prompt engineering to ensure the reliability and consistency of results obtained through such models in clinical trial retrieval.

5 Acknowledgments

This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020-EU.1.3.1., ID: 860721).

References

1. Georgios Peikos, Oscar Espitia, and Gabriella Pasi. Unimib at trec 2021 clinical trials track, 2022.
2. Craig Macdonald and Nicola Tonellotto. Declarative experimentation ininformation retrieval using pyterrier. In *Proceedings of ICTIR 2020*, 2020.