*Proceedings of Machine Learning Research* **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

# Clinical Trial Recommendation with LLM-Based Query Generation and Graph-Based Pairwise Re-ranking

Mehrnaz Senobari Vayghan[†], Emad A.Mohammed[‡,*], Behrouz H. Far[†]
[†] University of Calgary
[‡] Wilfrid Laurier University

**Abstract**

Clinical trials are essential for drug development and advancing medical treatments. However, many fail due to the challenges of patient recruitment, as identifying suitable participants is both expensive and time-consuming. Recent advances in large language models have demonstrated strong potential in healthcare settings, offering a promising way to automate this process. In this study, we propose an LLM-based recommendation pipeline that suggests a ranked list of clinical trials based on patient characteristics. In the first stage, a medical LLM generates focused search queries from patient notes via chain-of-thought prompting. These queries are used to retrieve a candidate set from a large-scale clinical trial corpus via dense semantic search. In the second stage, candidates are then re-ranked via pairwise re-ranking with graph aggregation. We evaluate our pipeline on the TREC Clinical Trials 2021 and 2022 benchmarks. Query-generated retrieval, achieves significant improvements over raw retrieval, with Recall@1000 improving by 33.8% and 46.1% on TREC 2021 and 2022. Pairwise re-ranking with graph aggregation further improves nDCG@10 by 12.8% and 11.9%, P@10 by 11.2% and 10.1%, and MRR by 18.2% and 12.3% on TREC 2021 and 2022 respectively. All results are obtained only by using an open-source 8B-parameter model, without task-specific fine-tuning or closed-source API dependence.

**Keywords:** Clinical trial recommendation, pairwise re-ranking, large language models, information retrieval, graph-based aggregation, medical NLP

## 1. Introduction

Clinical trials are critical for the discovery and validation of new medical treatments, yet recruitment often falls short of planned targets, leaving trials delayed, underpowered, or abandoned [1, 2]. A key contributor is the difficulty of identifying eligible patients at scale, as manual screening cannot be performed systematically across thousands of active trials [3].

Early automated approaches relied on rule-based systems such as ERGO [4] and EliXR [5], which were difficult to scale and sensitive to clinical language variation. Early machine learning methods, such as Criteria2Query [6] reduced manual effort but remained limited by their dependence on task-specific labeled data. The introduction of pre-trained biomedical language models then marked a more substantial shift in capability. BioBERT [7] and ClinicalBERT [8] established that domain-specific pre-training substantially benefits biomedical NLP tasks and became the standard backbone for end-to-end patient–trial matching. Both DeepEnroll [9] and COMPOSE [10] leveraged ClinicalBERT to encode Clinical trial criteria. DeepEnroll jointly embedded patient records and trial criteria in the same latent space, aligning them via attentive inference. COMPOSE introduced a pseudo-Siamese architecture with a composite loss function for cross-modal patient-trial matching. Despite these advances, prior approaches were constrained by their dependence on task-specific training data and their limited coverage of the diverse medical terminology and linguistic variation found in clinical trial criteria and patient records.

* Corresponding author email: emohammed@wlu.ca

Large language models have demonstrated enormous potential across a wide range of biomedical applications, from information retrieval and clinical decision support to question answering [11]. The emergence of large language models brought richer medical knowledge and zero-shot generalization to clinical trial recommendation. TrialGPT [12] combines LLM-driven retrieval with GPT-4-based criterion-level matching, achieving expert-level accuracy and reducing clinician screening time by over 40% in a prospective user study. Zero-shot GPT-4 has also been shown to match or exceed the best supervised systems without fine-tuning [13]. Despite these advances, these systems rely on closed-source models, raising serious concerns around cost, reproducibility, and data privacy for clinical deployment. Efforts to fine-tune LLaMA models via knowledge distillation from GPT-4-generated annotations [14] achieve competitive performance, yet this approach defers rather than eliminates the dependency on closed-source models, and risks inheriting systematic biases from the teacher model. OncoLLM [15] is a fine-tuned model that achieved strong performance by converting trial criteria into yes/no questions to assess patient eligibility, though its dependence on well-curated oncology EHR data, which is not broadly accessible and demands considerable effort to prepare for use, remains a key limitation.

To address these limitations, we propose a clinical trial recommendation pipeline built entirely on a lightweight, open-source 8B-parameter clinical LLM in a zero-shot setting. We evaluate our methodology on TREC Clinical Trials Track 2021 and 2022 [16, 17], to show strong performance without any task-specific fine-tuning or proprietary model dependence.

## 2. Methodology

### 2.1. Query Generation

Given a patient note, Med42-v2-8B [18], a lightweight open-source clinical LLM built on LLaMA-3, is prompted with a chain-of-thought instruction that guides it to reason about the patient's overall clinical situation and generate focused search queries. The model is explicitly instructed not to introduce any patient characteristics absent from the original note, though it may reformulate clinical information into more general or standardized medical language. Following this reasoning, the model generates a variable number of focused search queries covering different aspects of the patient's case, yielding an average of 9.4 and 8.8 queries per patient on TREC 2021 and 2022, respectively.

### 2.2. Dense Retrieval and Multi-Query Fusion

Each generated query is encoded using MedCPT [19], a biomedical dense retriever, and matched against an offline pre-encoded index of the clinical trial corpus. For each query, the top-$K$ candidate trials are retrieved based on semantic similarity. The per-query ranked lists are then aggregated using Reciprocal Rank Fusion (RRF):

$$\text{Score}(d) = \sum_{i=1}^{n} \frac{1}{k + \text{rank}_{q_i}(d)} \tag{2.1}$$

where $\text{rank}_{q_i}(d)$ is the rank of trial $d$ in the results for query $q_i$ and $k$ is a smoothing constant.

### 2.3. Pairwise Re-ranking with Graph Aggregation

While designing end-to-end retrieval systems, there is an inherent tradeoff between effectiveness and efficiency. Since applying an LLM directly to the full trial corpus is computationally infeasible, dense retrieval serves as an efficient first stage to narrow the candidate set. Ranking by embedding similarity alone, however, does not yield a sufficiently precise ordering. We therefore apply pairwise re-ranking with graph aggregation, adapted from

PRP-Graph [20], to the top 100 retrieved candidates. Pairwise re-ranking compares two candidate trials against each other given the patient note. This relative formulation allows the model to focus on the differences between the two candidates instead of producing an individual score for each trial. The cost of each comparison is kept low by reading the preference between the tokens "A" and "B" directly from the model's next-token log-probabilities. However, an exhaustive pairwise procedure requires $\binom{N}{2}$ comparisons over the candidate set, which grows rapidly with $N$. We therefore organize the comparisons as a Swiss-system tournament, which reduces the total number of comparisons required.

#### 2.3.1. **Pairwise Scoring**

For each pair of candidate trials $(d_i, d_j)$, we present the patient note alongside both trials' information (title, conditions, summary, inclusion and exclusion criteria) and ask Med42-v2-8B which trial is a better match. We extract preference scores from the model's output logits rather than the generated text. Specifically, the log-probabilities of tokens "A" and "B" are used to compute a calibrated preference via softmax:

$$s(d_i \succ d_j) = \frac{\exp(\log p_A)}{\exp(\log p_A) + \exp(\log p_B)} \tag{2.2}$$

To mitigate position bias, each pair is evaluated in both orderings, yielding forward and reverse scores used as separate directed edges in graph construction.

#### 2.3.2. **Swiss-System Tournament**

Instead of comparing all $\binom{N}{2}$ pairs of the $N$ candidates, we arrange the comparisons as a Swiss-system tournament over $R$ rounds. In each round, the candidates are sorted by their current score, and each candidate is paired with the nearest opponent it has not yet faced. After each comparison, both candidates' scores are updated using the model's preference and the opponent's current score, so that comparisons against higher-ranked opponents count more than comparisons against lower-ranked ones. With $N = 100$ and $R = 10$, this gives about 500 pairwise comparisons (1,000 bidirectional LLM calls), far fewer than the $\binom{100}{2} = 4{,}950$ pairs needed for a full all-pairs comparison.

#### 2.3.3. **Graph Aggregation.**

Each comparison produces two directed edges between the compared trials, one in each direction, weighted by the corresponding preference scores. After all rounds, these edges form a directed weighted graph over the candidates. We then aggregate the edge weights across the graph to compute a global relevance score for each candidate. The final score is a linear interpolation of the normalized graph-based and retrieval scores:

$$\hat{s}(c) = \lambda \cdot \widetilde{S}_{\text{rerank}}(c) + (1 - \lambda) \cdot \widetilde{S}_{\text{retr}}(c) \tag{2.3}$$

where $\widetilde{S}_{\text{rerank}}(c)$ and $\widetilde{S}_{\text{retr}}(c)$ are the normalized re-ranking and retrieval scores for candidate trial $c$, and $\lambda$ controls the relative contribution of each.

## 3. **Experimental Setup**

Our framework is evaluated on the TREC Clinical Trials 2021 and 2022 benchmarks, two widely used collections for clinical trial retrieval. Both benchmarks use synthetic patient topics created by individuals with medical training, simulating admission notes in the form of free-text patient case descriptions. TREC 2021 consists of 75 patient topics with 26,162 physician-judged patient–trial pairs, and TREC 2022 consists of 50 topics with 26,585 judged pairs. We adopt standard information retrieval metrics for evaluation. At the retrieval

stage, we report Recall@$K$ for $K \in \{500, 1000, 1500, 2000\}$ to assess how well the candidate pool retains relevant trials across different retrieval depths. At the re-ranking stage, we report nDCG@5 and nDCG@10 for ranking quality, P@5 and P@10 for precision, and Mean Reciprocal Rank (MRR) to measure the position of the first relevant trial in the ranked list.

## 4. **Results and Discussion**

Our results show consistent improvements at each stage of the pipeline across both benchmarks, demonstrating that query generation and graph-based re-ranking each contribute meaningfully to the final recommendation quality.

Figure 1 illustrates that query generation consistently improves retrieval coverage across all depths. With a small candidate pool per patient, the approach recovers above 80% of relevant trials on both datasets, ensuring that the re-ranking stage has access to the majority of relevant trials.

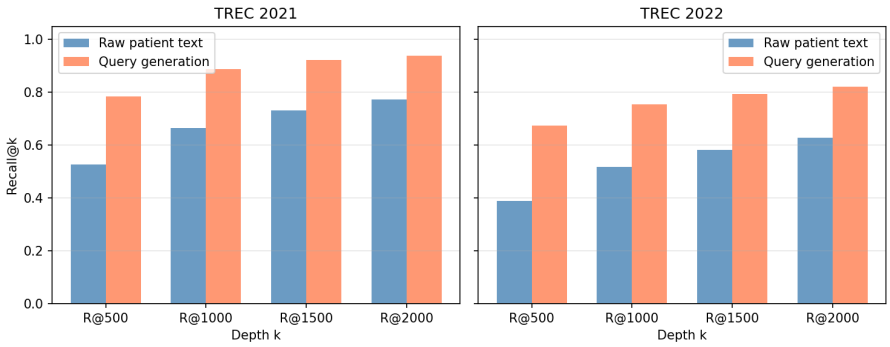


*Figure 1.* Retrieval recall at various depths on TREC 2021 and TREC 2022 (MedCPT retriever). Query generation consistently outperforms raw patient text across all depths.

The end-to-end results are presented in Table 1. Each stage of the pipeline contributes to the overall performance. Query generation improves all ranking metrics over the raw retrieval baseline, and the re-ranking stage provides additional gains across all metrics on both benchmarks. The nDCG improvements are larger at shallower depths, suggesting that the re-ranking stage is particularly effective at placing the most relevant trials at the very top of the ranked list. The high MRR on both benchmarks indicates that the first relevant trial consistently appears within the first few positions, which is important in practice as clinicians are unlikely to review beyond the top results. Precision gains at both P@5 and P@10 confirm that the quality of the ranked list is maintained throughout the top positions, not only at the very top.

| Dataset | Method | nDCG@5 | nDCG@10 | P@5 | P@10 | MRR |
|---|---|---|---|---|---|---|
| TREC 2021 | Raw Retrieval | 0.497 | 0.497 | 0.365 | 0.332 | 0.648 |
| | Query-Generated Retrieval | 0.654 | 0.639 | 0.654 | 0.633 | 0.768 |
| | + PRP-Graph Re-ranking | **0.743** | **0.721** | **0.732** | **0.704** | **0.908** |
| TREC 2022 | Raw Retrieval | 0.383 | 0.335 | 0.180 | 0.147 | 0.463 |
| | Query-Generated Retrieval | 0.598 | 0.581 | 0.592 | 0.573 | 0.765 |
| | + PRP-Graph Re-ranking | **0.678** | **0.650** | **0.666** | **0.631** | **0.859** |

*Table 1.* Evaluation result on TREC 2021 and TREC 2022 (75 and 50 patients respectively).

## 5. Conclusion and Future Work

This work shows that query generation and graph-based pairwise re-ranking, using a lightweight open-source LLM, each contribute to improved clinical trial recommendation. Results are consistent across both TREC 2021 and 2022 benchmarks, demonstrating that the pipeline generalizes well across different patient populations and trial collections. Future directions include improving the explainability of recommendations to support clinician trust and evaluating the pipeline on real EHR data to better assess its practical value in clinical settings.

### References


[1] M. Briel, K. K. Olu, E. Von Elm, B. Kasenda, R. Alturki, A. Agarwal, N. Bhatnagar, and S. Schandelmaier. "A systematic review of discontinued trials suggested that most reasons for recruitment failure were preventable". In: *Journal of clinical epidemiology* 80 (2016), pp. 8–15.

[2] D. B. Fogel. "Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review". In: *Contemporary clinical trials communications* 11 (2018), pp. 156–164.

[3] S. R. Thadani, C. Weng, J. T. Bigger, J. F. Ennever, and D. Wajngurt. "Electronic screening improves efficiency in clinical trial recruitment". In: *Journal of the American Medical Informatics Association* 16.6 (2009), pp. 869–873.

[4] S. W. Tu, M. Peleg, S. Carini, M. Bobak, J. Ross, D. Rubin, and I. Sim. "A practical method for transforming free-text eligibility criteria into computable criteria". In: *Journal of biomedical informatics* 44.2 (2011), pp. 239–250.

[5] C. Weng, X. Wu, Z. Luo, M. R. Boland, D. Theodoratos, and S. B. Johnson. "EliXR: an approach to eligibility criteria extraction and representation". In: *Journal of the American Medical Informatics Association* 18.Supplement 1 (2011), pp. i116–i124.

[6] C. Yuan, P. B. Ryan, C. Ta, Y. Guo, Z. Li, J. Hardin, R. Makadia, P. Jin, N. Shang, T. Kang, et al. "Criteria2Query: a natural language interface to clinical databases for cohort definition". In: *Journal of the American Medical Informatics Association* 26.4 (2019), pp. 294–305.

[7] 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". In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[8] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott. "Publicly available clinical BERT embeddings". In: *Proceedings of the 2nd clinical natural language processing workshop*. 2019, pp. 72–78.

[9] X. Zhang, C. Xiao, L. M. Glass, and J. Sun. "DeepEnroll: patient-trial matching with deep embedding and entailment prediction". In: *Proceedings of the web conference 2020*. 2020, pp. 1029–1037.

[10] J. Gao, C. Xiao, L. M. Glass, and J. Sun. "COMPOSE: Cross-modal pseudo-siamese network for patient trial matching". In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020, pp. 803–812.

[11] S. Tian, Q. Jin, L. Yeganova, P.-T. Lai, Q. Zhu, X. Chen, Y. Yang, Q. Chen, W. Kim, D. C. Comeau, et al. "Opportunities and challenges for ChatGPT and large language models in biomedicine and health". In: *Briefings in Bioinformatics* 25.1 (2024), bbad493.

[12] Q. Jin, Z. Wang, C. S. Floudas, F. Chen, C. Gong, D. Bracken-Clarke, E. Xue, Y. Yang, J. Sun, and Z. Lu. "Matching patients to clinical trials with large language models". In: *Nature communications* 15.1 (2024), p. 9074.

[13] M. Wornow, A. Lozano, D. Dash, J. Jindal, K. W. Mahaffey, and N. H. Shah. "Zero-shot clinical trial patient matching with LLMs". In: *NEJM AI* 2.1 (2025), AIcs2400360.

[14] M. Nievas, A. Basu, Y. Wang, and H. Singh. "Distilling large language models for matching patients to clinical trials". In: *Journal of the American Medical Informatics Association* 31.9 (2024), pp. 1953–1963.

[15] S. Gupta, A. Basu, M. Nievas, J. Thomas, N. Wolfrath, A. Ramamurthi, B. Taylor, A. N. Kothari, R. Schwind, T. M. Miller, et al. "PRISM: Patient Records Interpretation for Semantic

clinical trial Matching system using large language models". In: *NPJ digital medicine* 7.1 (2024), p. 305.

[16] K. Roberts, D. Demner-Fushman, E. M. Voorhees, S. Bedrick, and W. R. Hersh. "Overview of the TREC 2021 Clinical Trials Track". In: *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021)*. National Institute of Standards and Technology, 2021.

[17] K. Roberts, D. Demner-Fushman, E. M. Voorhees, S. Bedrick, and W. R. Hersh. "Overview of the TREC 2022 Clinical Trials Track". In: *Proceedings of the Thirty-First Text REtrieval Conference (TREC 2022)*. National Institute of Standards and Technology, 2022.

[18] C. Christophe, P. K. Kanithi, T. Raha, S. Khan, and M. A. Pimentel. "Med42-v2: A suite of clinical llms". In: *arXiv preprint arXiv:2408.06142* (2024).

[19] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, and Z. Lu. "MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval". In: *Bioinformatics* 39.11 (2023), btad651.

[20] J. Luo, X. Chen, B. He, and L. Sun. "Prp-graph: Pairwise ranking prompting to llms with graph aggregation for effective text re-ranking". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 5766–5776.