Streamlining Clinical Trial Recruitment: A Two-Stage Zero-Shot LLM Approach with Advanced Prompting

Mozhgan Saeidi MOZHGANS@STANFORD.EDU
Stanford University

Abstract

Identifying patient eligibility for clinical trials is a critical bottleneck hindering medical research progress because many clinical trials allow only small, specific patient cohorts to be included and require a certain number of participating patients to yield definitive results. Manually screening patients through unstructured medical records is time-consuming and expensive. This paper explores the potential of large language models (LLMs) enhanced with medical context to automate patient eligibility assessments for clinical trials. We first design a two-stage zero-shot LLM approach to analyze a patient's medical history (presented as unstructured text) and to determine their eligibility for a specific trial. We use advanced prompting strategies to guide the LLM toward faster and more targeted matches between trials and eligible patients. Additionally, a two-stage retrieval pipeline pre-filters potential trials using efficient retrieval techniques, reducing the number of trials considered for each patient. This two-way matching substantially improves processing speed and cost-effectiveness for clinical trial recruitment. Our method holds promise for streamlining clinical trial patient recruitment to accelerate medical advances.

Keywords: Clinical Trial Matching, Zero-Shot Learning, Advanced Prompting Techniques, Two-Stage Retrieval Pipeline, Patient Categorization, Transformer Embeddings, Relevance Scoring Functions, Pre-trained Models vs. LLMs, Balanced Relevance Ranking

Data and Code Availability We have two sources of data; one is the trials and clinical notes from the NIH [1] and the second is the 2018 n2c2 Clinical Trial Cohort Selection dataset, public benchmarks for patient matching Stubbs et al. (2019). The first covers diseases including glaucoma, anxiety, chronic obstructive pulmonary disease (COPD), breast cancer, COVID-19, rheumatoid arthritis, sickle cell anemia, and type 2 diabetes, with five patients' per disease. The patients' information (questioners) is available in XML format (Fig 1 and 2). For the trials, the May 2023 snapshot of `ClinicalTrials.gov` is used as the corpus, which includes 52,130 clinical trials and is provided as five single zip files. The patients' information (the questionnaires) is matched against these trials.

The second dataset has two parts: patients' information and eligibility criteria. There are 288 diabetic patients, split into a training group (70%) and a test group (30%). Each patient has several anonymized clinical notes (with an average of 2,711 words). The data comes from a combined database across two hospitals. As eligibility criteria, thirteen common criteria were used to determine if a patient would be eligible for a clinical trial Stubbs et al. (2019); Stubbs and Uzuner (2015); Kumar et al. (2015). There are no exclusion criteria in this simulated trial. These eligibility criteria are mentioned in Table 4.1.1. For labels, two doctors reviewed each patient's records and labeled them as "meeting" or "not meeting" each criterion. Our code is available on Github.

1. Introduction

Clinical trials are the engine driving the development of new and effective medical treatments and interventions Hardman et al. (2023). However, this engine often faces bottlenecks due to the critical roadblock of patient recruitment. Traditional methods, such as relying on specialists or sifting through electronic health records (EHRs), are slow and inefficient, leading to delays and even cancellations of important trials McMillan (2024). Fortunately, the landscape is changing. Patients are taking a more proactive approach to their health, actively seeking out clinical trials through online resources. In most cases,

[1] https://www.nia.nih.gov/health/clinical-trials-and-studies

participants struggle with the challenge of extracting relevant information from the complex and lengthy descriptions on `ClinicalTrials.gov`, a US registry of clinical trials. Finding the right fit is a demanding task for patients without medical training, as key details regarding eligibility often lie buried within intricate inclusion and exclusion criteria Patino and Ferreira (2018).

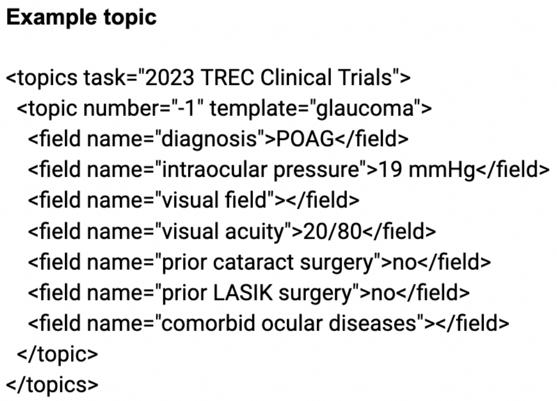
The current state of patient recruitment presents a challenge. On the one hand, participation in clinical trials offers patients undeniable benefits: access to cutting-edge therapies, closer monitoring by expert teams, and potentially improved health outcomes. Yet, a staggering 94% of patients remain unaware of trials they might qualify for, simply because their doctors have not informed them Wilson (2023); Gogtay et al. (2020). This knowledge gap stems from the highly manual and time-consuming nature of identifying eligible patients Ostropolets et al. (2020). Each trial has a stringent set of criteria that every participant must meet. Traditionally, trained coordinators spend hours meticulously combing through hundreds of patients' electronic health records (EHRs), which themselves are often a labyrinth of unstructured text, including progress notes, emails, radiology reports, and genetic data Honavar (2020). This unstructured format makes it extremely difficult to automate the process. As a result, screening a single patient for a late-stage cancer trial can take nearly an hour Ismail et al. (2023).

By leveraging the power of large language models (LLMs), we hope to streamline this arduous process Singhal et al. (2023). These sophisticated AI models, trained on massive datasets of text, hold the potential to analyze a patient's medical history from unstructured medical records and efficiently determine their eligibility for specific trials Papachristou et al. (2023). This approach has the potential to automate clinical trial recruitment, paving the way for faster medical progress and empowering patients to take a more active role in their health journey Chowdhury et al. (2024). Our proposed system goes beyond a simple yes/no for patient eligibility in clinical trials. We categorize patients into three distinct groups; 1)**Eligible**: Patients who meet all inclusion criteria and have no exclusion criteria. 2)**Excluded**: Patients who explicitly violate any exclusion criteria. 3)**Not Relevant**: Patients for whom there is not enough information in their history to definitively determine eligibility (neither exactly included nor excluded). This distinction separates those lacking sufficient data from those definitively disqualified.

To efficiently rank relevant clinical trials for a specific condition, we leverage the inclusion and exclusion criteria in a two-stage process; 1)**Filtering**: We use the criteria to filter out trials that clearly would not be a good fit based on exclusions or lack of relevant patient information. This substantially reduces the number of trials for further evaluation. 2)**Weighted Ranking**: For the remaining trials, we use transformer embeddings to calculate a *relevance score*. This score considers the importance of each inclusion and exclusion criteria, leading to a more nuanced ranking of potentially suitable trials Greco and Tagarelli (2023).

2. Related Work

Earlier attempts at automated clinical trial matching focused on transforming eligibility criteria written in natural language into structured queries that may be used to search electronic health records (EHRs) Meystre et al. (2023); Saeidi et al. (2023a). These systems, like EliXR Xu et al. (2023b) and EliIE Kang et al. (2017), rely on rule-based algorithms Margot and Luta (2021) or more advanced natural language processing (NLP) techniques like named entity recognition (NER) Nadeau and Sekine (2007). However, these systems are difficult to maintain and adapt to different trials. The focus then shifted towards "end-to-end" systems that directly match patients to free-text criteria. COMPOSE Gao et al. (2020) and DeepEnroll Zhang et al. (2020) used neural networks to analyze both enrollment criteria and patient records. These approaches only considered structured EHR data, ignoring rich information contained in unstructured clinical notes. TrialGPT Jin et al. (2023) emerged as the first system to leverage large language models (LLMs) for processing unstructured clinical text and eligibility criteria, achieving high accuracy. However, TrialGPT lacked zero-shot learning capabilities and only tested one specific version of an LLM. Recently, the ZeroShot model Wornow et al. (2024) filled this gap by evaluating some recent LLMs in this task. The remaining gap that we fill in is a system that uses LLMs for zero-shot patient matching that i) efficiently selects relevant patient notes before feeding them to the LLM and ii) evaluates the pre-trained language models as the backbone of the matching system when access to high compute machines is limited.



Example topic

```
<topics task="2023 TREC Clinical Trials">
 <topic number="-1" template="glaucoma">
  <field name="diagnosis">POAG</field>
  <field name="intraocular pressure">19 mmHg</field>
  <field name="visual field"></field>
  <field name="visual acuity">20/80</field>
  <field name="prior cataract surgery">no</field>
  <field name="prior LASIK surgery">no</field>
  <field name="comorbid ocular diseases"></field>
 </topic>
</topics>
```

Figure 1: An example of XML topic format

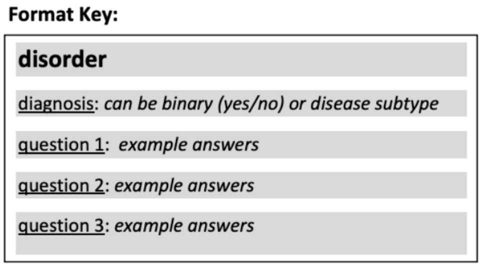
3. Problem Description

We focus on addressing challenges within the clinical trial matching process. We propose a system that aims to overcome current limitations by using a three-tiered approach to categorize patients and a two-stage ranking system for clinical trials. Here is a breakdown of the specific hurdles it tackles:

- Inefficiency: Traditional methods for identifying eligible patients are time-consuming, involving manual review of lengthy medical records.
- Accuracy: Classifying patients based solely on meeting all inclusion criteria might overlook potentially suitable candidates with missing information in their records.
- Limited Information: Current methods might not distinguish between patients who lack data and those explicitly excluded by the trial criteria.

Eligibility Criteria
ABDOMINAL, ADVANCED-CAD, ASP-FOR-MI, ALCOHOL-ABUSE, CREATININE, ENGLISH, DIETSUPP-2MOS , DRUG-ABUSE, HBA1C KETO-1YR, MAJOR-DIABETES, MAKES-DECISIONS, MI-6MOS

Table 1: The thirteen common eligibility criteria. The definition of each criterion is mentioned in Table 4.1.1.



Format Key:

disorder

diagnosis: *can be binary (yes/no) or disease subtype*

question 1: *example answers*

question 2: *example answers*

question 3: *example answers*

Figure 2: Questionnaire Template; the general format of the questionnaire

4. Method

For both datasets, we have a set of patients' notes, and a set of clinical trials. We map patients to the trials based on the inclusion and exclusion criteria, where applicable, and then provide a ranking of the trials based on their relevance to the patient's history. We compare this matching using the pre-trained LLM models of BERT Devlin et al. (2018), SBERT Reimers and Gurevych (2019), Clinical BERT (CBERT) Alsentzer et al. (2019); Saeidi et al. (2021a), as well as the popular open-source models GPT4 Achiam et al. (2023) and Llama-2-70b-32k Xu et al. (2023a).

BERT (Bidirectional Encoder Representations from Transformers) and SBERT (Sentence-BERT) are powerful pre-trained language models (PLMs) for NLP. BERT excels at understanding word context within a sentence using a masked language modeling objective. SBERT, a variant of BERT, focuses on sentence-level semantics by fine-tuning sentence similarity tasks. While BERT grasps word-level relationships, SBERT captures sentence-level meaning, making it ideal for applications requiring sentence comparison Saeidi et al. (2022); Saeidi. For clinical text, CBERT, a domain-specific BERT model, demonstrates superior performance compared to general-purpose embeddings. This highlights the importance of tailoring models to specific domains Alsentzer et al. (2019); Saeidi et al. (2021a).

When working with LLMs, prompts act as clear and concise questions that guide the LLM to sift through the unstructured data to identify the keywords and phrases to solve our matching task Alavi Naeini et al. (2024); Saeidi et al. (2021b). The prompt would instruct the LLM to analyze a patient's medical history to determine whether they meet the eligibility criteria for a particular clinical trial Saeidi et al. (2021a). In addition, the prompts plays a vital role by providing all of the necessary context for the LLM to understand the specific inclusion and exclusion criteria for a particular trial without prior training Saeidi et al. (2023b).

The model starts with a two-stage process. In the first stage, we identify all clinical trials that include the disorder. Then, we extract the negated and non-negated inclusion and exclusion criteria using negspaCy Pizarro (2023); Saeidi et al. (2021c) from the eligibility criteria column and convert them into UMLS Concept IDs Bodenreider (2004) using ScispaCy Neumann et al. (2019). By identifying both negated and non-negated criteria, the system captures the complete picture of the requirements for a clinical trial. For example, consider a clinical trial looking for participants with a specific disease but excluding pregnant women. The system would need to identify both the positive criterion ("Must have the disease") and the negated criterion ("NOT pregnant") to accurately determine a patient's eligibility.

We did not directly compare the text descriptions to identify potentially relevant clinical trials for a specific condition. Instead, we first do concept mapping, converting the medical terms in patient profiles and clinical trial descriptions into a standardized format using UMLS concept IDs. This ensures a consistent matching of the terminology across the system when the actual words used may differ. Second, we do similarity scoring, which is searching for matching Concept IDs between patient profiles and trial descriptions. Trials with more matching concepts were considered more relevant to the patient's condition. To rank these trials, we used a score based on the natural logarithm of the trial's position in the ranked list, compared to the total number of trials considered. This approach prioritizes trials with a higher number of concept matches. The outcome of this process is a ranked list of candidate clinical trials likely to be a good fit for the patient's specific situation.

In the second stage, we focus on the patient's perspective. Instead of wading through mountains of text descriptions, we use a smarter approach to find clinical trials that might be a good match. We translate the medical terms from both their records and the trial descriptions into a universal language doctors understand (UMLS Concept IDs). By unifying medical terminology through concept IDs, the UMLS promotes better communication and data exchange in healthcare, ultimately leading to more effective and interoperable information systems.

This ensures that the complex medical jargon and institution-specific terms are avoided in favor of a common language. Then, we find trials with concepts that are closely aligned with the patient's medical history. The trials with the best matches are considered the most relevant, and we prioritize them accordingly. This way, we get a ranked shortlist of trials that are likely to be a good fit for the specific needs of the patient.

We categorize the specialty of our model as follows:

- Standardized Language: We convert medical terms in both patient profiles and trial descriptions into a common language (UMLS Concept IDs). This ensures clear and consistent interpretation of medical terminology across the system.

- Match-Based Ranking: We search for matching concepts between patient profiles and trial descriptions. Trials with a higher number of concept matches are deemed more relevant to the patient's condition. We then rank them using a score that prioritizes trials with more matches.

- Pre-Trained Sentence Embeddings: The system uses pre-trained sentence embeddings, like those generated by a large model like Sentence-BERT, to represent the meaning of key concepts within the inclusion and exclusion criteria of a clinical trial. These pre-trained embeddings would capture semantic relationships between words and concepts, providing a rich understanding of the criteria.

- Prompt Design: The prompts used to guide the LLM to leverage these pre-trained sentence embeddings. For example, the prompt might ask the LLM to compare the sentence embeddings from the patient's medical notes with the pre-trained embeddings representing key concepts within the inclusion criteria. This comparison would allow the LLM to identify potential matches between the patient's medical history and the trial's criteria.

- Fine-Tuning a Sentence Transformer: The system fine-tunes a pre-trained sentence transformer model on a limited dataset of clinical trial criteria and relevant medical notes. This allows the model to adapt to the specific language used in the clinical trial domain, potentially improving performance in a few-shot setting.

- Prompt Engineering: Sophisticated prompt engineering techniques could be explored to leverage the LLM's capabilities even further. By crafting prompts that effectively reference pre-trained sentence embeddings or other knowledge sources, the system could potentially achieve good results in a true few-shot scenario.
- Using the SBERT, we loop through every patient and embed all of their notes using our embedding model. We also embed each trial, considering the tokens as sentences.

4.1. Ranking Metrics

We evaluate the ability of pre-trained and the LLM models to make a yes-or-no decision on patient eligibility for clinical trials based on the inclusion and exclusion criteria. The metrics for this evaluation are precision, recall, and overall Macro/Micro-F1 scores. A yes decision indicates if a patient met each of the inclusion criteria in the benchmark. For the trials' ranking, we evaluate the ranking with the following metrics:

4.1.1. Relevance Scoring Functions

To assess how well a patient's medical history aligns with a clinical trial's requirements, we use three different scoring methods. Each method relies on "similarity matrices" that capture how closely related different pieces of text are. We have two matrices; 1)Inclusion Matrix: This shows how similar each sentence extracted from the patient's medical history (topic sentence) is to each sentence within the trial's inclusion criteria. 2)Exclusion Matrix: This shows how similar each topic sentence from the patient's history is to each sentence within the trial's exclusion criteria.

Naive High Precision Score: We use a score named S_{naive} to assess clinical trials. This score is based on a predefined similarity threshold. Let's break down how it is calculated for each trial:

1. To assess how well a clinical trial aligns with the desired criteria, we compare each "must-have" point (inclusion criteria sentence) with the summaries (topic sentences) of the trial. We consider a requirement fulfilled if at least one summary demonstrates a high enough level of similarity (above a threshold called t) to the corresponding requirement.
2. For each requirement (inclusion criteria sentence) that a clinical trial successfully fulfills, we award a score of 1. We store this information in a $I_{satisfied}$ table. This table functions like a binary code, with a 1 indicating a met requirement and a 0 signifying an unmet one.
3. The system also considers factors that would disqualify a trial (exclusion criteria). If there are such factors (represented by the matrix E, which is not empty), we compare all the summaries (topic sentences) of the trial with each disqualification factor. If even one summary demonstrates a high level of similarity (above the threshold t) to any disqualification factor, the trial is excluded (exclusion score, E_{score}, is set to -1). An example of this similarity matrix is visualized in Figure 3.

The score (S_{naive}) for the clinical trial is computed as follows:

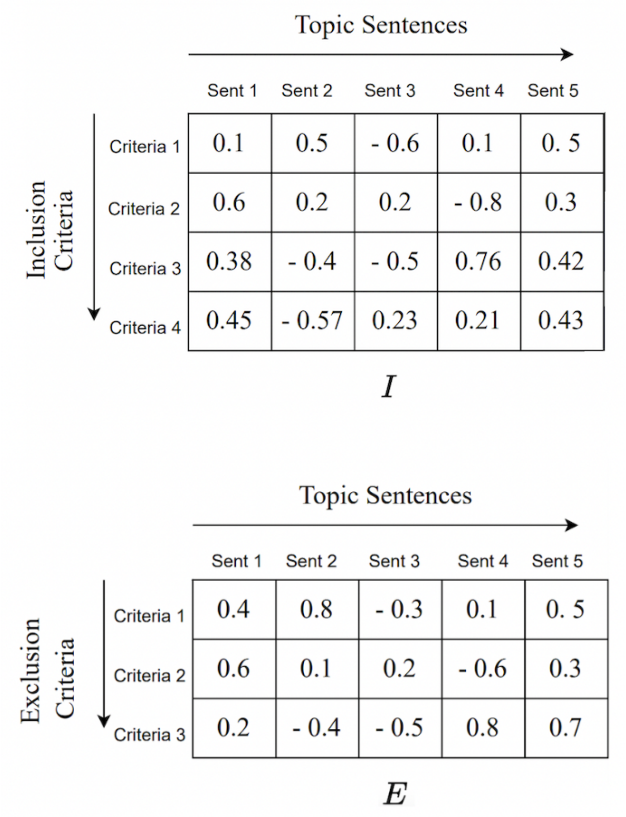
$$S_{naive} = \frac{1}{N} \sum_{i=1}^{N} I_{satisfied_i} + E_{score}$$

Where:

- N is the number of inclusion criteria sentences.
- $I_{satisfied_i}$ is 1 if the i-th inclusion criteria sentence is satisfied, 0 otherwise.
- E_{score} is -1 if any exclusion criteria are satisfied, 0 otherwise.

The score (S_{naive}) is a naive measure of the relevance of the clinical trial, where a negative score or zero score indicates exclusion due to the presence of exclusion criteria, and a positive score indicates inclusion based on the satisfaction of inclusion criteria. We use a threshold of 0.5 to compute this score.

Weighted Relevance Score: The $S_{weighted}$ score assesses clinical trial relevance without a pre-defined similarity threshold. It incorporates the relative importance of inclusion and exclusion criteria. Here is the approach: It calculates the average cosine similarity between topic sentences and each inclusion criteria sentence. It subtracts the maximum cosine similarity value found between any topic sentence and an exclusion criteria sentence. A negative score indicates


Topic Sentences
Inclusion Criteria
I

	Sent 1	Sent 2	Sent 3	Sent 4	Sent 5
Criteria 1	0.1	0.5	- 0.6	0.1	0. 5
Criteria 2	0.6	0.2	0.2	- 0.8	0.3
Criteria 3	0.38	- 0.4	- 0.5	0.76	0.42
Criteria 4	0.45	- 0.57	0.23	0.21	0.43

Topic Sentences
Exclusion Criteria
E

	Sent 1	Sent 2	Sent 3	Sent 4	Sent 5
Criteria 1	0.4	0.8	- 0.3	0.1	0. 5
Criteria 2	0.6	0.1	0.2	- 0.6	0.3
Criteria 3	0.2	- 0.4	- 0.5	0.8	0.7

Figure 3: *Example Cosine Similarity matrices*, We use two cosine similarity matrices to compute ranking score, I and E for {inclusion criteria, topic sentences} similarity and {exclusion criteria, topic sentences} similarity respectively.

that the topic sentences, on average, have a higher similarity to the exclusion criteria compared to the inclusion criteria, suggesting a weaker alignment with the desired characteristics.

1. The inclusion score reflects how well a clinical trial's summaries (topic sentences) align with the desired characteristics (inclusion criteria). We consider the cosine similarity matrix (I) to calculate this score. This matrix represents the similarity between each topic sentence (i) and each inclusion criteria sentence (j). We then compute the average cosine similarity across the entire matrix I. This average value represents the overall inclusion score.

$$I_{Score} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} I_{ij}$$

   where N is the number of inclusion criteria sentences and T is the number of topic sentences.

2. The exclusion score identifies the potential ineligibility criteria for a clinical trial. We calculate this score using the exclusion matrix (E), which stores the cosine similarity between each topic sentence (i) and each exclusion criteria sentence (j). Here, we are interested in the maximum cosine similarity value within the entire matrix E. This maximum value represents the exclusion score and signifies the highest degree of similarity between any summary and a disqualifying factor.

$$E_{Score} = \max_{i}(E_{ij})$$

3. Calculate the final score as the difference between the inclusion score and the exclusion score:

$$S_{weighted} = I_{Score} - E_{Score}$$

The $S_{weighted}$ score is a measure of how well a clinical trial aligns with the desired characteristics. A negative score indicates a stronger alignment between the summaries (topic sentences) and the exclusion criteria (characteristics that disqualify a trial) compared to the inclusion criteria (characteristics a trial must have). This scoring method emphasizes the importance of avoiding trials with disqualifying factors and penalizes them for such matches.

Balanced Relevance Score: The $S_{balanced}$ score offers a balanced assessment of clinical trial relevance. It avoids predefined matching thresholds and considers both inclusion and exclusion criteria with equal weight. This scoring function aims to strike a balance between these two sets of criteria. A negative score suggests the topic sentences, on average, have a higher similarity to the exclusion criteria compared to the inclusion criteria, indicating a potential mismatch with the desired characteristics.

The scoring process can be summarized as follows:

1. Compute the average cosine similarity in the inclusion matrix I, where I_{ij} represents the simi-

Eligibility Criteria	Definition
ABDOMINAL	If the patient has undergone abdominal surgery, including bowel resection or obstruction
ADVANCED-CAD	If the patient has a history of multiple cardiovascular risk factors or events
ASP-FOR-MI	If the patient's medication list includes aspirin for cardiovascular protection
ALCOHOL-ABUSE	If the patient's medical history indicates excessive alcohol consumption
CREATININE	If the patient's recent blood work indicates kidney function impairment
ENGLISH	If the patient speaks English fluently
DRUG-ABUSE	If the patient has a history of drug abuse
DIETSUPP-2MOS	If the patient has taken any dietary supplements (other than Vitamin D) in the past two months
HBA1C	If there is any lab results showing an HbA1c between 6.5% and 9.5%
KETO-1YR	If the patient had ketoacidosis in the past year
MAJOR-DIABETES	If the patient has any serious diabetes-related complications
MAKES-DECISIONS	If the patient is capable of independent decision-making
MI-6MOS	If the patient experienced a myocardial infarction within the past six months

Table 2: Description of the thirteen common eligibility criteria.

larity between topic sentence i and inclusion criteria sentence j. The inclusion score is given by:

$$I_{Score} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} I_{ij}$$

where N is the number of inclusion criteria sentences and T is the number of topic sentences.

2. Compute the average cosine similarity in the exclusion matrix E, where E_{ij} represents the similarity between topic sentence i and exclusion criteria sentence j. The exclusion score is given by:

$$E_{Score} = \frac{1}{KT} \sum_{i=1}^{K} \sum_{j=1}^{T} E_{ij}$$

We use K to denote the number of sentences outlining the factors that would disqualify a clinical trial (exclusion criteria). T represents the total number of sentences summarizing the key points of the trial (topic sentences).

3. The overall fit of a clinical trial is captured in the combined score. This score is calculated by taking the difference between the inclusion score (reflecting how well the trial matches the desired characteristics) and the exclusion score (indicating the degree of alignment with disqualifying factors).

$$S_{balanced} = I_{Score} - E_{Score}$$

The $S_{balanced}$ score offers a balanced perspective on a clinical trial's relevance by considering both inclusion and exclusion criteria with equal weight. This scoring function avoids pre-set thresholds for matching similarities. A negative score indicates the topic sentences, on average, have a higher similarity to the exclusion criteria compared to the inclusion criteria. This suggests a potential mismatch with the desired characteristics for the trial. Conversely, a positive score signifies a stronger alignment between the summaries and the inclusion criteria. This scoring method aims to provide a more nuanced assessment of how well a clinical trial aligns with the overall requirements, allowing for a more informed decision about its suitability.

Ideally, we would not want any matches with exclusion criteria in a clinical trial. However, the metrics that consider exclusion criteria serve a valuable purpose despite this seemingly contradictory aim.

5. Evaluation and Results

This work focuses on matching patients (topics) with suitable clinical trials. We identify a maximum of 100 trials for each patient, prioritizing those with the most relevant criteria. We evaluated four different methods (runs) to refine the initial trial suggestions generated in stage 1. Three of these methods involved re-ranking the candidate trials based on their fit for each patient. The first run simply presented the unordered list of trials from stage 1.

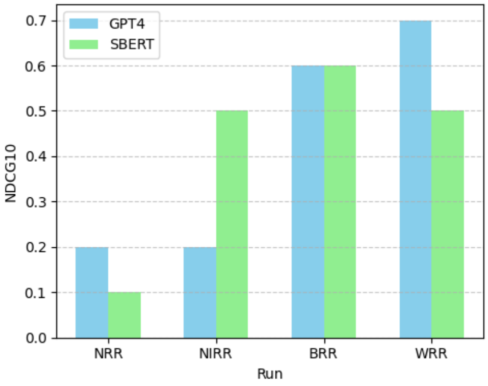


Figure 4: Distribution of NDCG Scores Across Topics. Our re-ranking methods significantly improved performance across all metrics, with "balanced relevance ranking" and "weighted relevance ranking" consistently outperforming others.

The outcomes of our experiments are presented in a concise format within Tables 3, and 4. The first and second tables showcase the employed evaluation metrics: Precision, recall, and overall Macro/Micro-F1 scores. To assess the variability in model performance, we ran each model 100 times and calculated the standard deviation (std) for each metric (precision, recall, F1 scores) using the population standard deviation formula.

Our re-ranking approaches were successful. Three of the four methods we tested significantly improved the results compared to the average performance across all the runs. This improvement applies to all three evaluation metrics: Precision at 10 (P-10), Reciprocal Rank, and Mean Average Precision (MAP). The most effective methods were “balanced relevance ranking” and “weighted relevance ranking”, they consistently outperformed all other runs on all metrics. We also evaluated the Normalized Discounted Cumulative Gain (NDCG) scores, which provide another measure of ranking quality. Here, the “balanced relevance ranking” strategy achieved the best results. Importantly, all our re-ranking methods improved the NDCG scores compared to the initial ranking (NRR output). Notably, both “naive relevance scoring” and “weighted relevance scoring” significantly improved over the median scores. The distribution of the evaluation metrics and NDCG scores across topics is presented in Figure 4.

Our next evaluation in continuing this study is on bias in the final results. We monitored bias in the ranked list output of the models by considering the top 50 ranked trials of GPT4 and SBERT to evaluate if they are biased toward a certain group or gender.

6. Conclusion and Discussion

Our approach to the clinical trial matching problem utilizes a two-stage process that leverages prompting and few-shot learning techniques to rank and re-rank clinical trials for a specific disorder effectively. Stage one is Initial Candidate Identification. When we use prompting for relevant trials, we don't simply search for keywords, but to guide the LLM to identify all relevant clinical trials. This involves crafting prompts that essentially ask the LLM to find trials where the “conditions” column mentions the target disorder. Once we have a list of potential trials, we need to extract the eligibility criteria. Here, we utilize a few-shot learning approach and provide the LLM with a few annotated examples of how to extract inclusion and exclusion criteria from eligibility text, allowing it to learn the patterns and generalize to unseen trials. This reduces the need for extensive manual annotation.

Stage two is Weighted Relevance Ranking. After identifying candidate trials, we rank them based on their relevance to the specific patient's needs. This stage involves two sub-steps; 2-1) Standardization and Matching: We first standardize medical terms within the extracted inclusion/exclusion criteria using UMLS Concept IDs. This creates a common language for comparison. We then compare these standardized terms to the UMLS Concept IDs found in the given patient's topics, generating a shortlist of highly relevant candidate trials. 2-2) Weighted Relevance Scoring: We calculate a weighted relevance score for each shortlisted trial. This score reflects the cosine similarity between contextual embeddings of the patient's topic sentences and the corresponding inclusion/exclusion criteria. Importantly, we consider three different weighting strategies for inclusion and exclusion criteria, acknowledging their varying importance in determining patient eligibility. Our proposed method achieves superior performance compared to the median scores in terms of Precision@10 and NDCG-10 metrics. Notably, “balanced relevance ranking”, a strategy that assigns balanced weights to inclusion and exclusion criteria, outperforms other approaches based on median metrics. This suggests

Model	P±std	R±std	MF1±std	mF1±std
GPT4	0.89±0.12	0.74±0.21	0.75±0.14	0.85±0.09
Llama	0.85±0.23	0.71±0.21	0.73±0.15	0.83±0.24
BERT	0.82±0.22	0.70±0.16	0.71±0.19	0.82±0.19
SBERT	0.85±0.31	0.73±0.25	0.74±0.26	0.83±0.18
CBERT	0.83±0.23	0.72±0.22	0.73±0.17	0.82±0.16

Table 3: Zero-shot results on the combined dataset with evaluation metrics Precision (P), Recall (R), Macro-F1 (MF1), and Micro-F1 (mF1) of the comparison models, including LLM of GPT4, Llama-2-70b-32k and pre-trained models of BERT, SBERT, and CBERT, before prompting. We calculated the standard deviation (std) to show the central tendency (average performance) of each model on each metric (precision, recall, and F1 scores) along with the variability around that central tendency.

Model	P±std	R±std	MF1±std	mF1±std
GPT4	0.92±0.22	0.93±0.28	0.83±0.23	0.94±0.25
Llama	0.89±0.18	0.91±0.21	0.81±0.22	0.92±0.18
BERT	0.87±0.16	0.88±0.17	0.80±0.11	0.89±0.18
SBERT	0.89±0.22	0.90±0.23	0.82±0.24	0.91±0.15
CBERT	0.88±0.17	0.89±0.21	0.81±0.28	0.90±0.11

Table 4: Precision (P), Recall (R), Macro-F1 (MF1), and Micro-F1 (mF1) of the comparison models, including LLM of GPT4, Llama-2-70b-32k and pre-trained models of BERT, SBERT, and CBERT, after prompting. Standard deviation (std) was calculated for each metric (precision, recall, F1 scores) to analyze both average performance and the degree of variation across the models.

that finding an appropriate balance when considering both inclusion and exclusion factors is crucial for accurate clinical trial matching.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Saeid Alavi Naeini, Raeid Saqur, Mozhgan Saeidi, John Giorgi, and Babak Taati. Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Ahmadul Karim Chowdhury, Md Saidur Rahman Sujon, Md Shirajus Salekin Shafi, Tasin Ahmmad, Sifat Ahmed, Khan Md Hasib, and Faisal Muhammad Shah. Harnessing large language models over transformer models for detecting bengali depressive social media text: A comprehensive study. *arXiv preprint arXiv:2401.07310*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Junyi Gao, Cao Xiao, Lucas M Glass, and Jimeng 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*, pages 803–812, 2020.

NithyaJ Gogtay, N Chaudhari, R Ravi, and UM Thatte. Recruitment and retention of the participants in clinical trials: Challenges and solutions. perspect. *Clin. Res*, 11:64–69, 2020.

Candida M Greco and Andrea Tagarelli. Bringing order into the realm of transformer-based language models for artificial intelligence and law. *Artificial Intelligence and Law*, pages 1–148, 2023.

Timothy C Hardman, Rob Aitchison, Richard Scaife, Jean Edwards, and Gill Slater. The future of clinical trials and drug development: 2050. *Drugs in Context*, 12, 2023.

Santosh G Honavar. Electronic medical records–the good, the bad and the ugly, 2020.

Abdalah Ismail, Talha Al-Zoubi, Issam El Naqa, and Hina Saeed. The role of artificial intelligence in hastening time to recruitment in clinical trials. *BJR—Open*, 5(1):20220023, 2023.

Qiao Jin, Zifeng Wang, Charalampos S Floudas, Jimeng Sun, and Zhiyong Lu. Matching patients to clinical trials with large language models. *ArXiv*, 2023.

Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. Eliie: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6):1062–1071, 2017.

Vishesh Kumar, Amber Stubbs, Stanley Shaw, and Özlem Uzuner. Creation of a new longitudinal corpus of clinical narratives. *Journal of biomedical informatics*, 58:S6–S10, 2015.

Vincent Margot and George Luta. A new method to compare the interpretability of rule-based algorithms. *Ai*, 2(4):621–635, 2021.

Lauren McMillan. *Artificial Intelligence–enabled self-healing infrastructure systems*. PhD thesis, UCL (University College London), 2024.

Stéphane M Meystre, Paul M Heider, Andrew Cates, Grace Bastian, Tara Pittman, Stephanie Gentilin, and Teresa J Kelechi. Piloting an automated clinical trial eligibility surveillance and provider alert system based on artificial intelligence and standard data models. *BMC Medical Research Methodology*, 23(1):88, 2023.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: Fast and robust models for biomedical natural language processing. *ArXiv*, abs/1902.07669, 2019. URL https://api.semanticscholar.org/CorpusID:67788603.

Anna Ostropolets, RuiJun Chen, Linying Zhang, and George Hripcsak. Characterizing physicians' information needs related to a gap in knowledge unmet by current evidence. *JAMIA open*, 3(2):281–289, 2020.

Marios Papachristou, Longqi Yang, and Chin-Chia Hsu. Leveraging large language models for collective decision-making. *arXiv preprint arXiv:2311.04928*, 2023.

Cecilia Maria Patino and Juliana Carvalho Ferreira. Inclusion and exclusion criteria in research studies: definitions and why they matter. *Jornal Brasileiro de Pneumologia*, 44:84–84, 2018.

Jeno Pizarro. negspacy. https://github.com/jenojp/negspacy, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Mozhgan Saeidi. Contextbert: Contextual graph representation learning in text disambiguation.

Mozhgan Saeidi, Evangelos Milios, and Norbert Zeh. Contextualized knowledge base sense embeddings in word sense disambiguation. In *Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 174–186. Springer, 2021a.

Mozhgan Saeidi, Evangelos Milios, and Norbert Zeh. Graph representation learning in document wikification. In *Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 509–524. Springer, 2021b.

Mozhgan Saeidi, Evangelos Milios, and Norbert Zeh. Graph convolutional networks for categorizing online harassment on twitter. In *2021 20th IEEE International Conference on Machine Learning* *and Applications (ICMLA)*, pages 946–951. IEEE, 2021c.

Mozhgan Saeidi, Evangelos Milios, and Norbert Zeh. Biomedical word sense disambiguation with contextualized representation learning. In *Companion Proceedings of the Web Conference 2022*, pages 843–848, 2022.

Mozhgan Saeidi, Aman Jaiswal, Abhishek Dhankar, Alan Katz, and Evangelos Milios. Malnis & ema3@ trec 2023 clinical trials track. TREC, 2023a.

Mozhgan Saeidi, Kaveh Mahdaviani, Evangelos Milios, and Norbert Zeh. Context-enhanced concept disambiguation in wikification. *Intelligent Systems with Applications*, 19:200246, 2023b.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Amber Stubbs and Özlem Uzuner. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29, 2015.

Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171, 2019.

F Perry Wilson. *How Medicine Works and when it Doesn't: Learning who to Trust to Get and Stay Healthy*. Grand Central Publishing, 2023.

Michael Wornow, Alejandro Lozano, Dev Dash, Jenelle Jindal, Kenneth W Mahaffey, and Nigam H Shah. Zero-shot clinical trial patient matching with llms. *arXiv preprint arXiv:2402.05125*, 2024.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*, 2023a.

Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Attila Kiraly, Sahar Kazemzadeh, Zakkai Melamed, et al. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv:2308.01317*, 2023b.

Xingyao Zhang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Deepenroll: patient-trial matching with deep embedding and entailment prediction. In *Proceedings of the web conference 2020*, pages 1029–1037, 2020.