Investigating RAG-based Approaches in Clinical Trial and Patient Matching

Daniel Leon Tramontini DANIEL.LEON-TRAMONTINI@STUDENT.UNI-TUEBINGEN.DE
Shrestha Ghosh SHRESTHA.GHOSH@UNI-TUEBINGEN.DE
Carsten Eickhoff CARSTEN.EICKHOFF@UNI-TUEBINGEN.DE
University of Tuebingen, Germany

Abstract

The task of matching clinical trials and patients involves predicting whether a patient meets the eligibility criteria of a clinical trial, via evidences from patient records, such as clinical notes. Given that both the trial eligibility criteria and the clinical notes of patients are unstructured texts, Large Language Models (LLMs) hold the potential to improve performance on this task. Nevertheless, LLMs come with their own challenges of transparency and accountability.

Current methods use Retrieval-Augmented Generation (RAG) in order to predict patient eligibility. In this work, we systematically investigate three aspects of these RAG-based approaches: (i) the complexity of the task, (ii) data retrieval for longitudinal records, and (iii) the effect of abstention on prediction quality. We show that criteria complexity, model abstention and chunking longitudinal patient records have noticeable effects on model performance. We also show that the choice of embedding models and ranking methods has little effect on the evidences retrieved from patient history. We hope that the findings of our study encourage research in improving the transparency and accountability of RAG approaches in clinical decision-making tasks.

Keywords: clinical trial and patient matching, retrieval-augmented generation, clinical decision-making

Data and Code Availability This paper uses the N2C2 cohort selection dataset by Stubbs et al. (2019). It comprises 288 de-identified patient records and their eligibility labels for 13 eligibility criteria. More details are provided in Section 4. The data is currently unavailable for download. Our code is available online at https://github.com/leontramontini97/clinical_trial-patient_matching .

Institutional Review Board (IRB) This research does not require an IRB approval.

1. Introduction

The task of matching clinical trials and patients is challenging. Key clinical information determining patient eligibility is often buried in the clinical notes of longitudinal patient records. Recent works by Jin et al. (2024); Wornow et al. (2025) and Li et al. (2025) highlight the potential of LLMs to predict patient eligibility per criterion and generate explanations for the same.

The state-of-the-art methods tackle the trial and patient matching problem via trial-centric or patient-centric approaches. Trial-centric approaches identify patients relevant to a particular clinical trial, and, patient-centric approaches identify clinical trials relevant to a patient. Trial-centric solutions, such as, Beattie et al. (2024); Wornow et al. (2025) and Li et al. (2025), use the N2C2 Cohort Selection benchmark (Stubbs et al., 2019). This benchmark focuses on predicting eligibility of diabetic patients with heart conditions at a criterion-level based on the evidences from longitudinal patient records. Patient-centric methods, such as, Jin et al. (2024); Rybinski et al. (2024), use the SIGIR (Koopman and Zuccon, 2016) and the TREC (TREC Biomedical Tracks) benchmarks. These benchmarks focus on retrieving and ranking relevant trials from large trial repositories with hundreds of thousands of clinical trials, such as, the ClinicalTrials.gov, given short patient descriptions and cohort keywords. All of the benchmarks discussed above

measure accuracy-based metrics, such as, precision, recall and F1 scores, for overall performance and for measuring ranked results, precision at k, recall at k, normalized discounted cumulative gain (nDCG) at k and mean reciprocal rank are utilized.

Bedi et al. (2025); Omar et al. (2024) and Nemati et al. (2025) have stressed that only accuracy-based measures do not provide any information about LLM uncertainty, fairness and bias. Clinical trial eligibility criteria vary widely in complexity, including assessing non-clinical criteria (such as, demographic data), determining the presence of particular conditions, handling temporal constraints, and handling ambiguous criteria. However, previous attempts provide little or no information regarding the relationship between criteria complexity and system performance. In their N2C2 cohort selection task report, Stubbs et al. (2019) note that the lowest performing criteria are those that require complicated reasoning with temporal modifiers and those that required inference. LLM-based systems deploy strategies from Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) to prompt engineering (Brown et al., 2020). Yet, it is unclear how and which of these strategies affect performance and what the role of criteria complexity is.

In this work, we set up a modular RAG pipeline for the task of trial and patient matching and investigate three aspects that affect systems tackling this task.

Firstly, we characterize the complexity of the task by annotating the number of entities and relations in the eligibility criteria. We then propose generalized strategies to infer implicit entities and report their effect on the final performance.

Secondly, we evaluate the effect of different embedding models and ranking strategies for retrieving longitudinal patient records.

Thirdly, we examine the effect of LLM abstention on prediction quality via accuracy and verbalized confidence.

Our aim with this line of investigative approach is to identify the factors that affect the behavior of LLMs and the challenges associated with the process. We hope that this sparks future interpretability research in identifying the underlying mechanisms that cause these LLM behaviors and ways to control them.

2. Related Work

Matching clinical trials and patients is a resource-intensive task and was performed manually by experts with limited automation until only a decade ago (Penberthy et al., 2012). Starting from early attempts with supervised machine learning on manually selected features and learned features (Zhang and Demner-Fushman, 2017; Vazquez et al., 2021), current transformer-based LLMs have created unprecedented leaps in trial recruitment tasks (Jin et al., 2024; Wornow et al., 2025).

Due to the sensitive nature of the task and the lack of public datasets of retrospective matches, a majority of LLM-based methods (Roberts et al., 2022; Jin et al., 2024; Rybinski et al., 2024) are evaluated on the patient-centric trial recommendation task from the Text REtrieval Conference (TREC) tracks on clinical trials[1]. The TREC benchmarks rank relevant trials per patient, but do not provide any criterion-level labels or longitudinal patient data as is encountered in real-world clinical settings. The notable trial-centric public dataset with longitudinal patient data is the 2018 N2C2 cohort selection task (Stubbs et al., 2019). This dataset provides criterion-level labels, but is limited to inclusion criteria for a single cohort. Another line of work transforms clinical eligibility criteria into logical formats that can be queried on structured patient databases (Yuan et al., 2019).

Previous benchmarks in matching clinical trials and patients Koopman and Zuccon (2016); Stubbs et al. (2019); TREC Biomedical Tracks, which use accuracy-based measures for evaluating model performance, do not address issues such as, transparency, accountability, fairness and uncertainty, which are crucial in LLM-based black-box stochastic systems Harrer (2023). A study on multidimensional evaluation of LLM applications in healthcare found that fairness, bias, and toxicity were reported only 15.8% of the time and uncertainty and calibration only 1.2% of the time compared to accuracy measures that were reported 95% of the time (Bedi et al., 2025). An-

1. https://www.trec-cds.org/

Method	Embedding Model	Generation Model	Prompting Technique	Output Complexity
Beattie et al. (2024)	ada-002 (Greene et al., 2022)	GPT-4 (Achiam et al., 2023)	Personified role Manually crafted tips for all criteria Response format: JSON object	Return criterion name, supporting evidence, and binary prediction.
Wornow et al. (2025)	MiniLM-L6-v2 (Reimers and Gurevych, 2019)	GPT-4	Personified role Manually modified criteria definitions (6 of 13) Response format: JSON object	Return criterion name, list of medications, rationale, binary prediction, and confidence rating
Li et al. (2025)	BioBERT (Lee et al., 2019)	Llama-3-8B-Instruct (Dubey et al., 2024)	AI assistant role Response format: Text	Return a binary prediction

Table 1: Break down of the methods by the models used for embedding and generation, the prompting techniques and the output complexity.

other work proposed new checklists for studies on the development and evaluation of LLMs in order to standardize reporting and improve reproducibility (Tripathi et al., 2025).

3. Trial-Patient Matching

We investigate recent methods, such as, Beattie et al. (2024); Wornow et al. (2025); Li et al. (2025), that use LLMs for the trial and patient matching task. These methods follow the RAG approach, where the generation model is prompted with relevant parts of the patient record as evidences to guide the prediction. Table 1 breaks down the methods by embedding models, generative models, prompting techniques and output complexity.

Model Size We see that all the reviewed methods use smaller embedding models (order of parameters less than a billion), with MiniLM (Reimers and Gurevych, 2019) and BioBERT (Lee et al., 2019) having 22.7M and 65M parameters, respectively. The generation models are much larger, from 8B parameters in Llama-3-8B-Instruct (Dubey et al., 2024) to estimated hundreds of billions of parameters in GPT-4.

Prompting Technique Two of the three methods supplement criteria definition with modifications to the original criteria (Wornow et al., 2025) or with tips on how to resolve each criterion (Beattie et al., 2024). They do not specify when and why these criteria modifications and tips work or how to adapt them to new unseen criteria.

Output Complexity The nature of output from the generation model ranges from simple text-based binary values (Li et al., 2025) to JSON objects comprising prediction, criteria name and text evidence (Beattie et al., 2024). Wornow et al. (2025) elicit additional items, such as medication list, rationale, and confidence rating (high, medium, low).

3.1. RAG Pipeline

Data Processing We divide each patient record into smaller chunks and create a vectorstore of the chunk embeddings. Additionally, we maintain the timestamp of the patient visit corresponding to each chunk. We achieve this by extracting the dates from the clinical record using a regular expression and normalizing them using Python's `dateutil.parser` module. The eligibility criteria are annotated (more details in Section 5), and prompts are adopted to make the criteria more explicit. In Section 5, we show how this affects the final performance.

Data Retrieval Each eligibility criterion is embedded using the same embedding model that we use for the patient clinical record. We follow standard information retrieval and fetch the top-k relevant chunks computed using the cosine similarity of the criterion and chunk embeddings. In Section 6, we investigate if the retrievers indeed return chunks with relevant information for eligibility prediction.

Answer Generation We formulate prompts with the criterion and the retrieved chunks for the LLM to predict eligibility. The prompt template comprises general instructions, the task and the output format.

Prompt template

Based on the following clinical record excerpt, determine if the patient meets this criterion:

Criterion: **[Criterion description]**

Relevant Clinical Context:
[Clinical note excerpt]

Provide your analysis in JSON format with the following structure:
{
"status": "met" or "not met",
"justification": "Brief explanation with specific evidence from the clinical context"
}
If the information is insufficient, you must still make a determination of either "met" or "not met" based on the available evidence.

In Section 7, we examine the uncertainty in the model via verbalized certainty values and how this changes when the model is given the option to abstain.

4. Experiment Setup

Dataset We work with the 2018 N2C2 cohort selection dataset (Stubbs et al., 2019). This public dataset comprises eligibility labels for 288 patients on 13 eligibility criteria. The patient records are de-identified longitudinal records, with an average of 2711 tokens per patient. The criteria labels and their definitions are in the Appendix A. We report our results on the test set of 86 patients, comprising (86×13) 1118 patient-criteria labels. We chose the N2C2 dataset over patient-centric datasets, such as, the SIGIR 2016 (Koopman and Zuccon, 2016) and the clinical trial tasks from the TREC Biomedical tracks (TREC Biomedical Tracks), since the N2C2 dataset has criterion-level labels on longitudinal patient datasets. This helps us to explore the relationship between complexity of the eligibility criteria and the system performance.

Baselines We compare our method with the following state-of-the-art (SOTA) LLM-based methods, Wornow et al. (2025), Li et al. (2025) and Beattie et al. (2024). We also report the best method from the original N2C2 task for completeness (Oleynik et al., 2019). All methods, except for Oleynik et al. (2019), use RAG. Li et al. (2025) also train a classifier on top of the LLM, but this provides only marginal improvement (macro-F1 increases by 0.01 to 0.86).

Evaluation Metrics We report the accuracy-based measurements, namely, the precision, recall and F1 scores. Since one or both macro and micro-F1 scores are reported in different studies, we report them both when available for the baselines in Table 2. Unlike micro-F1 which treats all criterion-patient pairs equally, macro-F1 truly reflects the criterion-level performance. For our results, we mean macro-F1, whenever we mention F1 without any prefix.

Implementation Details Each patient record is split into 500-character chunks with a 50-character overlap. We use OpenAI's GPT-4o model for answer generation. We use Open-AI's model text-embedding-ada-002 to generate patient embeddings and store them in a FAISS index. We also use the GPT-4o model for LLM-as-a-judge evaluations of the data retrieval results obtained from the smaller embedding model. Our code is available online[2].

5. Criteria Complexity

The eligibility criteria, expressed in natural language, are often modified to improve matching with patient data Beattie et al. (2024); Wornow et al. (2025). For instance, in the original N2C2 dataset, two criteria *"Advanced cardiovascular disease (CAD)"* and *"Major diabetes-related complication"* were further clarified, specifically the conditions satisfying the terms *"advanced"* and *"major complications"* were laid out for gold standard annotation by experts.

Criteria Characteristics Trial eligibility criteria vary in semantic complexity. Each criterion describes a central entity via relationships to attributes and logical operations. A criterion can range from being fully objective (HbA1c value between 6.5% and 9.5%) to subjective and ambiguous (major complications). Criteria can also be disease-specific or disease-agnostic, such as,

2. https://github.com/leontramontini97/clinical_trial-patient_matching

Methods	Strategy	Macro-F1	Micro-F1
Oleynik et al. (2019) (N2C2 2018 best)	Rule-based classifier	0.75	0.91
Beattie et al. (2024)[a]	RAG	0.75*	0.86
Wornow et al. (2025)	RAG	0.81	0.93
Li et al. (2025) (LLM-Match)	RAG with a trained classification head	0.86	-
Our asessment	RAG with original criteria	0.70	0.76
	RAG with improved criteria	0.81	0.86

**Derived from per criteria scores reported by Beattie et al. (2024).*

[a]Scores reported on a subset of the test set with 40 patients.

Table 2: SOTA performance on N2C2 cohort selection dataset of 288 patients and 13 criteria (train = 202; test = 86).

Criteria Label	#Entities	#Implicit	#Relations
ABDOMINAL	5	2	3
DIETSUPP-2MOS	4	2	3
ASP-FOR-MI	4	1	2
ALCOHOL-ABUSE	3	2	2
MAJOR-DIABETES	3	2	2
KETO-1YR	3	1	2
HBA1C	3	0	2
ADVANCED-CAD	2	1	1
CREATININE	2	1	1
DRUG-ABUSE	2	1	1
MI-6MOS	2	1	1
ENGLISH	1	1	-
MAKES-DECISIONS	1	1	-
Average	2.6	1.2	1.4

Table 3: Annotating defining characteristics of the N2C2 eligibility criteria.

patient demographics and decision-making capability.

As such, this variability in the criteria makes the comparison between trials quite challenging. Following the Chia annotation model Kury et al. (2020), we annotate the entities and relations in the N2C2 selection criteria (see B). Often, criteria assume implicit medical knowledge, such as measurements of a lab value *above normal levels*, adding another layer of complexity. Hence, we additionally annotate each entity as being implicit or explicit. In Table 3, we provide the number of entities, relations, and implicit entities in every criterion of the N2C2 dataset. The average criterion in the dataset has 2.6 entities, about half of which are implicit (1.2) with 1.4 relations between the entities. We find two criteria, ABDOMINAL and DIETSUPP-2MOS with the highest relations (3) and entities (5 and 4, respectively).

We identified the following classes of criterion-specific amendments. These amendments were tested on the same set of retrieved patient chunks, in order to not introduce any confounding effect from the retriever. The results of criterion-specific amendments are reported in Table 4.

Extended Description In the original task, the definitions of two criteria were extended with explicit conditions to reduce ambiguity and subjectivity for the expert human annotators. For instance, the term "advanced" in ADVANCED-CAD was defined to be constrained to two or more of four specific observations. Furthermore, the term "major complication" for MAJOR-DIABETES was confined to any of the six conditions that are strongly correlated with uncontrolled diabetes. We include these extended definitions in the criteria description. We also extend the definition of ABDOMINAL to include examples of intra-abdominal surgeries and rephrase the original criteria to improve clarity.

We see a huge jump in recall for ABDOMINAL, from 0.167 to 0.667 with the extended definition. Clarifying the original definition into two separate conditions - history of an intra-abdominal surgery or small bowel obstruction - and explicitly specifying examples of the former condition spanning types of intra-abdominal procedures, we see that the model becomes better at picking up the more instances of criterion eligibility.

We also see an improvement in the precision for MAJOR-DIABETES from 0.7 to 0.875.

Strategy	Criteria tag	Before			After		
		P	R	F1	P	R	F1
Extended Description	ABDOMINAL	1	0.167	0.286	0.870	0.667	0.755
	MAJOR-DIABETES	0.7	0.814	0.753	0.875	0.814	0.843
	ADVANCED-CAD	0.636	0.933	0.757	0.624	1	0.769
Default Decision	ENGLISH	0.957	0.603	0.740	0.913	1	0.954
	MAKES-DECISIONS	0.982	0.662	0.791	0.974	0.903	0.937
Temporal Tagging	KETO-1YR*	-	-	-	-	-	-
	MI-6MOS	0.417	0.625	0.5	1	0.875	0.933
	DIETSUPP-2MOS	0.813	0.886	0.847	0.875	0.814	0.843
Numerical limits	HBA1C	1	0.657	0.793	1	0.657	0.793
	CREATININE	1	0.627	0.769	1	0.627	0.769
	ALCOHOL-ABUSE	0.75	1	0.857	0.75	1	0.857
(None applied)	DRUG-ABUSE	0.3	1	0.462	0.3	1	0.462
	ASP-FOR-MI	0.9	0.926	0.913	0.9	0.926	0.913
	Overall average	0.788	0.742	0.706	0.840	0.857	0.819

Table 4: Performance per criterion before and after modification. KETO-1YR has no positive labels.

This makes sense since the original definition requires the model to infer whether a condition is diabetes-related or not and whether it is a major complication. The new definition simplifies this by specifically defining the conditions that qualify these conditions.

While we measure positive changes in two criteria, interestingly, ADVANCED-CAD does not improve much. Upon reexamining the modified version of the criterion, we hypothesize that the new definition, in fact, introduces more complexity. At least two of the four conditions must be satisfied for eligibility. The high false positive rate, 0.585, and a low false negative rate, 0.062, also support this. Upon examining the justifications returned by the LLM for some false positives, we find that even though the LLM names two conditions as met, these are, in fact, incorrect.

Explicit Default Decision The two criteria, ENGLISH and MAKES-DECISION, include an implicit assumption that the patient meets them unless evidence to the contrary is present in the dataset. Accordingly, the LLM is instructed to return "met" unless such evidence appears.

The recall for both criteria increased from 0.603 to 1 for ENGLISH, and from 0.662 to 0.903 for MAKES-DECISIONS. While both criteria are defined as the presence of an observation (speaks English; is capable of making decisions), they appear in the clinical notes only when they are not met. With non-English speakers, it is explicitly noted which language they speak and whether they have an interpreter. The evidence that a patient cannot make a decision is implicitly conveyed through the observations and diagnosis of mental health of the patient.

Explicit Temporal Tagging There are three criteria that require temporal decision making, KETO-1YR, MI-6MOS and DIETSUPP-2MOS. This process involves determining the most recent record of the patient, and then assessing whether the criterion is fulfilled within the mentioned time frame. In order to tackle this, the most recent patient visit date is appended with the relevant patient data for the criteria, KETO-1YR, MI-6MOS and DIETSUPP-2MOS. Further, we include specific instructions on how to handle temporal context.

Temporal-specific instructions

IMPORTANT TEMPORAL CONTEXT: This criterion has a time constraint. Pay special attention to dates and timing.The clinical record uses synthetic dates where 2-digit years should be interpreted as 2XXX (e.g., "2/16/51" = "2051-02-16" or "2051-02-16" depending on the

case, not "1951-02-16").The reference date (most recent clinical note) is: ⟨ FROM-PATIENT-RECORD ⟩. This is our present moment. When evaluating time-based criteria, calculate time intervals from events to this reference date. Example: If the reference date is 2151-04-11 and an event occurred on "2/16/51" (2151-02-16), that's about 2 months prior, which IS within The past ⟨ CRITERION-SPECIFIC-TIME ⟩.

While we cannot comment on the effect of the temporal tagging and instructions on KETO-1YR since it is not present in the test data, we see a sharp increase in precision and recall for MI-6MOS, from 0.417 to 1 precision and 0.625 to 0.875 recall. There is no overall effect in DIETSUPP-2MOS. Upon further inspection of the retrieved chunks and the model justifications, we find that the errors stem from the LLM's inability to identify dietary supplements, a rather broad term, and not from its temporal decision-making.

Explicit Numerical Limits Two criteria require an interpretation of lab results: one mentioned explicitly, *i.e.,* HBA1C, whose values must be between 6.5% and 9.5%, and one implicitly, *i.e.,* CREATININE, whose values should be *above normal limits.* A third criterion, ALCOHOL-ABUSE, also requires an inference of whether current alcohol consumption exceeds recommended limits.

The original task does not specify the actual limits for creatinine and alcohol consumption used by the annotators. Given the variance of these limits and the lack of common standards, it is difficult to apply explicit limits, but we can inspect the justifications that the LLM provides to infer the limits the LLM might have used to predict eligibility. In the case of creatinine, we find that sometimes, the LLM assumes the upper limit in the range of 1.3 to 1.5 mg/dL, while in other cases, it refuses to make a decision due to the lack of a specified upper limit. In the case of alcohol abuse, the model makes predictions only when the evidence is very clear, i.e the patient consumes alcohol multiple times daily. In many other cases, the model specifies the recommended limit as 7 or 14 drinks per week for men, but does not make a decision, citing the lack of specific limits.

Information per prompt	**F1**
Default (1 criterion, relevant chunks)	0.76
1 criterion, full patient record	0.63
all criteria, full patient record	0.67

Table 5: Effect of using the entire patient record vs relevant chunks on the F1 score.

It is interesting to note that the model defaults to known limits for men. This risks biased decisions when the model is not provided with clear specifications, especially in cases where men and women have different standard limits.

6. Data Retrieval

Due to the longitudinal nature of the patient data, it is stored in smaller chunks. Then, for matching a criterion, only the relevant chunks are retrieved and sent to the LLM to reduce noise.

Effect of Chunking In Table 5, we report the F1 scores of the model when it is prompted with the entire patient record under two conditions. Once with a single criterion per patient, and then with all criteria per patient. The LLM's performance decreases when provided with the entire patient record, to 0.67 F1 when all criteria are prompted at once, and to 0.63 when prompted with a single criterion. In a separate experiment Wornow et al. (2025) showed that as the number of relevant chunks provided keeps increasing, the performance plateaus, but never drops. This provides evidence of position bias in LLMs. The LLMs focus on the top evidences, hence showing a plateaued gain when presented with all chunks in a record, but ranked versus a drop in performance when the full record is passed as is.

Effect of Embedding Models We compare three embedding models used the RAG methods: Open AI's `text-embedding-ada-002`, Sentence Transformer's `all-MiniLM-L6-v2` and BioBERT model `dmis-lab/biobert-v1.1` and find no difference in the chunks and the order in which they were returned. We take the top 10 chunks returned by these models for every patient-criterion pair and measure the Jaccard index between all pairs of models, computed as the intersection over the union of two sets. The average Jaccard index of the three pairs is 0.298 $\pm$0.05, implying

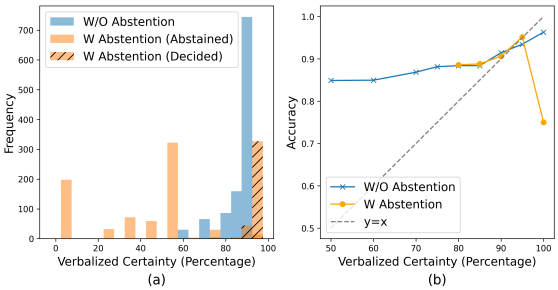

Figure 1: (a) Confidence distribution with and without abstention. (b) Accuracy at confidence thresholds, s.t. accuracy at x is computed for predictions with a conf. $\geq$ x.

that the models return quite different chunks for each patient.

Given that the N2C2 dataset does not include span-level evidence annotations, standard retrieval metrics such as recall@k or precision@k cannot be computed. We use an LLM-as-a-judge for rationale sufficiency assessment as a proxy for retriever sufficiency. This LLM-as-a-judge evaluates the top-10 chunks for sufficiency averaged over a random selection of 10 patients. It returns insufficient information 63% of the time for BioBERT, followed by MiniLM (61%) and ada-002 (60%). In the cases where there is enough information to make a decision, the accuracy of the ada-002 embedding model is the highest at 80%, followed by MiniLM at 78% and BioBERT at 77%.

Effect of Ranking We compare vector embeddings ranked using the FAISS similarity search with BM25 and diversity rankings using Maximal Marginal Rankings (MMR). These methods do return quite different chunks, the average Jaccard similarity of the retrieved chunks with the original vector embeddings is 0.18. When we ask an LLM-as-a-judge whether the top-10 returned chunks are sufficient to make a prediction, it returns yes 25% of the time for the FAISS similarity search, 18.9% of the time for MMR, and 16.5% of the time for BM25.

7. Answer Consistency

Certainty vs. Abstention We conduct two experiments in which the LLM is asked to additionally output a verbalized confidence of its prediction, between 0-100%. In the second run, we additionally provide the model the option to abstain. Figure 1 shows the main results. We also experiment with confidence expressed in levels between 1-5. We found a high Pearson correlation of 0.8 (p-value < 0.05) between the different modes of verbalization.

From Figure 1 (a) we can see that when forced to predict either "met" or "not met", the model mostly also labels its predictions as highly confident, with a mean confidence level of 88% $\pm$ 8.2%. The distribution is more spread out when the model is allowed to abstain. We now see that there are cases where the model is much less confident, with the average confidence equal to 54% $\pm$ 33.5%. Further, the model abstains from prediction if certainty drops below 75%.

Abstaining helps the model improve overall accuracy from 0.84 to 0.88, with a trade-off of an abstention rate of 66.2%. From Figure 1 (b), we see that, the model is overconfident only at 100% confidence level where the accuracy of the model drops to 0.75. The change in certainty in cases where the model does make a decision when given the option to abstain compared to non-abstention is 0.98% $\pm$ 2.4%. This means that the verbalized confidence is stable across prompts for high certainty values (>80%).

Abstention Justifications The justification provided for abstention in 59% of the cases is "insufficient evidence", followed by justifications that mention terms denoting lack of information or no evidence (38.1%). There were 7 instances where the model reasons that there is conflicting information. At a criterion level, the inability to determine if at least two conditions are met for ADVANCED-CAD (2.5%) and a lack of reference range for lab values for creatinine (1.7%) are common justifications for abstention.

8. Discussion and Limitations

Automated Criteria Complexity The results in Section 7, highlight the importance of criteria complexity and the need for criteria-class-specific strategies. We also show how implicit

criteria affect the overall LLM performance. Especially sensitive are cases where the model potentially assumes standards applicable only to certain groups, such as the normal upper creatinine values or weekly alcohol limits for men, putting other groups at a disadvantage. While our study covers only one clinical trial, extending this to other trials on a scale requires an automated method of annotating eligibility criteria and implicit information needs crucial for decision-making.

Evaluating LLM Rationale Although we provide qualitative anecdotes of LLM rationale, there is a need for a more systematic study of evaluating LLM rationale. The rationales, which report absence of sufficient information, indicate information gaps or criteria which depend heavily on lack of evidence as evidence itself, as was the case for ENGLISH and MAKES-DECISION criteria. They also expose model biases, as we saw in the cases of ALCOHOL-ABUSE and CREATININE, where the model defaults to using known information applicable to a particular group, such as recommended limits for men, putting other groups at a disadvantage.

Lack of Rich Data Our study on the N2C2 data comes with caveats: it is trial-specific and has high skew in the label distribution of some criteria. Our next step would be to check the generalizability of our findings regarding criteria importance and model stability on other datasets. In order to achieve this, it is important to develop automated criteria annotators, as discussed above. Although broader datasets such as SIGIR and TREC CT exist, they lack extensive criterion-level annotations, complicating direct evaluation of the components we intend to gain understanding of. Extending our investigation to such datasets requires new benchmarks with consistent ground truth, which we consider an important direction for future work. We also discussed the lack of ground-truth explanations and standard quality tests for evaluating LLM rationale in the preceding paragraph.

In our experiments on abstention in Section 7, the abstention rate was 66.2% with more than 97% of the justifications related to insufficient or no evidence. Under ordinary circumstances with no option of abstaining, the LLM would return highly confident answers even with insufficient evidence. This highlights the importance of tracking implicit information needs, which medical experts do not necessarily need but, that is required in LLM-based systems for a transparent decision-making.

9. Conclusion

In this work, we investigate RAG-based approaches for the task of matching clinical trial and patients via three aspects: criteria complexity, data retrieval and answer generation. We characterize the complexity of eligibility criteria by the number of entities and relations they contain and the number of implicit entities that need resolving. We show generalizable techniques to effectively tackle groups of implicit entities that can lead to gains in performance metrics (F1 0.706 $\rightarrow$ 0.819). We find that while different embedding models and ranking methods retrieve quite different chunks, LLM-as-a-judge evaluates the majority of cases as having insufficient evidence for predicting eligibility with little variation in overall accuracy. Finally, we show that self-reported confidence of LLMs can be unreliable, and abstention reveals decisions LLMs make in the face of insufficient information. Overall, our findings should be viewed as an empirical characterization of current RAG components. We hope that our findings encourage future research in improving LLM-based clinical decision-making and identifying underlying mechanisms affecting LLM behavior.

Acknowledgments

We thank Boehringer Ingelheim for generously supporting DLT as a student research assistant. We also thank our reviewers for their helpful comments.

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.

Jacob Beattie, Sarah Neufeld, Daniel Yang, Christian Chukwuma, Ahmed Gul, Neil Desai, Steve Jiang, and Michael Dohopolski. Utilizing large language models for enhanced clinical trial matching: A study on automation in patient screening. *Cureus*, 16(5), 2024.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA*, January 2025. doi: 10.1001/jama.2024.21700. URL https://doi.org/10.1001/jama.2024.21700.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan. New and improved embedding model. https://openai.com/index/new-and-improved-embedding-model/, 2022. Last accessed: 2025 Sep 03.

Stefan Harrer. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90, 2023.

Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. Matching patients to clinical trials with large language models. *Nature communications*, 15(1):9074, 2024.

Bevan Koopman and Guido 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*, pages 669–672, 2016.

Fabrício Kury, Alex Butler, Chi Yuan, Li-heng Fu, Yingcheng Sun, Hao Liu, Ida Sim, Simona Carini, and Chunhua Weng. Chia, a large annotated corpus of clinical trial eligibility criteria. *Scientific Data*, 7(1), 2020.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Xiaodi Li, Shaika Chowdhury, Chung Il Wi, Maria Vassilaki, Xiaoke Liu, Terence T Sio, Owen Garrick, Young J Juhn, James R Cerhan, Cui Tao, et al. Llm-match: An open-sourced patient matching model based on large language models and retrieval-augmented generation. *arXiv preprint arXiv:2503.13281*, 2025.

Ali Nemati, Mohammad Assadi Shalmani, Qiang Lu, and Jake Luo. Benchmarking large language models from open and closed source models to apply data annotation for free-text criteria in healthcare. *Future Internet*, 17(4): 138, 2025.

Michel Oleynik, Amila Kugic, Zdenko Kasáč, and Markus Kreuzthaler. Evaluating shallow and deep learning strategies for the 2018 n2c2

shared task on clinical text classification. *Journal of the American Medical Informatics Association*, 26(11):1247–1254, 2019.

Mahmud Omar, Girish N Nadkarni, Eyal Klang, and Benjamin S Glicksberg. Large language models in medicine: a review of current clinical trials across healthcare applications. *PLOS Digital Health*, 3(11):e0000662, 2024.

Lynne T Penberthy, Bassam A Dahman, Valentina I Petkov, and Jonathan P DeShazo. Effort required in eligibility screening for clinical trials. *Journal of Oncology Practice*, 8(6): 365–370, 2012.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, Steven Bedrick, and William R Hersh. Overview of the trec 2022 clinical trials track. In *TREC*, 2022.

Maciej Rybinski, Wojciech Kusa, Sarvnaz Karimi, and Allan Hanbury. Learning to match patients to clinical trials using large language models. *Journal of Biomedical Informatics*, 159:104734, 2024.

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.

TREC Biomedical Tracks. https://www.trec-cds.org/, 2024. Last accessed: 2024 Oct 31.

Satvik Tripathi, Dana Alkhulaifat, Florence X Doo, Pranav Rajpurkar, Rafe McBeth, Dania Daye, and Tessa S Cook. Development, evaluation, and assessment of large language models (deal) checklist: A technical report, 2025.

Janette Vazquez, Samir Abdelrahman, Loretta M Byrne, Michael Russell, Paul Harris, and Julio C Facelli. Using supervised machine learning classifiers to estimate likelihood of participating in clinical trials of a de-identified version of researchmatch. *Journal of Clinical and Translational Science*, 5(1):e42, 2021.

Michael Wornow, Alejandro Lozano, Dev Dash, Jenelle Jindal, Kenneth W Mahaffey, and Nigam H Shah. Zero-shot clinical trial patient matching with llms. *NEJM AI*, 2(1): AIcs2400360, 2025.

Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, and Chunhua Weng. Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, 26(4):294–305, February 2019. doi: 10.1093/jamia/ocy178. URL https://doi.org/10.1093/jamia/ocy178.

Kevin Zhang and Dina Demner-Fushman. Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations. *Journal of the American Medical Informatics Association*, 24(4):781–787, 2017.

Appendix A. Eligibility Criteria Definition

Table 6, lists the original criteria definitions for each criteria label from the 2018 N2C2 cohort selection dataset Stubbs et al. (2019). Further annotation guidelines were provided for two criteria.

The term "major complication" for MAJOR-DIABETES was defined as any of the following that are a result of (or strongly correlated with) uncontrolled diabetes: amputation, kidney damage, skin conditions, retinopathy, nephropathy, neuropathy.

The term "advanced" in ADVANCED-CAD was defined as having 2 or more of the following: Taking 2 or more medications to treat CAD; History of myocardial infarction (MI); Currently experiencing angina; Ischemia, past or present.

Criteria Label	Definition
DRUG-ABUSE	Drug abuse, current or past
ALCOHOL-ABUSE	Current alcohol use over weekly recommended limits
ENGLISH	Patient must speak English
MAKES-DECISIONS	Patient must make their own medical decisions
ABDOMINAL	History of intra-abdominal surgery, small or large intestine resection, or small bowel obstruction
MAJOR-DIABETES	Major diabetes-related complication.
ADVANCED-CAD	Advanced cardiovascular disease (CAD).
MI-6MOS	MI in the past 6 months
KETO-1YR	Diagnosis of ketoacidosis in the past year
DIETSUPP-2MOS	Taken a dietary supplement (excluding vitamin D) in the past 2 months
ASP-FOR-MI	Use of aspirin to prevent MI
HBA1C	Any hemoglobin A1c (HbA1c) value between 6.5% and 9.5%
CREATININE	Serum creatinine > upper limit of normal

Table 6: Criteria labels and their definitions for the N2C2 cohort selection dataset.

Appendix B. Eligibility Criteria Annotation

In Table 7, we annotate the 2018 N2C2 cohort selection eligibility criteria using the Chia Annotation Model (Kury et al., 2020). Additionally, we label an entity as expressed directly (D) or implicitly (I).

Criteria Label	Entities			Relations	
	Item	Entity (Text)	Expression	Item	Relation (arg1, arg2)
DRUG-ABUSE	T1	Condition (drug abuse)	D	R1	has_temporal (T1, T2)
	T2	Temporal (current or past)	I/D		
ALCOHOL-ABUSE	T1	Condition (alcohol use)	D	R1	has_temporal (T1, T2)
	T2	Temporal (Current)	I/D	R2	has_qualifier (T1, T3)
	T3	Qualifier (over weekly recommended limits)	I		
ENGLISH	T1	Observation (speak English)	I/D		
MAKES-DECISIONS	T1	Observation (make their own medical decision)	I		
ABDOMINAL	T1	Observation (History of)	I/D	R1	has_temporal (T2, T1)
	T2	Procedure (intra-abdominal surgery)	I	*	or (T3, T4)
	T3	Procedure (small or large intestine resection)	D	R2	subsumes (T2, T5)
	T4	Condition (small bowel obstruction)	D		
	T5	Scope (T3, T4)			
MAJOR-DIABETES	T1	Observation (complication)	D	R1	has_qualifier (T1, T2)
	T2	Qualifier (major)	I	R2	has_qualifier (T1, T3)
	T3	Qualifier (diabetes-related)	I		
ADVANCED-CAD	T1	Condition (cardiovascular disease (CAD))	D	R1	has_qualifier(T1, T2)
	T2	Qualifier (advanced)	I		
MI-6MOS	T1	Condition (MI)	D	R1	has_temporal (T1, T2)
	T2	Temporal (past 6 months)	I		
KETO-1YR	T1	Condition (ketoacidosis)	D	R1	has_temporal (T1, T2)
	T2	Temporal (past year)	I	R2	has_context (T1, T3)
	T3	Context (diagnosis)	D		
DIETSUPP-2MOS	T1	Observation (dietary supplement)	I	R1	has_negation (T3, T4)
	T2	Temporal (past 2 months)	I	R2	has_temporal (T1, T2)
	T3	Context (vitamin D)	D	R3	has_context (T1, T3)
	T4	Negation (excluding)			
ASP-FOR-MI	T1	Drug (Aspirin)	D	R1	has_context (T2, T3)
	T2	Condition (MI)	D	R2	has_scope (T1, T4)
	T3	Context (prevent)	I/D		
	T4	Scope (to prevent MI)			
HBA1C	T1	Measurement (hemoglobin A1c (HbA1c))	D	R1	has_value (T1, T2)
	T2	Value (value between 6.5% to 9.5%)	D	R2	has_qualifier (T1, T3)
	T3	Qualifier (Any)	D		
CREATININE	T1	Measurement (Serum creatinine)	D	R1	has_value (T1, T2)
	T2	Value (> upper limit of normal)	I		

Table 7: Entity and relation annotations for N2C2 eligibility criteria according to the Chia Annotation Model. We additionally label an entity as expressed directly (D) or implicitly (I).