A Survey on LLM-Assisted Clinical Trial Recruitment

Shrestha Ghosh[1] **Moritz Schneider**[2] **Carina Reinicke**[2] **Carsten Eickhoff**[1]

[1]University of Tübingen, Germany
[2]Boehringer Ingelheim, Germany
[1]{first.last}@uni-tuebingen.de [2]{first.last}@boehringer-ingelheim.com

Abstract

Clinical trials are designed in natural language and the task of matching them to patients, represented via both structured and unstructured textual data, benefits from knowledge aggregation and reasoning abilities of LLMs. LLMs with their ability to consolidate distributed knowledge hold the potential to build a more general solution than classical approaches that employ trial-specific heuristics. Yet, adoption of LLMs in critical domains, such as clinical research, comes with many challenges, such as, the availability of public benchmarks, the dimensions of evaluation and data sensitivity. In this survey, we contextualize emerging LLM-based approaches in clinical trial recruitment. We examine the main components of the clinical trial recruitment process, discuss existing challenges in adopting LLM technologies in clinical research and exciting future directions.

1 Introduction

Clinical trials evaluate the effects of an intervention on human health. Selecting the precise and required size of patient population is crucial for trial completion. According to various estimates, more than 50% of aborted clinical trials fail due to low accrual rates, and 80% of all clinical trials do not manage to recruit the required patient cohorts within the allotted time (Clinical Trials Arena, 2012; Williams et al., 2015; Pharmaceutical Technology, 2019). Although this trend has steadily declined over the past decade with the intensive use of technology-aided solutions, efficient patient recruitment remains the most crucial bottleneck in clinical trial research (Clinical Trials Arena, 2022). As electronic health records (EHRs) of patients become more accessible, clinical researchers adopt machine intelligence and develop explainable systems to correctly interpret model predictions (Murdoch and Detsky, 2013; Payrovnaziri et al., 2020; von Itzstein et al., 2021).

There has been a rapid development of methods leveraging LLMs for cohort retrieval and modeling (Fang et al., 2022; Tian et al., 2023; Park et al., 2024; Liu et al., 2025a; Wang et al., 2025), trial design (Reinisch et al., 2024; Curran et al., 2024; Bornet et al., 2025; Neehal et al., 2025), trial search (White et al., 2023; Rybinski et al., 2020), trial matching (Jin et al., 2024; Nievas et al., 2024; Wornow et al., 2025), trial outcomes and duration prediction (Reinisch et al., 2024; Yue et al., 2024a,b; Liu et al., 2025c), risk of bias assessment (Lai et al., 2024; Ji et al., 2025), and clinical trial results extraction (Lee et al., 2024), while the community catches up with recommended practices for responsible use of AI throughout the drug development process (Geraci et al., 2025).

Figure 1 shows the components in clinical trial recruitment, namely, data sourcing, information extraction, matching, and evaluation. An expert reviews several hundred patients per trial and can end up spending hours on one patient, hence incurring significant costs (Penberthy et al., 2012; Ni et al., 2015). Even simple automation using table queries and lexical searches saves between 165 hours to 1,329 hours of reviewing time when compared to manual evaluation (Penberthy et al., 2010). In the past, the patient recruitment process has seen relatively low adoption of the pretrained language models (He et al., 2020; Harrer, 2023; Lu et al., 2024). Generative LLMs serve as knowledge aggregators, and through their reasoning and instruction-following capabilities, they have revived research in the task of trial and patient matching (Jin et al., 2024; Nievas et al., 2024; Rybinski et al., 2024; Wornow et al., 2025).

Difference to Prior Work. Despite the rapidly evolving landscape of LLM technology, there is no prior work surveying this area. Gueguen et al. (2025) evaluate public trial matching tools and Layne et al. (2025) compare the efficacy of open

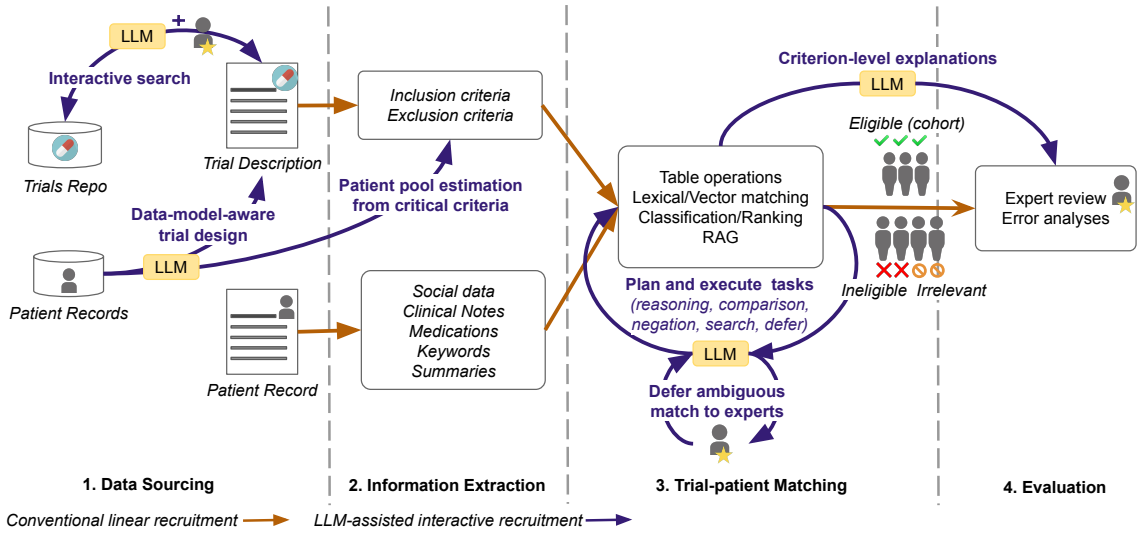

Figure 1: Components in a patient recruitment process: conventional linear flow (in orange) vs. our proposed LLM-assisted interactive flow (in purple).

and proprietary LLM-assisted trial and patient matching in oncology. Systematic reviews on this topic, bound by strict selection criteria and highly specific research question, do not capture the broad perspectives related to dataset challenges and evaluation beyond accuracy (Kim and Quintana, 2022; Idnay et al., 2022; Kantor and Morzy, 2024b; Chen et al., 2025) (See Appendix A). Although it provides an overview of LLM approaches used in clinical trial matching, Chen et al. (2025) only briefly discusses associated challenges and future work.

Our Contribution. We examine the main components of a trial recruitment process, presented in Figure 1. We formalize the problem of trial and patient matching. We analyze existing approaches via the tasks (classification vs. ranking), the directionality (trial-centric and patient-centric), the benchmarks used (longitudinal vs. short patient descriptions, single vs. multi-trial) and the evaluation metrics reported. We present a taxonomy of errors for a consistent evaluation of LLM-generated responses. We discuss the critical challenges associated with the use of LLMs in trial recruitment research. Finally, we present actionable steps towards interactive patient recruitment (illustrated in purple in Figure 1).

2 Background

Clinical Trial Recruitment. Also known as patient recruitment/enrollment/(pre-)screening, it is the process of matching patients (or a cohort) to a clinical trial via its eligibility criteria. Clinical trials have a dual nature, consisting of universal and trial-specific requirements, making it challenging to design generalized approaches (Idnay et al., 2023). This has traditionally resulted in linear trial-centric matching processes (illustrated in orange in Figure 1) with limited scope for interaction and feedback. Standard approaches have an initial data filter on structured EHRs followed by keyword matches and concept identification (Penberthy et al., 2010; Tun et al., 2023) or cohort-specific classifiers (Zhang and Demner-Fushman, 2017). Ni et al. (2015) represented both trials and patients as feature vectors supporting both trial and patient retrieval.

Biomedical NLP and LLMs. The biomedical NLP landscape is shifting from specialized pretrained language models, such as, BioBERT (Lee et al., 2020), BioLM (Lewis et al., 2020), PubmedBERT (Gu et al., 2021), BioGPT (Luo et al., 2022), MedCPT (Jin et al., 2023), among many others (Wang et al., 2023), towards instruction-following and chat-enabled LLMs (commonly termed generative AI), used as is (Nori et al., 2023; Kung et al., 2023) or fine-tuned for domain alignment, such as, Med-PaLM (Singhal et al., 2023), Med-Alpaca (Han et al., 2023) and LLaVA-Med (Li et al., 2023a). We point our reader to Thirunavukarasu et al. (2023) and Liu et al. (2025b) for further reading. The Journal of American Medical Informatics Association (JAMIA), which published 41 articles on biomedical health and LLMs in a focus issue, also observed this shift towards generative AI (Lu et al., 2024). They additionally report that the OpenAI (Achiam et al., 2023) family of proprietary

Search period	2019 - May 2025
Initial search pool *(135 SERPS, 9.8 results/SERP on average)*	1332
Excluded *(not related to trial and patient matching/duplicates)*	1207
Excluded *(full text not available / not a methodology, dataset, evaluation)*	98
Included	27
Included *(via citation network of Jin et al. (2024) and Wornow et al. (2025))*	25
Total included	52
Publication venues covered	27

Table 1: Survey Selection Protocol

models (GPT3.5/4 being the most common) is used much more often than open-sourced models (Touvron et al., 2023; Bai et al., 2023; Jiang et al., 2023). Dorfner et al. (2024) show that fine-tuning generative LLMs on the biomedical domain offers limited performance gain. Another study by Alber et al. (2025) shows that models are more prone to propagate medical misinformation encountered during fine-tuning despite performing well on benchmarks. Bedi et al. (2025) report that real patient data is used in less than 5% of the 519 biomedical studies using LLMs, and fairness, bias, uncertainty and deployment considerations are rarely assessed.

State of Adoption of NLP Advancements. Kantor and Morzy (2024b)'s study on adoption of AI for parsing eligibility criteria reports low adoption rates of generative AI, with BERT-based models being the most popular and generative models being used only since 2024. A systematic review of the role of NLP systems in patient recruitment in 2022 identified only 11 studies (Idnay et al., 2022). Heterogeneous outcomes, diverse results, a dependence on small retrospective data and a lack of common standardized benchmarks drive the gap in NLP research and their adoption in real-world settings (Idnay et al., 2024; Kantor and Morzy, 2024b). A study by Idnay et al. (2023), investigating how clinical researchers screened patients, highlights the challenges of universal and domain-specific nature of the eligibility criteria and makes recommendations to build interactive, flexible and transparent recruitment strategies. Interestingly, when Corbaux et al. (2024) categorize tools for oncological trial matching, they indicate that the automatic methods still fall in the research and development phase, yet to be commercially available.

3 Methodology

Given the interdisciplinary nature of the task and to reduce confirmation bias of known venues, we opted for a broad-scope search via Google Scholar. We queried with the keywords "clinical trial", "cohort discovery", "patient recruitment", "trial recruitment", "trial matching" in conjunction with "llm", "language model", "gpt" and inspected the first ten results pages, between the years 2019 and 2025 (both inclusive). We additionally used Jin et al. (2024) and Wornow et al. (2025) as seeds and recursively traced their citations to efficiently capture the evolution of predictive methods in this task. We finally included 52 papers in our survey of which 6 are from the arXiv and the rest span 27 venues in medical (e.g., JAMIA, Cureus, AMIA), computer science (r.g., TREC, SIGIR) and interdisciplinary (e.g., NEJM AI, Nature Communiations) publications. Table 1 provides details of the search protocol. We organized the papers into data sourcing (Section 4), information extraction and parsing (Section 5), trial and patient matching (Section 6) and evaluation (Section 7). Via our discussion on critical limitations in Section 8 and on promising directions towards interactive patient recruitment in Section 9, we provide a holistic discussion of the challenges of LLM-based trial recruitment pipelines and exciting future directions.

4 Data Sourcing: Public Benchmarks

We start with data sourcing, the first component of the recruitment pipeline (Figure 1). We analyze five trial and patient matching benchmarks.

- 2018 N2C2 Cohort Selection (Stubbs et al., 2019) for criterion-level eligibility prediction.[1]
- Koopman and Zuccon (2016) for ranking trials.
- Text REtrieval Conference Clinical Trial (TREC CT)[2] tracks 2021, 2022, 2023 for ranking trials.

Table 2 provides an overview of the benchmarks.

4.1 Trial Data

The TREC CT benchmarks and Koopman and Zuccon (2016) source the `ClinicalTrials.gov` reg-

[1]Currently unavailable as of 2024 Nov 6.

[2]Available at `https://www.trec-cds.org/`.

Task	Benchmark	Size	Eval. Metrics	Best Score	#P	LLM Usage (#P)
Cohort selection	2018 N2C2 (Stubbs et al., 2019)	#patients = 288 #records per patient = 2-5 #tokens/patient = 2711 #trials = 1 #criteria = 13	Micro-averaged: P, R, F1	0.91 micro F1	47	*(not applicable)*
Trial ranking	Koopman and Zuccon (2016)	#patient summaries (22 words avg) = 60 #patient description (78 words avg) = 60 #keywords (4.4 words avg) = 489 #trials = 204,855 #relevant matches = 685 #relevance judgments = 4000	MRR, P@5, adaptive precision	0.3 MRR $<$ 0.2 P@5	- †	*(not applicable)*
	TREC CT 2021 (Soboroff, 2021)*	#patient descriptions = 75 #trials = 375,580 #relevant matches = 5,570 #relevance judgments = 35,832	NDCG@10, P@10, MRR, R-Precision	0.71 NDCG@10 0.59 P@10 0.82 MRR 0.26 R-Precision	26	BERT-based keyword extraction (9) / query summarization (1), Transformer-based rankers (10)

	TREC CT 2022 (Roberts et al., 2022)	#patient descriptions = 50
		trials = 375,580
		#relevant matches = 3,949
		#relevance judgments = 35,394
			NDCG@10, P@10, MRR, R-Precision	0.61 NDCG@10
				0.50 P@10
				0.72 MRR
				0.32 R-Precision
					12	Query reformulation using BERT / sequence-to-sequence models (3), Transformer-based rankers (3)

	TREC CT 2023 (Rybinski et al., 2024)**	#patient tables = 40 #disease templates = 8 #trials = 451,538 #relevant matches = 11,963 #relevance judgments = 34,931	NDCG@10, P@10, MRR	0.81 NDCG@10 0.73 P@10 0.78 MRR	11	Query reformulation using LLMs (5), Transformer-based rankers (6), LLM prompt-based relevance prediction (4)

*Best scores are aggregated from the Appendix of the runs in the TREC Browser. †No participants, since this was not a challenge.
**Borrowed from (Rybinski et al., 2024) as TREC CT 2023 does not have a published track overview.

Table 2: Overview of the public trial and patient matching benchmarks. #P is the number of participating teams. LLM usage (#P) tracks the number of participants using LLMs that we could verify from the proceedings.

istry. Meanwhile, the N2C2 benchmark focuses on a single trial. `ClinicalTrials.gov` is one of the largest online databases of clinical studies submitted by investigators from over 200 countries. As of June 2025, it lists more than 400,000 clinical trials, thousands of which are active. The `euclinicaltrials.eu` is another public online registry comprising over 50,000 trials from the European Union of which, 7000 are active. Every trial comprises a study description, eligibility criteria and study plan among other details. All the LLM-based methods we analyze use the `ClinicalTrials.gov` registry and one additionally uses EudraCT, the internal database a subset of which is publicly available via the EU Clinical Trials Registry (refer Table 5 in Appendix).

4.2 Patient Data

The N2C2 benchmark contains 288 de-identified longitudinal records of patients. The trial ranking benchmarks use up to 75 synthetic patient profiles, comprising keywords defining cohorts in (Koopman and Zuccon, 2016), short textual patient descriptions: 75 in TREC CT 2021 and 50 in TREC CT 2022, and 40 questionnaire templates in TREC CT 2023. The MIMIC database (Johnson et al., 2016, 2020) is the largest anonymized public database of structured and unstructured data of 299,712 patients. While this dataset is popular for training biomedical LLMs (as mentioned in Section 2), it is not used in any of the five benchmarks, possibly due to significant challenges in obtaining eligibility labels on this scale. Of the 20 LLM-based methods reported in Table 5, 15 used one of the benchmarks from Table 2, 2 used synthetic patient profiles (EHR) made publicly available, 3 used private patient data: 2 clinical notes and 1 structured patient data.

4.3 Annotated Labels

Two medical experts annotated 3,744 criterion-level labels in the N2C2 benchmark. The TREC CT annotations were created by pooling the top-k results from all participating teams. Medical experts manually annotated this pool of trial and patient matches (Koopman and Zuccon, 2016; Soboroff, 2021; Roberts et al., 2022). Out of a total of 35,394 pooled trial and patient matches in the 2022 edition, 11% were judged as *eligible*, 9% as *ineligible* and 80% as *not relevant* with an average of 700 trials judged per patient (Roberts et al., 2022). The relevance judgments were more balanced between the three labels (Rybinski et al., 2024) in TREC CT 2023. All benchmarks depend on manual anno-

tation from experts, which is time-consuming and challenging to scale (Kim and Quintana, 2022).

5 Information Extraction and Parsing

Extraction of medical entities evolved from a combination of rule-based heuristics and feature-based supervised sequence labeling models (Kang et al., 2017; Yuan et al., 2019) via embedding-based neural models (Khan et al., 2019; Tseo et al., 2020) to transformer-based pretrained models and generative AI (Liu et al., 2021c; Zeng et al., 2020; Tian et al., 2021; Li et al., 2022; Murcia et al., 2024; Kantor and Morzy, 2024a). Datta et al. (2024) use disease-specific prompting to extract structured information about entities and its attributes from criteria text. Gao et al. (2020); Zhang et al. (2020) and Theodorou et al. (2023) use BERT-based embeddings to encode eligibility criteria and patient data. Patient data converted to search queries, via reformulation and expansion using LLMs, are particularly effective in trial retrieval (Peikos et al., 2023; Rybinski et al., 2024; Peikos et al., 2024; Jin et al., 2024). Yuan et al. (2019); Tian et al. (2023); Park et al. (2024); Mugambi et al. (2024) and Ziletti and D'Ambrosi (2025) use semantic parsing to translate eligibility criteria into logical forms ready for querying structured patient databases.

6 Trial and Patient Matching

6.1 Formalization

Given sets of inclusion and exclusion criteria (C_{inc}, C_{exc}) from a trial and a set of patient data, P, we formalize the trial and patient matching problem, M, to predict one of the labels *Ineligible (Inel.)*, *Irrelevant (Irr.)* or *Eligible (Eli.)* by aggregating criterion-level binary matches $M'(c, P)$.

$$M(C_{inc}, C_{exc}, P) = \begin{cases} \textit{Inel.}, & \exists c \in C_{exc}, M'(c, P) \\ \textit{Irr.}, & \exists c \in C_{inc}, \neg M'(c, P) \\ \textit{Eli.}, & \neg(\textit{Inel.} \vee \textit{Irr.}) \end{cases} \quad (1)$$

This induces a priority, such that, a patient satisfying any exclusion criteria becomes ineligible, regardless of inclusion criteria matches. Section 6.2 explores the task at varying levels of granularity, starting from criterion-level via trial-level predictions to trial ranking. Criterion-level binary decisions become too restrictive, such that one unmet criterion M' due to lack of data can render the entire match M as ineligible. In such cases, where missing information is expected, ranking trials is a feasible first step. Just ranking trials is insufficient as it does not provide the degree of criteria coverage. In Section 8, we elaborate on the importance of formalization and its effects on trial-level aggregation.

Classical	LLM-based
! Direction-specific approaches, applicable to a set of trials or a cohort	* Direction-agnostic criteria- & trial-level prediction
! Trial-specific heuristics: filters and features	* Generalizable across trials
! Evaluated on private patient data	* Public benchmarks more common

Table 3: Differences between classical and LLM-based trial matching approaches.

Despite the task in Equation 1 being direction-agnostic, there exist two directional approaches to tackle the matching problem, mainly due to data availability. First is the **trial-centric** approach, taken by a trial investigator, that matches longitudinal patient records to a specific trial. The 2018 N2C2 cohort selection is a trial-centric benchmark with criterion-level predictions. Alternatively, a **patient-centric** approach, taken by a patient or their healthcare provider, matches relevant trials from a trial registry to a short description of the patient. Koopman and Zuccon (2016) and the TREC CT 2021, 2022, and 2023 are patient-centric benchmarks for ranking trials.

6.2 LLM-based Approaches

Unlike classical approaches (Penberthy et al., 2010; Ni et al., 2015; Zhang and Demner-Fushman, 2017; Yuan et al., 2019; Tun et al., 2023), LLM-based approaches do not rely on trial-specific heuristics. Table 3 lists the primary differences between the classical and LLM-based approaches (see Appendix C for a full comparison). We group the LLM-based approaches by the granularity of the matches, starting with criterion-level prediction via trial-level prediction to trial ranking. While criterion- and trial-level predictions are direction-agnostic, trial ranking is a patient-centric task. With enough patients, we could evaluate patient ranking, similar to the bidirectional implementation in Ni et al. (2015)'s work, though this has not yet been addressed.

Criterion-Level Prediction. Here, methods utilize the reasoning capability of LLMs to obtain ratio-

nale and other context data in addition to the eligibility label. Hamer et al. (2023) use 1-shot prompt, where given the patient profile and the eligibility criteria, the LLM first labels each criterion as either being applicable to the patient or not, followed by a list of rationales and finally, the eligibility labels for the applicable criteria. Unlu et al. (2024); Beattie et al. (2024) and Wornow et al. (2025) chunk longitudinal patient data and store them as vector embeddings. They prompt LLMs with criteria and relevant patient data chunks under zero-shot setting, with Beattie et al. (2024) providing expert criterion-level strategy and Wornow et al. (2025) providing criteria modifications in the prompts.

Unlu et al. (2024) generate only the decision label from GPT4 and Beattie et al. (2024) and (Wornow et al., 2025) generate criterion-level JSON objects, comprising the criteria label, the eligibility label and a rationale among other context. Both work with proprietary OpenAI models (GPT-3.5 and GPT-4) and Wornow et al. (2025) additionally used open-sourced Llama (Jiang et al., 2023) and Mixtral models (Jiang et al., 2024). Ferber et al. (2024) prompt GPT-4o to generate criterion-level boolean predictions to be reviewed by experts.

Trial-Level Prediction. Wong et al. (2023) used GPT3.5 and GPT4 to extract and structure eligibility criteria into logical expression to be matched locally with structured patient information. Yuan et al. (2023) use LLMs to augment eligibility criteria and pass the BERT-based embeddings of these criteria and patient data through a fully connected classification layer to predict patient-criterion eligibility. The classifier and embedding models are trained jointly on classification loss and a contrastive loss function derived from Equation 1.

Trial Ranking. These methods formulate effective queries, retrieve trials and re-rank them. Query processing involve generating sentence queries from patient descriptions using a fine-tuned T5 model (Pradeep et al., 2022) or zero-shot LLM prompts (Saeidi et al., 2023; Kusa et al., 2023b), generating patient descriptions from trial data via 1-shot LLM prompts (Zhuang et al., 2023), generating noSQL queries via LLMs (Ferber et al., 2024), using LLMs to reformulate and expand queries (Rybinski et al., 2024; Peikos et al., 2024; Datta et al., 2025), and using LLMs to extract keywords (Jullien et al., 2024; Jin et al., 2024; Nievas et al., 2024).

Next comes retrieval, using embeddings similarity (Lahiri et al., 2023; Richmond and Deshpande, 2023; Saeidi et al., 2023; Ferber et al., 2024; Saeidi, 2025) or a multi-stage retrieval with neural re-rankers (Zhuang et al., 2023; Rybinski and Karimi, 2023; Rybinski et al., 2024; Jin et al., 2024; Datta et al., 2025). Some re-rank the top trials using GPT models (Zhuang et al., 2023; Rybinski et al., 2024; Datta et al., 2025), some prompt LLMs for relevance labels (Pradeep et al., 2022), while others prompt LLMs to generate trial- or criterion-level eligibility labels (Rybinski et al., 2024; Peikos, 2023; Jin et al., 2024; Nievas et al., 2024; Jullien et al., 2024). Criterion-level labels are then aggregated using set-based reasoning mechanisms (Jullien et al., 2024), variations of the matching Equation 1 (Jin et al., 2024; Nievas et al., 2024; Saeidi, 2025), and by prompting LLMs (Jin et al., 2024).

7 Evaluation

Standard metrics, such as, precision, recall, F1, normalized discounted cumulative gains (NDCG@k) and mean reciprocal rank (MRR) are popular metrics used by the benchmarks (Table 2). When we compare the different systems side by side (see Table 5 in Appendix) the lack of standardized reporting becomes apparent. Five of the eight (62.5%) methods that tackle criterion-level and trial-level eligibility prediction introduce their own datasets of which three use private patient data.

In the criterion-level prediction task, 3 of the six reported methods use the N2C2 benchmark, namely, Beattie et al. (2024); Wornow et al. (2025) and Saeidi (2025). Even between these three methods, comparison is difficult, since Beattie et al. (2024) report their best results on a subset of the benchmark, and Saeidi (2025) report the performance averaged on N2C2 and TREC CT 2023.

13 of the 14 methods for trial ranking use the TREC CT benchmarks of which 9 evaluate their performance on the 2023 edition, 5 on the 2022 edition and 5 on the 2021 edition. The remaining method ran evaluations on their own dataset of 51 synthetic EHR profiles and 15 trials. 2 methods for trial ranking ran additional evaluations on the SIGIR benchmark.

7.1 Evaluation Beyond Accuracy

Some methods evaluate justification quality of the LLMs via manual evaluation, but differ in the subsets evaluated and the metrics used (Jin et al., 2024; Nievas et al., 2024; Wornow et al., 2025). Notably, researchers in the medical domain have stressed the

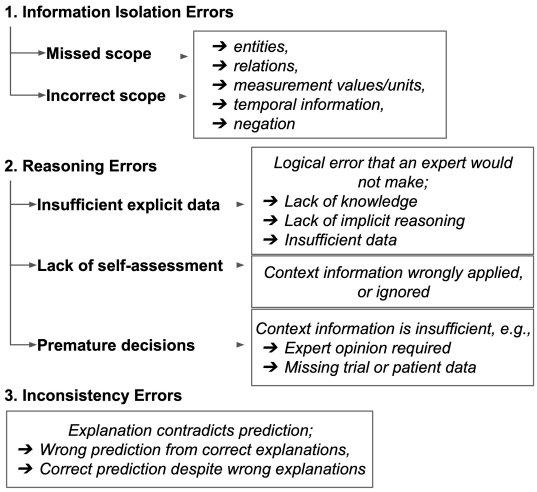

Figure 2: Taxonomy of errors in LLM-generations.

lack of consistent benchmark data, primarily, due to the dependence on manual review which also restricts the size of the data (Kim and Quintana, 2022; Kantor and Morzy, 2024b).

Rybinski et al. (2024) report query latency, with GPT models having 15 times higher latency (47.8s) compared to smaller pre-trained models for a 12% increase in NDCG@10. Wornow et al. (2025) report the cost per patient in terms of money (up to 11.88$), API calls (up to 57) and token used (up to 103,000). Both Hamer et al. (2023) and Jin et al. (2024) report a workload reduction, from a 90% reduction in the criteria to be screened by Harrer (2023) to a 42.6% reduction in screening time by Jin et al. (2024), when using LLM-generated predictions and explanations.

A broader audit on faithfulness of the explanations, logical completeness of the explanations with respect to the criteria, handling missing information, robustness to counterfactuals, uncertainty and bias is lacking in the current methods.

7.2 Taxonomy of LLM Errors

Since there is no one accepted taxonomy of errors, we often come across inconsistent and semantically overlapping categories of errors in LLM generated explanations. For e.g., *incorrect reasoning* and *lack of knowledge* are recognized as independent error types by Jin et al. (2024); Hamer et al. (2023); Nievas et al. (2024), even though the latter often leads to the former (full list in Appendix B). In the proposal by Liévin et al. (2024), reasoning errors are separate from reading comprehension, even though the instances of incorrect reading comprehension occur when the model ignores contextual information and *"reasons"* using its learned knowledge. Here, we identify where errors occur and break them down further by an identifiable source of the error (see Figure 2).

Information Isolation Errors. These are errors in information extraction from patient data or trial criteria falls. This includes ***missed*** or ***incorrect*** NER, measurements (numeric or unit), temporal scopes and negations. LLMs are good at recognizing entities and measurements, but still struggle with negations (Nievas et al., 2024).

Reasoning Errors. An error in the explanations provided by LLMs falls into this category. The most common source is ***insufficient explicit data***, which occurs when LLMs fail to draw logical conclusions from given data, while a human expert can. This stems from previously unseen data (*"lack of knowledge"*) or the inability to recall prior information (*"lack of implicit reasoning"*) or the inability to infer from context (*"insufficient data"*). The second source of reasoning errors is ***lack of self-assessment***. The LLM contradicts explicit information in the prompt. The error occurs when knowledge is wrongly recalled or knowledge is correctly recalled, but contextual information is not applied, resulting in wrong reasoning. This is often referred to as *"incorrect knowledge"*. The third source are ***premature decisions*** made by an LLM when the criterion or patient data is ambiguous and it is necessary to defer to an expert opinion.

Inconsistency Errors. Generating explanations to arrive at final answers unlock LLM's reasoning capacities (Wei et al., 2022). Even so, the explanation and the prediction can be inconsistent. LLMs may predict incorrectly despite correct explanation (reported as explanation-output mismatch in Nievas et al. (2024)). The opposite situation, when a prediction is correct despite an incorrect explanation, is more difficult to identify without human evaluation of automated verification checks. Wornow et al. (2025) report between 3%-8% of incorrect or partially correct justifications despite the LLM making a correct eligibility prediction. This shortcut, similar to cognitive biases in humans, hints towards a bias that the model picked up during training. Both cases negatively impact transparency and accountability of the matching system.

8 Discussion

Annotated Corpora and their Size. Despite attempts to structure clinical trials (Chen et al.,

2022), finding similar trials and analyzing systematic failure cases is notoriously difficult (Rybinski et al., 2020; White et al., 2023). Criterion-level annotations require significant manual annotations, thereby limiting the size of the corpus. For instance, Chia Kury et al. (2020) and LCT (Dobbins et al., 2022), are manually annotated corpora of 1000 trials each, while supervised annotators such as, EliIE (Kang et al., 2017) and Criteria2Query (Yuan et al., 2019), require expensive manual labels (230 disease-specific clinical trials in this case). Eligibility labels on real patient data from real enrollment status, necessary to discount reviewer bias, are only available on a small-scale, if at all, due to the trade-off between scale and privacy (Kim and Quintana, 2022). The systems thus evaluated on private retrospective data (Wong et al., 2023; Yuan et al., 2023; Unlu et al., 2024) cannot be transparently compared. Kantor and Morzy (2024b) stress on standardized benchmarks as the dependence on manual evaluation hinders meta-analyses and comparison between different studies. Public annotations, such as the TREC CT tracks have an average of 700 trial annotations per patient for less than 100 patients, while the N2C2 has only a few thousands of criterion-level annotations. Jointly, the major public corpora, e.g., *ClinicalTrials.gov* and *MIMIC* datasets, present an opportunity to build on the proposal by Kim and Quintana (2022) to generate large-scale data using automated methods.

Dimensions of Evaluation. In Section 7 we see that the majority of methods focus on model accuracy, corroborating the result from Bedi et al. (2025), which reports that more than 95% of studies use accuracy as the primary dimension of evaluation, while fairness, bias and uncertainty are measured less frequently. Omar et al. (2024) reviewed 27 clinical trials evaluating LLMs in healthcare also found the accuracy and reliability standards for LLM use to be undefined. Further results from Nemati et al. (2025), who benchmarked the annotation ability of LLMs across 9 performance metrics, show that while LLMs consistently score high on precision, recall and F1 (lowest being 0.8), their scores highly vary on semantic similarity, factual consistency, relevance, fluency, consistency and coherence (ranging from 0.1 to 0.9) highlighting the need for multiple dimensions of evaluation.

Formalization. Equation 1 highlights the importance of aggregation and priority. Surprisingly, very few works explicitly formalize the matching task. This lack of formalization coincides with an absence of aggregation strategies for trial-level predictions (Hamer et al., 2023; Wornow et al., 2025; Unlu et al., 2024) and others. Formalization guides Yuan et al. (2023) to design a loss function that accounts for the contrastive requirements of inclusion and exclusion criteria, and Jullien et al. (2024) and Saeidi (2025) to define re-ranking scores.

Societal Impact. Recent research measured distinct bias in disease diagnosis across gender, age and disease in popular LLM models (GPT4, ChatGPT and Qwen) (Zhao et al., 2024b) and found that GPT4 tends to stereotype demographic presentations when generating diagnoses (Zack et al., 2024). Alber et al. (2025) show that LLMs are prone to making medical misjudgments by replacing just 0.001% of the training data with medical misinformation. All LLM-based systems, except the trial ranking models, are evaluated on a few disease-specific trials (see Table 5 in Appendix C), the largest being 146 trials, covering 10 cancer types, evaluated by Hamer et al. (2023), making the generalizability of LLMs across diseases unclear. In addition to generalizability tests across diseases, we recommend risk and bias assessments on demographic slices (Benkirane et al., 2025).

Data Sensitivity. Several methods deploy GPT models on Azure AI to comply with privacy regulations (Unlu et al., 2024; Wong et al., 2023; Wornow et al., 2025). Yet, processing patient data with LLMs raises serious ethical challenges due to lack of HIPAA compliance (Edemekong et al., 2024). As already discussed, creating large-scale realistic patient records while protecting their privacy is particularly challenging. We recommend locally deploying open-sourced models or set up a Business Associate Agreement (BAA) with cloud API providers for HIPAA compliance. There is interest in generating synthetic data, as digital twins of patients, with limited access to real patient data as a viable privacy protected alternative (Das et al., 2023; Wang et al., 2024).

Core Limitations of LLMs. According to Harrer (2023) the core limitations affecting LLM adoption are **unfiltered pre-training**, which does not differentiate between facts, opinions, or misinformation; **lack of self-assessment**, where a model generates invalid but syntactically and semantically coherent sentences; **non-determinism**: where surface-form prompt variations lead to drastic changes in the output and repeatability is not guaranteed under

consistent input conditions; and, **knowledge recall**: where updating outdated data or injecting new information requires expensive retraining since the mechanisms of memory in LLMs are not well understood. These limitations pose a direct challenge to the transparency and accountability principles of AI for health laid down by the World Health Organization (Guidance, 2021).

In Section 7, we discuss how knowledge recall and lack of self-assessment surface through reasoning errors. Ways to mitigate non-determinism could include model robustness evaluations to prompt modifications and model settings like temperature and decoding strategies. Using domain-specific LLMs (Singhal et al., 2023), grounding LLMs with external knowledge (Alber et al., 2025), and verifying LLM reasoning could be useful in handling unfiltered pre-training from affecting inference time decision-making

9 Future Directions

As interest in LLMs orchestrating an end-to-end pipeline and incorporating human interactions is gaining more attention (Gao et al., 2024; Qiu et al., 2024), we focus on four promising directions.

Trial Search. Past trials that share a target population are important for designing new trials, recruiting patients, systematic reviews and meta-analyses. This problem has seen little activity since clustering using lexical features (Hao et al., 2014). Newer search methods include constructing a clinical trial knowledge graph Chen et al. (2022), searching via patient EHRs (Wu et al., 2018) and designing features for similarity matchingSun et al. (2022). Jullien et al. (2023) use textual entailment in LLMs to find trials that match short descriptions. However, these models falter on inferences that require numerical reasoning and could not surpass a BM25 baseline for ranking evidence.

Interactive Trial Design. LLM agents have the potential of bridging the semantic gap between eligibility criteria and patient data by suggesting data models underlying patient data for structuring eligibility criteria early on. This collaborative idea is not new (Luo et al., 2013), yet, designing eligibility criteria is a big challenge and has been handled post-hoc by optionally considering patient data models (Kang et al., 2017; Sun and Loparo, 2019; Liu et al., 2021a; Dasgupta et al., 2020). Small patient pools, which ultimately affect the successful completion of a trial, are often the result of restrictive criteria (Clinical Trials Arena, 2022). LLMs can improve trial design by identifying restrictive criteria for trial investigators to relax them and create larger patient pools, especially for trials tackling diseases with a high mortality rate (Liu et al., 2021b).

Collaborative Trial Planning. Liu et al. (2025c) propose an iterative feature discovery model using LLM agents for interpretable trial outcome prediction. Markey et al. (2025) showed promising results on content relevance and suitability of trial protocols generated using LLMs, with room for improvement in logical reasoning and provenance. Similar to Shi et al. (2024), who propose collaboration of agents for knowledge-augmentation and reasoning, LLM agents can identify and distribute tasks and aggregate them to a final result. Another important operation is learning to defer to experts (Mozannar and Sontag, 2020), which can separate operable criteria, such as those that require tabular operations (e.g., via structured queries) or reasoning (temporal, numerical, negation), from criteria that require expert feedback.

Explainable Matches. Explaining black-box LLM predictions in human-understandable form is very challenging (Zhao et al., 2024a) and in clinical trial matches, this is limited to chain-of-thought generations, which is only one of the many facets of explainability (Nauta et al., 2023; Bodria et al., 2023; Chen and Eickhoff, 2024). Consistency and completeness checks, logical component matches, robustness tests on negative, numeric and temporal logic can greatly improve the reliability of guided explanations. Wong et al. (2023), for instance, could potentially explain eligibility via logical component (mis)matches. We hope that the error taxonomy discussed in Section 7.2 assists in systematic evaluation of explanations.

10 Conclusion

The task of clinical trial recruitment, that matches patients to a clinical trial via its participation eligibility criteria, benefits from knowledge aggregation and reasoning abilities of LLMs. In this survey, we critically examine the evolving role of LLM technologies in clinical research. We analyze the main components in a clinical trial recruitment process and provide a modern perspective on the challenges in adopting LLMs to clinical research, such as the benchmarks used, the dimensions of evaluation and data sensitivity. We hope that this serves as a valuable resource for future research.

11 Limitations

This survey focuses on the role of advances in NLP in the critical domain of clinical trial recruitment. Given that this is a rapidly evolving field, we have made our best effort to include a comprehensive view of available resources and methods. It is possible that more sophisticated methods using the latest technology already exist (e.g., in the form of proprietary products), but are not yet made public or are only available as abstracts, as is common in some medical communities, for example, the Annual Meeting of the American Society of Clinical Oncology (ASCO). Clinical trial and patient matching involves sensitive data, and is therefore vulnerable to dual-use risk, which must be challenged and debated by experts on ethics, governance, and technology. Such extensive discussion is out of the scope of this work, but we point our readers to dedicated research on these topics (Braun, 2021; Li et al., 2023b).

Acknowledgements

We would like to thank our reviewers for their thoughtful comments and feedback.

References

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

ACM Computing Surveys. https://dl.acm.org/journal/csur/author-guidelines. Last accessed: 2024 Dec 6.

Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Valliani, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, et al. 2025. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, pages 1–9.

ARR. https://aclrollingreview.org/cfp. Last accessed: 2024 Dec 6.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

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. 2025. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*.

Kenza Benkirane, Jackie Kay, and Maria Perez-Ortiz. 2025. How can we diagnose and treat bias in large language models for clinical decision-making? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2263–2288. Association for Computational Linguistics.

Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2023. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778.

Alban Bornet, Philipp Khlebnikov, Florian Meer, Quentin Haas, Anthony Yazdani, Boya Zhang, Poorya Amini, and Douglas Teodoro. 2025. Analysis of eligibility criteria clusters based on large language models for clinical trial design. *Journal of the American Medical Informatics Association*, 32(3):447–458.

Matthias Braun. 2021. Represent me: please! towards an ethics of digital twins in medicine. *Journal of Medical Ethics*, 47(6):394–400.

Catherine Chen and Carsten Eickhoff. 2024. Evaluating search system explainability with psychometrics and crowdsourcing. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1051–1061.

Hongyu Chen, Xiaohan Li, Xing He, Aokun Chen, James McGill, Emily C Webber, Hua Xu, Mei Liu, and Jiang Bian. 2025. Enhancing patient-trial matching with large language models: A scoping review of emerging applications and approaches. *JCO Clinical Cancer Informatics*, 9:e2500071.

Ziqi Chen, Bo Peng, Vassilis N Ioannidis, Mufei Li, George Karypis, and Xia Ning. 2022. A knowledge graph of clinical trials (ctkg). *Scientific reports*, 12(1):4724.

Clinical Trials Arena(2012). 2012. Clinical Trials Arena. Clinical Trial Delays: America's Patient Recruitment Dilemma. https://www.clinicaltrialsarena.com/marketdata/featureclinical-trial-patient-recruitment/. Last accessed: 2024 Oct 31.

Clinical Trials Arena(2022). 2022. Clinical Trials Arena. Trial termination analysis unveils a silver lining for patient recruitment. https://www.clinicaltrialsarena.com/

features/clinical-trial-terminations/. Last accessed: 2024 Oct 31.

P Corbaux, A Bayle, S Besle, A Vinceneux, H Vanacker, K Ouali, B Hanvic, C Baldini, PA Cassier, C Terret, et al. 2024. Patients' selection and trial matching in early-phase oncology clinical trials. *Critical Reviews in Oncology/Hematology*, page 104307.

Corey Curran, Nafis Neehal, Keerthiram Murugesan, and Kristin P Bennett. 2024. Examining trustworthiness of llm-as-a-judge systems in a clinical trial design benchmark. In *2024 IEEE International Conference on Big Data (BigData)*, pages 4627–4631. IEEE.

Trisha Das, Zifeng Wang, and Jimeng Sun. 2023. Twin: Personalized clinical trial digital twin generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 402–413.

Tirthankar Dasgupta, Ishani Mondal, Abir Naskar, and Lipika Dey. 2020. Extracting semantic aspects for structured representation of clinical trial eligibility criteria. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 243–248.

Surabhi Datta, Kyeryoung Lee, Liang-Chin Huang, Hunki Paek, Roger Gildersleeve, Jonathan Gold, Deepak Pillai, Jingqi Wang, Mitchell K Higashi, Lizheng Shi, et al. 2025. Patient2trial: From patient to participant in clinical trials using large language models. *Informatics in Medicine Unlocked*, 53:101615.

Surabhi Datta, Kyeryoung Lee, Hunki Paek, Frank J Manion, Nneka Ofoegbu, Jingcheng Du, Ying Li, Liang-Chin Huang, Jingqi Wang, Bin Lin, et al. 2024. Autocriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *Journal of the American Medical Informatics Association*, 31(2):375–385.

Nicholas J Dobbins, Tony Mullen, Özlem Uzuner, and Meliha Yetisgen. 2022. The leaf clinical trials corpus: a new resource for query generation from clinical trial eligibility criteria. *Scientific Data*, 9(1):490.

Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, et al. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*.

Peter Edemekong, Pavan Annamaraju, Muriam Afzal, and Micelle Haydel. 2024. Health insurance portability and accountability act (hipaa) compliance. *StatPearls*.

EudraCT. https://eudract.ema.europa.eu/results-web/. Last accessed: 2024 Dec 3.

Yilu Fang, Betina Idnay, Yingcheng Sun, Hao Liu, Zhehuan Chen, Karen Marder, Hua Xu, Rebecca Schnall, and Chunhua Weng. 2022. Combining human and machine intelligence for clinical trial eligibility querying. *Journal of the American Medical Informatics Association*, 29(7):1161–1171.

Dyke Ferber, Lars Hilgers, Isabella C Wiest, Marie-Elisabeth Leßmann, Jan Clusmann, Peter Neidlinger, Jiefu Zhu, Georg Wölflein, Jacqueline Lammert, Maximilian Tschochohei, et al. 2024. End-to-end clinical trial matching with large language models. *arXiv preprint arXiv:2407.13463*.

Junyi Gao, Cao Xiao, Lucas M Glass, and Jimeng Sun. 2020. 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.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151.

Joseph Geraci, Prasanna Rao, Cheryl Grandinetti, Bessi Qorri, Patrick Nadolny, Kassa Ayalew, Lisbeth Bregnhøj, Lindsay Edwards, Karen Hofmann, Sean Khozin, et al. 2025. Current opportunities for the integration and use of artificial intelligence and machine learning in clinical trials: Good clinical practice perspectives. *Journal of the Society for Clinical Data Management*, 5(2).

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Lilia Gueguen, Louise Olgiati, Clément Brutti-Mairesse, Alric Sans, Vincent Le Texier, and Loic Verlingue. 2025. A prospective pragmatic evaluation of automatic trial matching tools in a molecular tumor board. *npj Precision Oncology*, 9(1):28.

WHO Guidance. 2021. Ethics and governance of artificial intelligence for health. *World Health Organization*.

Danny M den Hamer, Perry Schoor, Tobias B Polak, and Daniel Kapitan. 2023. Improving patient pre-screening for clinical trials: assisting physicians with large language models. *arXiv preprint arXiv:2304.07396*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Tianyong Hao, Alexander Rusanov, Mary Regina Boland, and Chunhua Weng. 2014. Clustering clinical trials with similar eligibility criteria features. *Journal of biomedical informatics*, 52:112–120.

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

Zhe He, Xiang Tang, Xi Yang, Yi Guo, Thomas J George, Neil Charness, Kelsa Bartley Quan Hem, William Hogan, and Jiang Bian. 2020. Clinical trial generalizability assessment in the big data era: a review. *Clinical and translational science*, 13(4):675–684.

Betina Idnay, Caitlin Dreisbach, Chunhua Weng, and Rebecca Schnall. 2022. A systematic review on natural language processing systems for eligibility prescreening in clinical research. *Journal of the American Medical Informatics Association*, 29(1):197–206.

Betina Idnay, Yilu Fang, Caitlin Dreisbach, Karen Marder, Chunhua Weng, and Rebecca Schnall. 2023. Clinical research staff perceptions on a natural language processing-driven tool for eligibility prescreening: an iterative usability assessment. *International journal of medical informatics*, 171.

Betina Idnay, Emily R Gordon, Aubrey S Johnson, Jordan G Nestor, Karen Marder, and Chunhua Weng. 2024. Clinical researchers' insights on key data for eligibility screening in clinical studies. *Journal of Clinical and Translational Science*, 8(1):e167.

Changkai Ji, Bowen Zhao, Zhuoyao Wang, Yingwen Wang, Yuejie Zhang, Ying Cheng, Rui Feng, and Xiaobo Zhang. 2025. Robguard: Enhancing llms to assess risk of bias in clinical trial documents. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1258–1277.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

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

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. Mimic-iv. *PhysioNet. Available online at:* https://physionet.org/content/mimiciv/1.0/, pages 49–55.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Mael Jullien, Alex Bogatu, Harriet Unsworth, and Andre Freitas. 2024. Controlled llm-based reasoning for clinical trial retrieval. *arXiv preprint arXiv:2409.18998*.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and André Freitas. 2023. Nli4ct: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764.

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

Klaudia Kantor and Mikolaj Morzy. 2024a. Fine-tuned transformers and large language models for entity recognition in complex eligibility criteria for clinical trials. In *Harnessing Opportunities: Reshaping ISD in the post-COVID-19 and Generative AI Era (ISD2024 Proceedings)*.

Klaudia Kantor and Mikołaj Morzy. 2024b. Machine learning and natural language processing in clinical trial eligibility criteria parsing: a scoping review. *Drug Discovery Today*, 29(10):104139.

Md Abdullah Al Hafiz Khan, Md Shamsuzzaman, Sadid A Hasan, Mohammad S Sorower, Joey Liu, Vivek Datla, Mladen Milosevic, Gabe Mankovich, Rob van Ommering, and Nevenka Dimitrova. 2019. Improving disease named entity recognition for clinical trial matching. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2541–2548. IEEE.

Jeongeun Kim and Yuri Quintana. 2022. Review of the performance metrics for natural language systems for clinical trials matching. *MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation*, pages 641–644.

Bevan Koopman and Guido Zuccon. 2016. 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.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

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

Wojciech Kusa, Óscar E Mendoza, Petr Knoth, Gabriella Pasi, and Allan Hanbury. 2023a. Effective matching of patients to clinical trials using entity extraction and neural re-ranking. *Journal of biomedical informatics*, 144:104444.

Wojciech Kusa, Patrick Styll, Maximilian Seeliger, Óscar Espitia Mendoza, and Allan Hanbury. 2023b. Dossier at trec 2023 clinical trials track. In *TREC*.

Aritra Kumar Lahiri, Emrul Hasan, Qinmin Vivian Hu, and Cherie Ding. 2023. TMU at trec clinical trials track 2023. In *TREC*.

Honghao Lai, Long Ge, Mingyao Sun, Bei Pan, Jiajie Huang, Liangying Hou, Qiuyu Yang, Jiayi Liu, Jianing Liu, Ziying Ye, et al. 2024. Assessing the risk of bias in randomized clinical trials with large language models. *JAMA Network Open*, 7(5):e2412687–e2412687.

Ethan Layne, Claire Olivas, Jacob Hershenhouse, Conner Ganjavi, Francesco Cei, Inderbir Gill, and Giovanni E Cacciamani. 2025. Large language models for automating clinical trial matching. *Current opinion in urology*, 35(3):250–258.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kyeryoung Lee, Hunki Paek, Liang-Chin Huang, C Beau Hilton, Surabhi Datta, Josh Higashi, Nneka Ofoegbu, Jingqi Wang, Samuel M Rubinstein, Andrew J Cowan, et al. 2024. Seetrials: Leveraging large language models for safety and efficacy extraction in oncology clinical trials. *Informatics in medicine unlocked*, 50:101589.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In *Proceedings of the 3rd clinical natural language processing workshop*, pages 146–157.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. 2023b. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 5(6):e333–e335.

Jianfu Li, Qiang Wei, Omid Ghiasvand, Miao Chen, Victor Lobanov, Chunhua Weng, and Hua Xu. 2022. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. *BMC medical informatics and decision making*, 22(Suppl 3):235.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Changshuo Liu, Lingze Zeng, Kaiping Zheng, Shaofeng Cai, Beng Chin Ooi, and James Wei Luen Yip. 2025a. Neuralcohort: Cohort-aware neural representation learning for healthcare analytics. In *Forty-second International Conference on Machine Learning*.

Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Yining Hua, Peilin Zhou, et al. 2025b. Application of large language models in medicine. *Nature Reviews Bioengineering*, pages 1–20.

Fengze Liu, Haoyu Wang, Joonhyuk Cho, Dan Roth, and Andrew W Lo. 2025c. Autoct: Automating interpretable clinical trial prediction with llm agents. *arXiv preprint arXiv:2506.04293*.

Hao Liu, Yuan Chi, Alex Butler, Yingcheng Sun, and Chunhua Weng. 2021a. A knowledge base of clinical trial eligibility criteria. *Journal of biomedical informatics*, 117:103771.

Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping, et al. 2021b. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature*, 592(7855):629–633.

Xiong Liu, Greg L Hersch, Iya Khalil, and Murthy Devarakonda. 2021c. Clinical trial information extraction with bert. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 505–506. IEEE.

Zhiyong Lu, Yifan Peng, Trevor Cohen, Marzyeh Ghassemi, Chunhua Weng, and Shubo Tian. 2024. Large language models in biomedicine and health: current research landscape and future directions. *Journal of the American Medical Informatics Association*, 31(9):1801–1811.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Zhihui Luo, Riccardo Miotto, and Chunhua Weng. 2013. A human–computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. *Journal of biomedical informatics*, 46(1):33–39.

Nigel Markey, Ilyass El-Mansouri, Gaetan Rensonnet, Casper van Langen, and Christoph Meier. 2025. From rags to riches: Utilizing large language models to write documents for clinical trials. *Clinical Trials*, page 17407745251320806.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and t PRISMA Group*. 2009. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269.

Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pages 7076–7087. PMLR.

Purity Mugambi, Alexandra Meliou, and Madalina Fiterau. 2024. Leveraging foundation language models (flms) for automated cohort extraction from large ehr databases. *arXiv preprint arXiv:2412.11472*.

Victor M Murcia, Vinod Aggarwal, Nikhil Pesaladinne, Ram Thammineni, Nhan Do, Gil Alterovitz, and Rafael B Fricks. 2024. Automating clinical trial matches via natural language processing of synthetic electronic health records and clinical trial eligibility criteria. *AMIA Summits on Translational Science Proceedings*, 2024:125.

Travis B Murdoch and Allan S Detsky. 2013. The inevitable application of big data to health care. *Jama*, 309(13):1351–1352.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.

Nafis Neehal, Bowen Wang, Shayom Debopadhaya, Corey Curran, Keerthiram Murugesan, Soham Dan, Vibha Anand, and Kristin Bennett. 2025. Are large language models effective in clinical trial design? a study on baseline feature generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5557–5570.

Ali Nemati, Mohammad Assadi Shalmani, Qiang Lu, and Jake Luo. 2025. 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.

Yizhao Ni, Jordan Wright, John Perentesis, Todd Lingren, Louise Deleger, Megan Kaiser, Isaac Kohane, and Imre Solti. 2015. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. *BMC medical informatics and decision making*, 15:1–10.

Mauro Nievas, Aditya Basu, Yanshan Wang, and Hrituraj Singh. 2024. Distilling large language models for matching patients to clinical trials. *Journal of the American Medical Informatics Association*, pages 1953–1963.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

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

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Jimyung Park, Yilu Fang, Casey Ta, Gongbo Zhang, Betina Idnay, Fangyi Chen, David Feng, Rebecca Shyu, Emily R Gordon, Matthew Spotnitz, et al. 2024. Criteria2query 3.0: Leveraging generative large language models for clinical trial eligibility query generation. *Journal of Biomedical Informatics*, 154:104649.

Seyedeh Neelufar Payrovnaziri, Zhaoyi Chen, Pablo Rengifo-Moreno, Tim Miller, Jiang Bian, Jonathan H Chen, Xiuwen Liu, and Zhe He. 2020. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27(7):1173–1185.

Georgios Peikos. 2023. UNIMIB at TREC 2023 Clinical Trials Track. In *TREC*.

Georgios Peikos, Daria Alexander, Gabriella Pasi, and Arjen P de Vries. 2023. Investigating the impact of query representation on medical information retrieval. In *European Conference on Information Retrieval*, pages 512–521. Springer.

Georgios Peikos, Pranav Kasela, and Gabriella Pasi. 2024. Leveraging large language models for medical information extraction and query generation. In *2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 367–372. IEEE.

Lynne Penberthy, Richard Brown, Federico Puma, and Bassam Dahman. 2010. Automated matching software for clinical trials eligibility: measuring efficiency and flexibility. *Contemporary clinical trials*, 31(3):207–217.

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

Pharmaceutical Technology(2019). 2018. Pharmaceutical Technology. Enrolment Issues are the Top Factor in Clinical Trial Terminations. https://www.pharmaceutical-technology.com/analyst-comment/reasons-for-clinical-trial-termination/?cf-view. Last accessed: 2024 Oct 31.

Ronak Pradeep, Yilin Li, Yuetong Wang, and Jimmy Lin. 2022. Neural query synthesis and domain-specific ranking templates for multi-stage clinical trial matching. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2325–2330.

Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J Topol. 2024. Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, pages 1–3.

Michael Reinisch, Jianfeng He, Chenxi Liao, Sauleh Siddiqui, and Bei Xiao. 2024. Ctp-llm: Clinical trial phase transition prediction using large language models. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3667–3672. IEEE.

Luke Richmond and Priya Deshpande. 2023. Leveraging openai's ada embedding model for zero-shot classification at trec 2023 clinical trials. In *TREC*.

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

Maciej Rybinski and Sarvnaz Karimi. 2023. Matching of patient questionnaires to clinical trials with large language models. In *TREC*.

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

Maciej Rybinski, Jerry Xu, and Sarvnaz Karimi. 2020. Clinical trial search: Using biomedical language understanding models for re-ranking. *Journal of Biomedical Informatics*, 109:103530.

Mozhgan Saeidi. 2025. Streamlining clinical trial recruitment: A two-stage zero-shot llm approach with advanced prompting. In *Machine Learning for Health (ML4H)*, pages 886–896. PMLR.

Mozhgan Saeidi, Aman Jaiswal, Abhishek Dhankar, Alan Katz, and Evangelos Milios. 2023. MALNIS & EMA3 TREC 2023 Clinical Trials Track. In *TREC*.

Hanwen Shi, Jin Zhang, and Kunpeng Zhang. 2024. Enhancing clinical trial patient matching through knowledge augmentation and reasoning with multi-agents. *arXiv preprint arXiv:2411.14637*.

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. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Ian Soboroff. 2021. Overview of trec 2021. In *TREC*.

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

Yingcheng Sun and Kenneth Loparo. 2019. Knowledge-guided text structuring in clinical trials. In *Advances in Data Mining - Applications and Theoretical Aspects, 19th Industrial Conference, ICDM 2019, Poster Proceedings*, pages 211–219.

Yingcheng Sun, Jiaqi Tang, Alex M. Butler, Cong Liu, Yilu Fang, and Chunhua Weng. 2022. Interactive similarity-based search of clinical trials. *Studies in health technology and informatics*, 290.

TACL. https://transacl.org/index.php/tacl/about/submissions. Last accessed: 2024 Dec 6.

Brandon Philip Theodorou, Cao Xiao, and Jimeng Sun. 2023. Treement: Interpretable patient-trial matching via personalized dynamic tree-based memory network. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Shubo Tian, Arslan Erdengasileng, Xi Yang, Yi Guo, Yonghui Wu, Jinfeng Zhang, Jiang Bian, and Zhe He. 2021. Transformer-based named entity recognition for parsing clinical trial eligibility criteria. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–6.

Shubo Tian, Pengfei Yin, Hansi Zhang, Arslan Erdengasileng, Jiang Bian, and Zhe He. 2023. Parsing clinical trial eligibility criteria for cohort query by a multi-input multi-output sequence labeling model. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4426–4430. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yitong Tseo, MI Salkola, Ahmed Mohamed, Anuj Kumar, and Freddy Abnousi. 2020. Information extraction of clinical trial eligibility criteria. *arXiv preprint arXiv:2006.07296*.

Pyae Phyo Tun, Jiawen Luo, Jiecheng Xie, Sandi Wibowo, and Chen Hao. 2023. Automatic assessment of patient eligibility by utilizing nlp and rule-based analysis. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE.

Ozan Unlu, Jiyeon Shin, Charlotte J Mailly, Michael F Oates, Michela R Tucci, Matthew Varugheese, Kavishwar Wagholikar, Fei Wang, Benjamin M Scirica, Alexander J Blood, et al. 2024. Retrieval-augmented generation–enabled gpt-4 for clinical trial screening. *NEJM AI*, page AIoa2400181.

Mitchell S. von Itzstein, Melanie Hullings, Helen Mayo, M. Shaalan Beg, Erin L. Williams, and David E. Gerber. 2021. Application of information technology to clinical trial evaluation and enrollment: A review. *JAMA Oncology*, 7(10):1559–1566.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023. Pretrained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52.

Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. 2024. Twin-gpt: digital twins for clinical trials via large language model. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Zifeng Wang, Qiao Jin, Jiacheng Lin, Junyi Gao, Jathurshan Pradeepkumar, Pengcheng Jiang, Benjamin Danek, Zhiyong Lu, and Jimeng Sun. 2025. Trialpanorama: Database and benchmark for systematic review and design of clinical trials. *arXiv preprint arXiv:2505.16097*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Renee White, Tristan Peng, Pann Sripitak, Alexander Rosenberg Johansen, and Michael Snyder. 2023. Clinidigest: a case study in large language model based large-scale summarization of clinical trial descriptions. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, pages 396–402.

Rebecca J Williams, Tony Tse, Katelyn DiPiazza, and Deborah A Zarin. 2015. Terminated trials in the clinicaltrials. gov results database: evaluation of availability of primary outcome data and reasons for termination. *PLoS ONE*, 10(5).

Cliff Wong, Sheng Zhang, Yu Gu, Christine Moung, Jacob Abel, Naoto Usuyama, Roshanthi Weerasinghe, Brian Piening, Tristan Naumann, Carlo Bifulco, et al. 2023. Scaling clinical trial matching using large language models: a case study in oncology. In *Machine Learning for Healthcare Conference*, pages 846–862. PMLR.

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

Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, et al. 2018. Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association*, 25(5):530–537.

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

Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2023. Large language models for healthcare data augmentation: An example on patient-trial matching. In *AMIA Annual Symposium Proceedings*, volume 2023, pages 1324–1333. American Medical Informatics Association.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024a. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10.

Ling Yue, Sixue Xing, Jonathan Li, Md Zabirul Islam, Bolun Xia, Jintai Chen, and Tianfan Fu. 2024b. Trialdura: Hierarchical attention transformer for interpretable clinical trial duration prediction. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–1.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

Kun Zeng, Zhiwei Pan, Yibin Xu, and Yingying Qu. 2020. An ensemble learning strategy for eligibility criteria text classification for clinical trial recruitment: algorithm development and validation. *JMIR Medical Informatics*, 8(7):e17832.

Kevin Zhang and Dina Demner-Fushman. 2017. 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.

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

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024a. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Yutian Zhao, Huimin Wang, Yuqi Liu, Wu Suhuang, Xian Wu, and Yefeng Zheng. 2024b. Can llms replace clinical doctors? exploring bias in disease diagnosis by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13914–13935.

Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2023. Team IELAB at TREC Clinical Trial Track 2023: Enhancing Clinical Trial Retrieval with Neural Rankers and Large Language Models. In *TREC Clinical Trials Track 2023*.

Angelo Ziletti and Leonardo D'Ambrosi. 2025. Generating patient cohorts from electronic health records using two-step retrieval-augmented text-to-sql generation. *arXiv preprint arXiv:2502.21107*.

A Systematic Reviews versus Surveys

A systematic review is a common practice in the medical research community, with standardized reporting guidelines (Moher et al., 2009; Page et al., 2021). This guideline has a checklist of required items in the title, abstract, introduction, methods, results, discussion and other information that every study needs to fulfil. These studies are bound by strict inclusion and exclusion criteria applied to the scientific literature considered for any assessment. This rigor is necessary in evidence-based analysis of a specific research question. A survey is more flexible and provides a board coverage of the topic being discussed. For instance, the guidelines for writing surveys in our community as outlined by ACM Computing Surveys, TACL and ARR aim to draw perspectives on an evolving topic of interest.

ACM Computing Surveys (long paper). A paper that summarizes and organizes recent research results in a novel way that integrates and adds understanding to work in the field. A survey article assumes a general knowledge of the area; it emphasizes the classification of the existing literature, developing a perspective on the area, and evaluating trends.

TACL (excerpt). They should thus not simply be a descriptive enumeration of the contents of papers, but draw broad themes and (importantly) provide new insights on the topic. These insights should be major contributions of the submission.

ARR (note). all papers are expected to include reviews of related literature. This category is meant for the papers that go beyond that, e.g. in scope or in establishing new interdisciplinary connections.

B Reported Error Types

In the following we illustrate the the error distributions by percentages as reported by previous studies.

Errors reported in (Jin et al., 2024) supplementary:
- Incorrect reasoning: 30.7%
- Lack of medical knowledge: 15.4%
- Ambiguous label definition: 26.9%
- other errors: 26.9% (self-conflicting)

Errors reported in (Hamer et al., 2023):
- Incorrect reading: 6.5%
- Insufficient knowledge: 2.2%
- Incorrect reasoning: 91.3%

Errors reported in (Nievas et al., 2024):
- Lack of knowledge 55%
- Implicit criteria 15%
- Wrong reasoning
- Accurate reasoning, wrong decision
- Lack of restraint when expert opinion is needed
- Negated criteria error

C Trial and Patient Matching Systems

Table 4 gives an overview of all the classical systems in covered in this survey by their direction of approach, method, source data characteristics and limitations. Table 5 covers all the LLM-based systems discussed in this survey by their task, method, data source characteristics and limitations. We notice the shift in the availability of patient sources towards public data in the LLM-based approaches. We also notice a shift towards trial ranking due to publicly available data.

Direction	Method	Patient Source	Trial Source	Metrics (Best)	Limitations
Trial-centric	Penberthy et al. (2010) Discrete-data filter and sub-word matching	**Availability**: Private **Size** 282-2,112 **Mode**: text, tables **Additional annotation**: expert review	**Availability**: public (1); unknown (4) **Size**: 5 **Type**: Specific	Yield (% cross trials): 17.2 to 73.8 Efficiency (ratio manual to automated screening time): 0.8 to 19.4	Manually coded eligibility criteria. String similarity for keyword search.
	Yuan et al. (2019) (Criteria2Query) NL to structured queries Information extraction, normalization Mapping to pre-defined cohort templates	No patient data	**Availability**: Public* **Size**: 10 **Type**: Varying **Additional annotation**: 125 criteria sentence, 215 entities, 34 relations, 137 negations, 20 attributes	F1 (NER): 0.79 F1 (RE): 0.80 Accuracy (across tasks): 0.51 to 0.98	Disease-specific NER. Query templates cannot deal with missing attributes. Entity normalization into a small set of 2000 concepts.
	Tun et al. (2023) Rule-based filters Similarity scores per criteria as features to a classifier	**Availability**: Private **Size**: 40,000 **Mode**: text, tables **Additional annotation**: 109 patients labels	**Availability**: Unknown **Size**: 1 **Type**: Specific	Sensitivity (across models): 0.4 to 1.0 Precision (across models): 0.23 to 0.78	Limited to a single trial Rules-based models have high sensitivity (1.0) and low precision (0.23). Hybrid classifier makes a good sensitivity and precision trade-off, but is not robust to criterion changes.
	Ni et al. (2015) Discrete-data filter Index trial and patient bag-of-words vectors Return top vector matches for a trial	**Availability**: Private **Size**: 215 **Mode**: text, tables **Additional annotation**: historical match, expert review	**Availability**: Public* **Size**: 55 **Type**: Specific	Workload reduction: 85% Precision: 12.5% Specificity: 89.9%	Majority of false positives due to lack of semantic knowledge
Patient-centric	Ni et al. (2015) Return top vector matches for a patient	*(same as previous row)*		Workload reduction: 54.7% to 92.8% Precision: 4% to 35.7& Specificity: 65.5% to 95.5%	Majority of false positives due to lack of semantic knowledge
	Zhang and Demner-Fushman (2017) Bag-of-words feature vector SVM classifier	No patient data	**Availability**: Public* **Size**: 2461 **Type**: 891 Specific; 1570 Varying **Additional annotation**: Trial labels	Precision: 0.90 Recall: 0.86 F2: 0.87	Cohort-specific model No real patient data considered Closest to keyword search (trials to cohort)

* Public source for clinical trials: https://clinicalTrials.gov

Table 4: Overview of classical systems covered in this survey: direction, methods, data sources and limitations.

Task	Method	Patient Source	Trial Source	Metrics (Best)	Limitations
Criterion-level prediction	Hamer et al. (2023) 1-shot prompt to LLM for criteria level prediction with explanation	**Availability**: Public (synthetic patient profile) **Size**: 10 **Mode**: Short text **Type**: Specific	**Availability**: Mixed (clinicalTrials.gov, EudraCT*) **Size**: 146 clinical trials **Type**: Specific **Additional Annotation**: Expert review	Criterion-level accuracy: 72% Trial-level precision: 0.71 Trial-level recall: 0.5 Workload reduction: 90% Stochasticity of precision / recall (10 runs): 0.03 / 0.02 SD	Majority of criterion-level errors due to incorrect reasoning (91%) which also occur for true positives
	Unlu et al. (2024) RAG approach	**Availability**: Private **Size**: 3,000 **Mode**: Textual clinical notes	**Availability**: Public** **Size**: 1 **Type**: Specific	Accuracy: 0.92 Correlation coeff: 0.81 Precision: 98.1% Recall: 92.3% Specificity: 93.9%	RAG pipeline evaluated on a single trial.
	Ferber et al. (2024) Sequential GPT-4o requests to create structured query to trial DB; remove irrelevant trials; and make criterion-wise boolean predictions	**Availability**: Public **Size**: 51 **Mode**: EHR	**Availability**: Public** **Size**: 15 **Type**: Specific	Trial Recall: 93.3% Accuracy: 92.7% to 97.8%	Refined evaluation with updated human judgment risks circular evaluation via target leakage or confirmation bias.
	Beattie et al. (2024) RAG approach with criterion-specific guidance prompts	**Availability**: Public (2018 N2C2: Cohort Selection)		Accuracy: 0.86 Sensitivity: 0.86 Specificity: 0.90 Precision: 0.87 Micro F1: 0.85	Best performance metrics is obtained on test subset (40 of 182). Expert guidance requires manual expertise for every criterion.
	Wornow et al. (2025) Zero-shot RAG approach with criteria modifications	**Availability**: Public (2018 N2C2: Cohort Selection)		Precision: 0.91 Recall: 0.92 Overall macro-F1: 0.81 Overall micro-F1: 0.93 Cost / patient: 0.87 USD to 11.88 USD API calls / patient: 1 to 57 Tokens (10^3) / patient: 8 to 103	67% of incorrect decision had correct reasoning indicating potential shortcuts taken by LLMs. Best strategy ICAN is 10 times more expensive than the least expensive strategy ACAN.
	Saeidi (2025) Few-shot LLM prompts with fine-tuned BERT-based concept embeddings of patient and criteria to predict binary eligibility labels.	**Availability**: Public (2018 N2C2: Cohort Selection)		Precision: 0.92 Recall: 0.93 Macro-F1: 0.83 Micro-F1: 0.94	Reported metrics are combined across N2C2 and TREC datasets which tackle different tasks. No ablation on components: concept matching, prompt design, fine-tuning embedding models

Continues to the next page.

Task	Method	Patient Source	Trial Source	Metrics (Best)	Limitations
Trial-level prediction	Wong et al. (2023) Few-shot prompts to LLM to generate structured forms of eligibility criteria using provided templates	**Availability**: Private **Size**: Unknown **Mode**: Structured **Additional Annotation**: 523 trial and patient historical labels; 68,485 trial and patient new labels	**Availability**: Public** **Size**: 53 **Type**: Specific	Recall (historical data): 76.8% Precision: 0.88 Recall: 67.3 F1: 76.1	Clinical Trial eligibility limited to the first 40 lines. Trial-level performance is very low (F1 score range:[29.6-48])
	Yuan et al. (2023) Prompt LLMs to reformulate eligibility criteria Train patient and criterion encoders contrastively on inclusion and exclusion criteria to predict trial-level match	**Availability**: Private **Size**: 825 **Mode**: Longitudinal text records	**Availability**: Public** **Size**: 6 **Type**: Specific **Additional Annotation**: 100,000 criterion-patient labels	Criterion-level: Precision: 0.96 Recall: 0.86 F1: 0.91 Trial-level: Precision: 0.80 Recall: 0.83 F1: 0.81	Criterion-level performance metrics 10 points higher than trial-level performance. High variance between different trials: F1 range: 0.48 - 0.98, even with trials concerning the same condition. Data separation methods for testing generalizability is unknown.
Trial ranking	Pradeep et al. (2022) Synthesize queries for initial trial retrieval. Fine-tune T5 to generate relevance label based on patient description and trial data	**Availability**: Public (TREC CT 2021)		NDCG@10: 0.71 P@10: 0.59 RR: 0.81	Zero-shot relevance ranking is only slightly better than BM25 with synthetic queries. Model prediction is difficult to interpret as signals come from trial condition, description and criteria.
	Zhuang et al. (2023) Hybrid sparse-dense retriever for top-1000. Cross-encoder re-ranker GPT4 for top-20 re-rank	**Availability**: Public (TREC CT 2023)		TREC CT 2022: P@10: 0.56 R@1000: 0.65 NDCG@10: 0.65 TREC CT 2023: P@10: 0.51 R@1000: 0.38 NDCG@10: 0.73	Precision (0.51) and recall (0.38) on TREC CT is very low. Validation on textual patient note does not transfer well to testing on structured patient note
	Embedding cosine similarity of patient and trial embeddings obtained using GPT Richmond and Deshpande (2023).	**Availability**: Public (TREC CT 2023)		MAP: 0.02	Direct embeddings of entire patient and trial documents are ineffective.
	Embedding cosine similarity of patient and trial embeddings obtained using Sentence Transformer by Lahiri et al. (2023).	**Availability**: Public (TREC CT 2023)		NDCG@10: 0.03 P@10: 0.09 MAP@10 and R@10 < 0.01	Direct embeddings of entire patient and trial documents are ineffective.

Continues to the next page.

Task	Method	Patient Source	Trial Source	Metrics (Best)	Limitations
	Kusa et al. (2023b) Sentence query formulation using GPT-3.5. Query enrichment using Kusa et al. (2023a). Trial-level prediction of re-ranked pairs using GPT-3.5.	**Availability**: Public (TREC CT 2023)		NDCG@10: 0.68 P@10: 0.58 RR: 0.65	Zero-shot LLM prompts provide marginal improvement over neural rerankers
	Peikos (2023) Utilize GPT3.5 to obtain trial-level labels for final re-ranking on top of lexical and neural re-rankers	**Availability**: Public (TREC CT 2023)		NDCG@10: 0.65 P@10: 0.44 RR: 0.58	Query processing is template-based and excludes negative information.
	Ferber et al. (2024) No-SQL query formulation from patient EHR using GPT-4o for initial retrieval, followed by vector embedding re-ranking.	**Availability**: Public **Size**: 51 **Mode**: EHR	**Availability**: Public** **Size**: 15 **Type**: Specific The authors consider an initial pool of 105,600 trials of which only 15 are annotated at trial- and criterion-level	%age of trials recalled at top-10: 93.3%	Refined evaluation with updated human judgment risks circular evaluation via target leakage or confirmation bias. Performance metrics used is very different to standard accuracy, precision and recall metrics.
	Rybinski and Karimi (2023) Multi-stage retriever with fine-tuned GPT3.5-turbo	**Availability**: Public (TREC CT 2021, 2022, 2023)		TREC CT 2023: NDCG@10: 0.73 P@10: 0.52 RR: 0.66	
	Rybinski et al. (2024) Multi-stage retriever using GPT3.5/GPT4 LLM-based relevance scoring and filtering Chain-of-thought (CoT) prompts with GPT4	**Availability**: Public (TREC CT 2021, 2022, 2023)		TREC CT 2023: NDCG@10: 0.78 P@10: 0.69 RR: 0.84	High latency of GPT4 trade-off for the performance boost Gains over (Rybinski and Karimi, 2023) is mainly due to reranker used (TCRR) Additional marginal gains via CoT of the larger GPT4 model.
	Jullien et al. (2024) LLM-guided basic attribute extraction for trial retrieval and filtering. Criterion-level LLM predictions fed to set-reasoning-based re-rank scoring functions.	**Availability**: Public (TREC CT 2022)		NDCG@10: 0.69 P@10: 0.73 P@25: 0.63 MRR: 0.86	LLMs underperform in exclusion criteria labeling.

Continues to the next page.

Task	Method	Patient Source	Trial Source	Metrics (Best)	Limitations
	Jin et al. (2024) (TrialGPT) Hybrid trial filtering with BM25 and MedCPT (Jin et al., 2023) Criterion-level LLM prediction aggregated to trial scores	**Availability**: Public (Koopman and Zuccon (2016), TREC CT 2021, 2022) **Additional annotation**: 1,015 criterion-patient labels		NDCG@10: 0.72 P@10: 0.66 AUROC (Excluding): 0.79 Time savings: 42.6% Reasoning correctness: 87.8% correct, 9.7% partially correct, 2.6% incorrect	The overall metric that averages over NDCG, P@10 and AUROC is meaningless Retrieval on prefiltered could bias results to be more favorable Longitudinal patient data not tested LLM aggregation assumes LLMs perform mathematical reasoning
	Nievas et al. (2024) Extends (Jin et al., 2024) to open-source LLMs	*(same as Jin et al. (2024))* **Additional annotation:** Patient sentence supporting eligibility label 500 criteria labels on: eligibility and difficulty.		NDCG@10: 66.3 P@10: 58.8 AUROC: 65.2 AURPC: 65.15 Implicit criterion-level accuracy (CLA): 68.7 Explicit CLA: 59.9 Win-rate: 68.9% Faithfulness (P/R/F1): *Exact scores not reported.*	Significantly high fine-tuning costs Reported aggregated score over metrics that span precision, recall, accuracy and AUC is meaningless.
	Datta et al. (2025) (Patient2Trial) Lexical retrievers use LLM generated query expansion. LLM predicts trial-level label and a criterion-level rationale with a matching score. Final ranking based on matching score.	**Availability**: Public (TREC CT 2023)		NDCG@10: 0.81 NDCG@30: 0.82 P@10: 0.73 P@30: 0.73 MRR: 0.78 Bpref: 0.30 P-precision: 0.24	Trials prefiltered by disorder-specific keywords. Manually curated disorder-specific instructions. Model predicted trial-level label is not evaluated
	Saeidi et al. (2023); Saeidi (2025) Embed patients and trials to concept vector space. Use variations of Equation 1 to compute relevance scores.	**Availability**: Public (TREC CT 2023)		Precision: 0.92 Recall: 0.93 Macro-F1: 0.83 Micro-F1: 0.94	Reported metrics are combined across N2C2 and TREC datasets. No direct connection between criterion-level LLM predictions and trial-level embeddings-based relevance score computations

* *EudraCT (European Union Drug Regulating Authorities Clinical Trials) is the European clinical trials database (EudraCT).*
** *Public source for clinical trials:* `clinicalTrials.gov`.

Table 5: Overview of LLM-based systems covered in this survey: tasks, methods, data sources used and limitations.