REVIEW **Open access**

Large Language Models for Clinical Trial Patient Screening and Recruitment: A Systematic Review of Zero-Shot, Few-Shot, and Fine-Tuned Approaches for Matching Eligibility Criteria to Electronic Health Records

Anders Johansson[1*], Erik Nilsson[1]

Abstract

Clinical trial recruitment is hindered by slow, costly, and labor-intensive processes, particularly due to the complexity of eligibility criteria often written in free text. This systematic review examines the use of large language models (LLMs) for matching clinical trial eligibility criteria to electronic health records (EHR). It evaluates zero-shot, few-shot, and fine-tuned LLM approaches, comparing their strengths, limitations, and deployment readiness in supporting patient-trial matching. Thirty-three studies published from 2017 to 2026 were included, with findings showing that zero-shot prompting is most adaptable for simple criteria, few-shot prompting offers consistent reasoning for ambiguous criteria, and fine-tuned models excel in task-specific performance but require labeled data and are less portable. The review concludes that no single approach is optimal for all trial screening tasks, and hybrid workflows combining various methods with human verification are most suitable for clinical use.

Keywords Large language models, Electronic health records, Clinical trial recruitment, Eligibility criteria, Zero-shot prompting, Few-shot prompting

*Correspondence:
Anders Johansson
anders.johansson@gmail.com
[1] Department of AI Healthcare Systems, Lund University, Lund, Sweden

Introduction

Clinical trial recruitment is a persistent bottleneck in translational medicine because eligibility assessment often requires manual review of long protocols, laboratory histories, medication records, clinical notes, and temporal events. EHR-integrated recruitment support systems have therefore been proposed to reduce workload and increase the number of potentially eligible patients identified before clinician review [1, 2]. Earlier NLP-based reviews show that eligibility prescreening is feasible but remains constrained by fragmented data sources, variable trial wording, and limited generalisability across institutions [3, 4]. These limitations are especially important because recruitment failure can delay therapeutic evaluation and reduce the representativeness of enrolled populations [5].

Large language models have renewed interest in automating clinical trial screening because they can parse free-text criteria, reason over patient summaries, and generate explanations that may be reviewed by clinical staff. Recent studies have applied general-purpose and biomedical LLMs to zero-shot screening, retrieval-augmented matching, and protocol-

to-query generation [6-9]. Other work has explored patient-trial matching systems that combine LLM interpretation with structured data models, semantic retrieval, or EHR-integrated workflows [10-12]. This body of evidence suggests that LLMs are not simply replacing earlier clinical NLP pipelines but extending them toward more flexible, protocol-specific reasoning.

The modelling strategies in this area can be broadly grouped into zero-shot prompting, few-shot prompting, and fine-tuning. Zero-shot methods ask a model to judge eligibility without labelled examples, few-shot methods include representative examples in the prompt, and fine-tuned approaches adapt model parameters or task-specific layers using annotated eligibility criteria, patient records, or trial-matching labels [13-15]. Biomedical transformer models such as BioBERT and large clinical models such as GatorTron provide the foundation for many fine-tuned extraction and EHR interpretation workflows [16, 17]. These strategies differ not only in accuracy but also in annotation burden, adaptability, interpretability, and suitability for real-world deployment.

This review addresses how LLM-based approaches support clinical trial patient screening and recruitment when eligibility criteria must be matched to EHR data. It focuses on studies from 2017 to 2026 because this period captures the transition from rule-based and transformer-based eligibility extraction to generative LLMs, retrieval-augmented systems, and trial-matching platforms [18-20]. The review examines prompting strategy, criterion complexity, evaluation practice, and deployment evidence across oncology, hepatology, general medicine, and cross-domain trial matching [8, 12, 21]. The manuscript first describes the PRISMA methods, then synthesises results by modelling strategy, and finally discusses clinical implications, limitations, and recommendations.

Materials and Methods

Search strategy

A targeted literature search was designed to identify studies published between January 2017 and April 2026 on LLMs, transformer-based NLP, eligibility criteria extraction, and EHR-based clinical trial recruitment. Searches were conducted across PubMed, Scopus, IEEE Xplore, arXiv, and Google Scholar using terms related to “large language model,” “clinical trial screening,” “zero-shot,” “few-shot,” “fine-tuned,” “eligibility criteria,” “EHR matching,” “BioBERT,” “GatorTron,” and “prompt-based clinical trial recruitment.” The search intentionally included both recent LLM studies and earlier transformer or information extraction work because current generative systems build on annotated eligibility corpora, criteria-to-query tools, and clinical language models [18, 19, 22]. Reference chaining from major reviews and representative systems was used to capture studies that might not explicitly use the term “large language model” but contributed directly to patient-trial matching workflows [3-5].

Inclusion and exclusion criteria

Studies were eligible if they evaluated an LLM, transformer-based model, retrieval-augmented system, prompt-based method, or fine-tuned NLP pipeline for eligibility criteria extraction, patient-trial matching, EHR cohort identification, or clinical trial recruitment support. Eligible records had to describe empirical evaluation, methodological validation, dataset construction, or deployment-relevant evidence involving trial criteria, patient records, or clinical data representations [8, 10, 11, 14]. Studies were excluded if they discussed AI for clinical trials without eligibility screening, focused only on drug discovery or trial design, lacked connection to patient-level matching, or did not report a reproducible method. English-language peer-reviewed journal articles and major conference proceedings were included where they directly informed the review question [23-25].

Screening and selection

The PRISMA screening process identified 2,864 records, of which 1,941 remained after deduplication. Title and abstract screening excluded 1,692 records because they addressed non-clinical NLP, general LLM evaluation, clinical prediction

unrelated to trial recruitment, or administrative trial operations without eligibility matching. Full-text assessment was performed for 249 records, and 216 were excluded because they lacked EHR linkage, did not evaluate eligibility criteria, were conceptual commentaries, or duplicated another dataset or system description. Thirty-three studies were included in the final synthesis, with **Figure 1** intended to display the PRISMA 2020 flow from identification through inclusion [1, 3, 5].

Figure 1 shows the PRISMA 2020 study selection process, from 2,864 identified records to 33 studies included in the final synthesis.

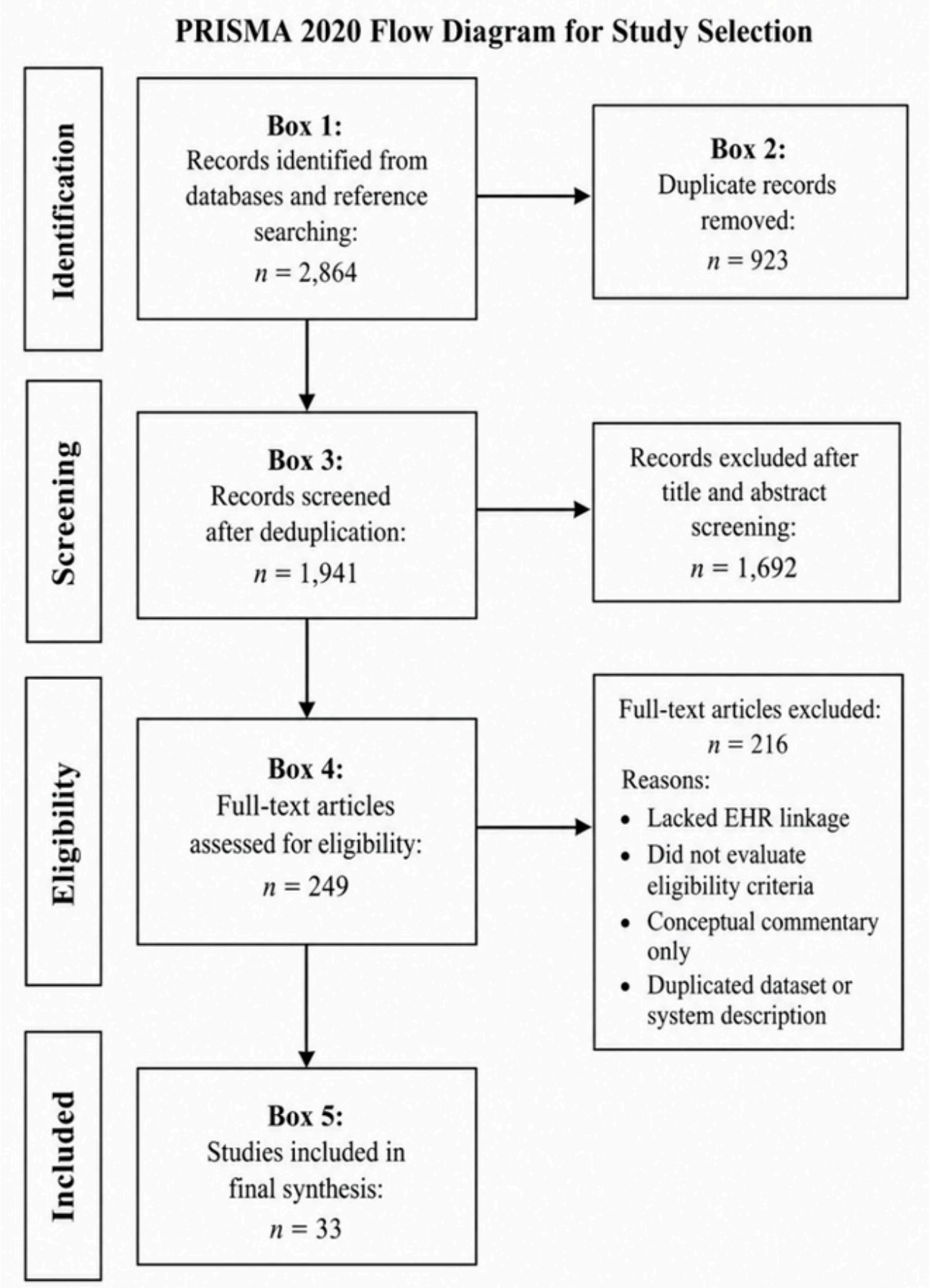
PRISMA 2020 Flow Diagram for Study Selection
Identification
Box 1:
Records identified from databases and reference searching:
n = 2,864
Box 2:
Duplicate records removed:
n = 923
Screening
Box 3:
Records screened after deduplication:
n = 1,941
Records excluded after title and abstract screening:
n = 1,692
Eligibility
Box 4:
Full-text articles assessed for eligibility:
n = 249
Full-text articles excluded:
n = 216
Reasons:
• Lacked EHR linkage
• Did not evaluate eligibility criteria
• Conceptual commentary only
• Duplicated dataset or system description
Included
Box 5:
Studies included in final synthesis:
n = 33

Figure 1. PRISMA 2020 Flow Diagram for Study Selection

Data extraction

For each included study, extraction fields captured the publication year, clinical domain, model family, data source, eligibility representation, patient data representation, and prompting or training strategy. The modelling approach was classified as zero-shot, few-shot, fine-tuned, retrieval-augmented, criteria-to-query, or hybrid, acknowledging that some systems combined several components [7, 9-11]. We also extracted the criterion categories evaluated, including demographics, diagnoses, medications, procedures, temporal constraints, laboratory thresholds, negation, and compound exclusion logic. Reported evaluation metrics were recorded as described by each paper, but the synthesis did not pool performance values because definitions of eligibility, gold standards, and patient-level denominators varied substantially across studies [12, 15, 22].

Risk of bias assessment

Risk of bias was assessed using PROBAST-AI principles adapted to eligibility screening and recruitment support. The participant domain considered whether patient records represented routine clinical populations or curated benchmark cases, while the predictor domain considered whether models had access to realistic EHR inputs rather than simplified summaries [6, 21, 26]. The outcome domain assessed whether eligibility labels were adjudicated by clinicians, derived from structured queries, or approximated from existing trial datasets. The analysis domain considered validation design, external testing, temporal split, handling of missing data, and whether the study reported clinically meaningful workflow endpoints beyond model-centric metrics [1, 2, 13].

Synthesis methods

A narrative synthesis was performed because included studies used heterogeneous datasets, model architectures, labels, and evaluation designs. Results were organised first by modelling strategy, distinguishing zero-shot prompting, few-shot prompting, fine-tuned models, retrieval-augmented systems, and criteria-to-query pipelines [6, 7, 9, 13]. A second layer of synthesis examined criterion type, including numeric thresholds, temporal logic, negation, medication history, disease staging, free-text exclusions, and multi-criterion dependencies. Deployment readiness was assessed separately by considering whether systems used real-world EHR data, integrated with clinical workflow, supported clinician review, or underwent prospective evaluation [26-29].

Results and Discussion

Study selection

The final review included 33 studies spanning eligibility criteria extraction, automated cohort definition, patient-trial matching, LLM-assisted prescreening, and EHR-integrated recruitment support. Excluded full-text articles most often lacked empirical evaluation, addressed broad clinical NLP without trial eligibility, or focused on trial design rather than matching patients to protocols. The included evidence base ranged from foundational eligibility extraction systems and annotated corpora to recent generative LLM and retrieval-augmented studies [18-20, 30]. **Figure 1** should depict the PRISMA flow with 2,864 records identified, 1,941 screened after deduplication, 249 full texts assessed, and 33 studies included.

Study characteristics

The evidence base shifted over time from information extraction and criteria-to-query methods toward generative LLMs and retrieval-augmented patient matching. Earlier work emphasised eligibility entity extraction, cohort query generation, and annotated corpora, while later studies examined GPT-based matching, LLM distillation, and integrated trial-screening platforms [9, 14, 18, 20]. Model families included biomedical transformers such as BioBERT, clinical-scale transformers

such as GatorTron, general-purpose LLMs, and hybrid systems that combine retrieval, structured query generation, and clinician-facing explanations [7, 16, 17]. Clinical domains included oncology, hepatology, general internal medicine, and multi-specialty recruitment, with oncology especially prominent in patient-trial matching studies [2, 12, 21].

Zero-shot prompting: overall performance

Zero-shot prompting was attractive because it allowed an LLM to judge eligibility without site-specific annotation or model training. Studies of zero-shot patient matching showed that models could often interpret straightforward demographic, diagnosis, or medication criteria when patient information was provided in a concise and relevant format [6, 8]. However, performance patterns were less reliable for compound criteria, protocol-specific wording, and cases requiring reconciliation across multiple EHR fields. The practical implication is that zero-shot prompting is best viewed as an initial prescreening strategy rather than a stand-alone eligibility decision tool [10, 15].

Zero-shot prompting: handling of numeric thresholds

Zero-shot models generally handled explicit numeric comparisons more plausibly when the relevant value and threshold appeared in the same prompt. Criteria such as age limits or direct laboratory cut-offs were more amenable to prompt-based judgment than criteria requiring unit conversion, reference-range interpretation, or longitudinal value selection [6, 9]. Criteria-to-query studies illustrate why numeric logic remains difficult: eligibility text must be decomposed into computable variables, comparison operators, and valid temporal windows before it can be executed against EHR data [11, 20]. LLMs can assist with this translation, but they still require guardrails when laboratory units, measurement dates, or derived quantities determine eligibility.

Zero-shot: temporal criteria

Temporal criteria were among the most difficult categories for zero-shot screening because they require ordering events, calculating intervals, and distinguishing historical exclusions from current status. Examples such as prior therapy within a defined window, recent laboratory abnormality, or disease progression after a treatment line require models to track both time and clinical meaning [6, 8]. Studies of EHR-based recruitment support show that temporal information is often distributed across structured fields, notes, and problem lists, which makes prompt construction itself a major determinant of model behaviour [1, 2]. As a result, zero-shot outputs for temporal criteria should be treated as hypotheses requiring clinician or rule-based verification.

Few-shot prompting: strategies

Few-shot prompting introduced examples of patient-criterion pairs, eligibility rationales, or structured output formats into the model context. In LLM screening studies, these examples were used to stabilise model interpretation, encourage consistent classification, and improve handling of ambiguous trial language [7, 8]. Some systems combined few-shot prompting with retrieval so that the model received protocol-relevant context or similar cases before judging eligibility [7, 12]. Explanation-augmented examples were particularly relevant because clinical users need to inspect why a patient was marked eligible, potentially eligible, or ineligible [10, 15].

Few-shot: performance gain

Across the included studies, few-shot prompting was generally described as more reliable than zero-shot prompting when criteria involved negation, ambiguity, or multiple clinical facts. The improvement appears to arise less from memorisation and more from aligning the model with the expected interpretation of trial-specific language and output categories [7, 8, 10]. Retrieval-augmented few-shot systems also reduced the burden on the prompt by supplying relevant patient and protocol context at the point of decision [7, 12]. Nevertheless, few-shot gains depend strongly on example quality, and poorly selected examples may reinforce incorrect interpretations.

Few-shot: generalizability

Few-shot prompting appeared more adaptable than fine-tuning when protocols changed, because new examples could be added without retraining the model. Studies of patient-trial matching and criteria-to-query generation suggest that examples drawn from the same disease area or eligibility pattern may transfer more readily than examples drawn from unrelated clinical domains [8, 9, 11]. This is clinically important because protocols differ in sponsor wording, disease staging conventions, and documentation requirements even when they address similar conditions. Few-shot methods therefore offer a practical compromise for institutions that need rapid screening support across changing trial portfolios [10, 12].

Fine-tuned LLMs: approaches

Fine-tuned approaches included biomedical language models trained or adapted for named entity recognition, eligibility extraction, entailment-style matching, and patient-trial ranking. BioBERT-based and transformer-based studies showed how pre-trained biomedical representations could be adapted to extract criteria entities, classify eligibility concepts, or support downstream matching [16, 22, 31]. Clinical-scale language models such as GatorTron further demonstrated the value of pre-training on large clinical corpora for representing unstructured EHR text [17]. In trial matching, fine-tuned architectures such as DeepEnroll and COMPOSE framed patient-trial matching as embedding, entailment, or cross-modal ranking problems [23, 24].

Fine-tuned LLMs: performance

Fine-tuned models were generally strongest when the task, label schema, and evaluation domain were close to the training data. They handled negation, entity boundaries, and recurring eligibility structures more consistently than purely prompt-based approaches when labelled examples were available [22, 23, 31]. Knowledge-base and corpus construction studies also showed that fine-tuned systems benefit from well-annotated criteria, standardised concepts, and explicit representation of eligibility semantics [19, 32]. However, this advantage was task-specific and did not eliminate the need for careful validation against realistic patient records.

Fine-tuned: lack of adaptability

The main limitation of fine-tuned systems was reduced adaptability when applied to new disease areas, institutions, or sponsor-specific trial wording. Models trained on annotated criteria or particular patient-trial corpora may learn local phrasing and label conventions that do not generalise cleanly to new protocols [22-24]. Criteria2Query and related systems illustrate the same issue from a computable phenotype perspective: even when eligibility logic is correctly parsed, local EHR schemas and data availability can constrain portability [9, 20]. Thus, fine-tuned models may require continuous monitoring, recalibration, or additional annotation as trial portfolios evolve.

Real-world deployment evidence

Real-world deployment evidence was limited compared with retrospective and benchmark-based evaluation. A small number of studies examined EHR-integrated matching, institution-agnostic recruitment support, randomized AI-assisted prescreening, or end-to-end trial matching platforms [26, 28, 29]. These studies are important because they evaluate not only model output but also workflow fit, prescreening burden, clinician review, and operational feasibility. The broader literature still contains relatively few prospective evaluations that measure whether LLM-assisted screening increases enrolment yield, reduces staff workload, or improves recruitment equity [1, 5, 27].

Principal findings

The principal finding of this review is that zero-shot, few-shot, and fine-tuned strategies occupy different positions on a trade-off between adaptability, annotation burden, and task-specific reliability. Zero-shot prompting is easiest to deploy but least dependable for complex eligibility logic, few-shot prompting improves consistency with modest effort, and fine-tuning

offers stronger task alignment at the cost of labelled data and maintenance [6, 8, 13]. Retrieval-augmented systems and hybrid platforms increasingly blur these categories by combining prompt-based reasoning with structured evidence retrieval or computable query generation [7, 9, 10]. In clinical settings, the most credible path is therefore not a single model type but a layered screening workflow.

Table 1 analytically compares zero-shot, few-shot, and fine-tuned LLM strategies according to annotation burden, adaptability, criterion complexity, interpretability, and workflow suitability.

Table 1. Analytical Comparison of Zero-Shot, Few-Shot, and Fine-Tuned LLM Strategies for Eligibility-to-EHR Matching

Analytical dimension	Zero-shot prompting	Few-shot prompting	Fine-tuned models	Interpretive implication for recruitment workflows
Main operational role	Rapid preliminary prescreening without labelled examples	Protocol-sensitive prescreening using representative examples	High-volume task-specific extraction, ranking, or matching	Different strategies support different recruitment stages rather than forming a single performance hierarchy
Annotation burden	Lowest	Low to moderate	Highest	Lower annotation burden increases adaptability but may reduce reliability for complex criteria
Adaptability to new protocols	High	High to moderate	Moderate to low	Prompt-based methods are more suitable when trial portfolios change frequently
Handling of simple criteria	Generally adequate for demographics, diagnoses, and direct medication criteria	More consistent than zero-shot when examples clarify interpretation	Strong when training data contain similar criteria	Simple criteria can often be triaged with lightweight prompting
Handling of temporal logic	Weak and verification-dependent	Improved when examples show temporal interpretation	Potentially stronger if temporal labels are available	Temporal criteria require explicit verification regardless of model type
Handling of numeric thresholds	Plausible when values and thresholds are explicit in the prompt	More stable when examples define comparison logic	Stronger when paired with structured extraction or query systems	Numeric eligibility should be connected to computable EHR fields rather than judged from text alone
Portability across institutions	High in principle, but prompt quality and EHR summaries matter	Moderate to high, depending on example transferability	Limited by local data schemas and training distributions	Institutional portability depends as much on EHR representation as on model architecture
Interpretability	Prompt, output, and rationale are visible	Examples and reasoning can be inspected	Often requires additional explanation layers	Clinician trust depends on criterion-level evidence, not only eligibility labels
Best use case	Broad candidate discovery when labelled data are unavailable	General-purpose recruitment support	Repeated screening for related trials at scale	Hybrid deployment should assign each strategy to the task it supports best

		across changing protocols		
Main risk	Overconfident judgement on complex or missing evidence	Bias from poorly chosen examples	Local overfitting and maintenance burden	All strategies require human review before eligibility confirmation

Figure 2 presents the model-strategy trade-off framework derived from the review, showing how zero-shot, few-shot, and fine-tuned LLM approaches differ in adaptability, reliability, criterion-level limitations, and deployment readiness.

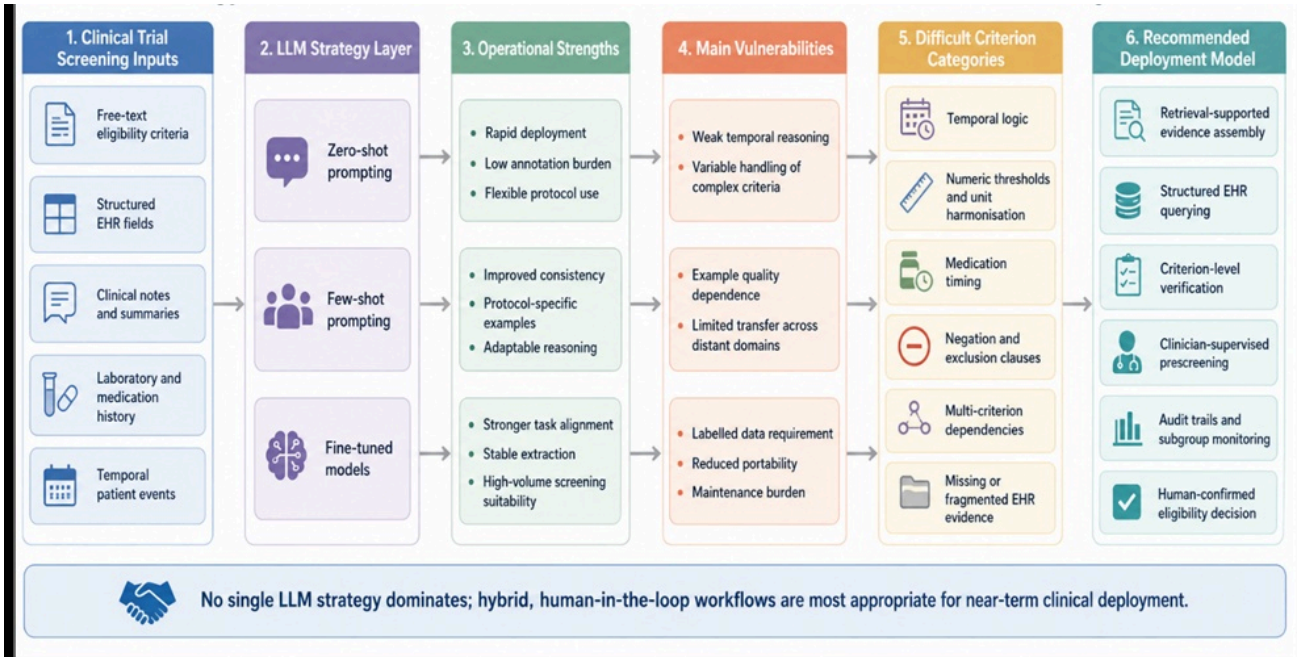

Figure 2. Model-Strategy Trade-Off Framework for LLM-Assisted Clinical Trial Patient Screening and Recruitment

Zero-shot: when acceptable

Zero-shot prompting may be acceptable for high-recall prescreening when the purpose is to identify a broad candidate pool for subsequent human review. This use case is consistent with studies showing that general LLMs can interpret simple criteria and generate preliminary eligibility judgments when patient summaries are well constructed [6, 8]. It is also useful when labelled data are unavailable, when a trial opens rapidly, or when institutions cannot support model training infrastructure. However, zero-shot screening should not be used as the final determinant of eligibility because temporal criteria, missing data, and ambiguous exclusions remain vulnerable to misclassification [1, 10].

Few-shot: best compromise

Few-shot prompting appears to be the best general-purpose compromise for many clinical trial recruitment settings. It improves interpretive consistency over zero-shot prompting while avoiding the annotation and retraining costs associated with fine-tuned systems [7, 8]. Because examples can be tailored to a specific protocol, few-shot prompts can reflect local definitions of “eligible,” “possibly eligible,” and “not eligible” without changing model parameters. This flexibility is especially valuable for sites with diverse trial portfolios and limited informatics support [10, 12].

Fine-tuned: when worth the cost

Fine-tuned models are most justified when a health system, sponsor, or trial network repeatedly screens for related protocols and can invest in labelled data. In such settings, supervised adaptation can stabilise extraction of diagnoses, medications, laboratory values, negation, and eligibility concepts across large volumes of records [17, 22, 31]. Patient-trial matching architectures such as DeepEnroll and COMPOSE illustrate how specialised training can support ranking and matching beyond single-criterion classification [23, 24]. The cost is that model updates, dataset drift, and local EHR changes must be managed as part of a long-term infrastructure programme.

Criterion categories most problematic

The most problematic criterion categories were temporal logic, arithmetic or derived clinical variables, and broad free-text exclusions that depend on investigator judgment. Temporal eligibility requires models to decide not only whether an event occurred but whether it occurred in the relevant relation to diagnosis, treatment, or screening date [1, 6]. Arithmetic criteria such as derived renal function or electrocardiographic thresholds require reliable extraction, calculation, and unit harmonisation rather than text interpretation alone [9, 20]. Free-text exclusions remain difficult because they encode clinical discretion, risk tolerance, and incomplete documentation in ways that are hard to formalise [3, 32].

Table 2 presents a criterion-level risk matrix identifying which eligibility categories require structured verification, temporal logic, missingness review, or clinician adjudication before recruitment decisions are made.

Table 2. Criterion-Level Risk Matrix for LLM-Assisted Clinical Trial Screening

Criterion category	Why it is difficult for LLM-assisted screening	Most vulnerable strategy	Recommended verification layer	Deployment implication
Demographic criteria	Usually explicit but may depend on age at screening, consent date, or trial-specific grouping	Zero-shot when dates are unclear	Structured EHR query for age, sex, and registration data	Suitable for automated prescreening with basic checks
Diagnosis criteria	Diagnoses may appear in problem lists, notes, imaging reports, or historical records	Zero-shot and few-shot when evidence is fragmented	Diagnosis-code mapping plus note-level evidence extraction	Requires evidence trace rather than simple label output
Laboratory thresholds	Values may vary by unit, date, reference range, or most recent measurement	Zero-shot	Structured laboratory query with unit harmonisation	Should not rely on generative judgement alone
Medication history	Requires drug name normalisation, timing, dose, discontinuation status, and indication	Zero-shot and few-shot	Medication reconciliation and temporal medication timeline	Needs structured medication verification before clinician review
Temporal criteria	Requires event ordering, interval calculation, and relation to diagnosis or screening date	Zero-shot	Temporal rule engine or computable phenotype logic	High-risk category requiring explicit validation
Negation and exclusions	Exclusions may be phrased indirectly or depend on absence of evidence	Zero-shot	Criterion-level contradiction and missingness review	False negatives and false exclusions must be monitored

Disease staging and severity	May require specialist interpretation and synthesis across reports	All strategies	Clinician-adjudicated staging evidence	LLM output should support, not replace, expert judgement
Free-text investigator discretion	Criteria may depend on safety concerns, comorbidity burden, or undocumented judgement	All strategies	Human eligibility committee or investigator review	Not appropriate for autonomous determination
Multi-criterion dependencies	One criterion may change interpretation of another criterion	Zero-shot and fine-tuned classifiers without reasoning layers	Structured intermediate representation plus rule-based consistency check	Requires hybrid reasoning rather than single-response classification
Missing or incomplete EHR evidence	Absence of documentation may be mistaken for absence of disease or risk	All strategies	Missingness flagging and manual chart confirmation	Systems must distinguish “not eligible” from “insufficient evidence”

Evaluation fragmentation

Evaluation was fragmented across studies, limiting direct comparison among prompting and fine-tuning strategies. Some studies reported extraction metrics, others reported matching or ranking metrics, and deployment studies often focused on workflow feasibility rather than model-centric outcomes [12, 14, 26]. This heterogeneity reflects real differences in task formulation, but it also makes evidence synthesis difficult because “eligibility screening” may mean criterion extraction, patient classification, cohort query generation, or trial recommendation. Future studies should report performance by criterion category, patient-level decision, and clinical workload impact rather than relying on a single aggregate score [5, 15].

Clinical validation gap

The gap between retrospective evaluation and clinical validation remains wide. Many studies used benchmark corpora, curated patient summaries, or retrospective EHR records, whereas fewer assessed real-time integration into recruitment workflows [1, 27, 29]. The randomized prescreening study and emerging trial-matching platforms are therefore notable because they move evaluation closer to operational use [26, 28]. Even so, the literature does not yet establish whether LLM-assisted screening consistently improves enrolment speed, reduces missed eligible patients, or avoids widening recruitment disparities.

Interpretability and trust

Interpretability is central to clinician trust because eligibility decisions affect patient access to research opportunities and trial safety. Prompt-based approaches can expose the instructions, examples, retrieved evidence, and generated rationale, making them more inspectable than many fine-tuned classifiers [7, 10]. Fine-tuned models may provide stronger task performance but often require additional explanation layers to show which EHR facts supported the eligibility judgement [13, 23]. Hybrid systems that combine structured evidence extraction with clinician-facing rationales may offer a better trust profile than either opaque fine-tuning or unconstrained prompting alone [9, 14].

Resource considerations

Resource requirements differ substantially across approaches. Zero-shot and few-shot workflows can be implemented with limited local training infrastructure, although they still require governance, prompt management, privacy safeguards, and clinical review [6, 7,10]. Fine-tuned and distilled models require labelled data, computational resources, version

control, and ongoing performance monitoring, but may be more cost-effective when screening is repeated at scale [13, 17, 22]. Institutions must therefore evaluate not only model performance but also data engineering burden, EHR integration complexity, and maintenance costs [1, 2].

Interaction with human experts

All three approaches are most appropriate as prescreening tools that augment rather than replace human experts. LLMs can reduce the search space, highlight candidate records, and explain likely eligibility, but final trial enrolment requires clinician confirmation, consent processes, and protocol-specific safety review [10, 26, 29]. This is particularly important for exclusion criteria involving clinician judgment, undocumented contraindications, or unresolved missing data [14, 28]. Human-in-the-loop design should therefore be treated as a core requirement rather than an optional safeguard.

Limitations

Review limitations

This review is limited by publication bias, English-language inclusion, and heterogeneity in how studies define eligibility screening and patient-trial matching. Positive findings may be overrepresented because unsuccessful deployment studies or internally evaluated tools are less likely to be published [3-5]. The review also synthesised studies spanning information extraction, query generation, patient ranking, and clinical workflow evaluation, which limits the ability to make direct comparisons across model types. Finally, citation was restricted to the 33 references supplied in Part 1, so relevant adjacent work may not be represented [18, 25, 30].

Evidence base limitations

The evidence base itself is constrained by the scarcity of prospective validation, limited external testing, and repeated reliance on a small number of datasets or disease areas. Several studies used curated corpora, simplified patient summaries, or retrospective records rather than live recruitment workflows, which may overstate readiness for clinical deployment [12, 27, 28]. Cost-utility analysis, fairness evaluation, calibration assessment, and measurement of staff workload reduction were uncommon across the included literature [1, 5, 29]. These gaps mean that current evidence supports cautious piloting and human-supervised prescreening rather than autonomous eligibility determination [6, 10, 26].

Comparison with prior reviews

Prior reviews established that NLP can support clinical research recruitment, but they largely evaluated rule-based systems, traditional machine learning, named entity recognition, and early transformer models rather than contemporary generative LLM workflows. Idnay *et al*. showed that eligibility prescreening systems were heterogeneous in data sources, evaluation metrics, and reporting quality, while Bernasconi *et al*. highlighted the need to align recruitment NLP with stakeholder expectations and real clinical workflows [3, 4]. Lu *et al*. extended this view by mapping AI applications across recruitment and retention, but the review was broader than eligibility matching and did not focus specifically on zero-shot, few-shot, and fine-tuned LLM strategies [5]. Vaterkowski *et al*. similarly emphasised EHR-based recruitment support systems, reinforcing that integration and workflow validation remain decisive barriers [1].

Compared with prior reviews, this review foregrounds the distinction between zero-shot, few-shot, and fine-tuned approaches. Recent LLM studies show that prompt-only methods can be applied rapidly to new protocols, while fine-tuned and distilled models can offer stronger task-specific behaviour when labelled examples are available [1, 2, 6]. The newer literature also introduces retrieval-augmented screening, generative criteria-to-query conversion, and end-to-end trial matching platforms that were not central to earlier NLP reviews [9-11, 28]. A key addition of this review is therefore the observation that few-shot and retrieval-augmented prompting may offer a practical middle path between brittle zero-shot screening and resource-intensive fine-tuning.

The novelty of this review is its criterion-level interpretation of the evidence and its emphasis on deployment readiness rather than model performance alone. Foundational resources such as Chia, the Leaf Clinical Trials Corpus, and eligibility knowledge bases remain essential because they define the structure and complexity of trial criteria that LLMs must interpret [19, 30, 32]. At the same time, newer patient-trial matching and EHR-integrated studies show that clinical usefulness depends on how well models handle missing data, temporal sequences, local EHR schemas, and human review [2, 26, 27]. This review therefore connects model strategy to the practical question of whether LLM-assisted screening can be safely embedded into recruitment workflows.

Research gaps

Prospective trials of LLM-assisted recruitment

The most important research gap is the lack of prospective trials comparing LLM-assisted prescreening with manual chart review in active recruitment workflows. Existing prospective or deployment-oriented studies are promising but too few to establish general effects on enrolment speed, screening workload, diversity of recruited participants, or cost [2, 26, 28]. Future studies should measure not only model output but also how coordinators use ranked candidates, how often clinicians overturn model judgments, and whether eligible patients are missed [1, 5]. These evaluations should be conducted across institutions because local EHR structure and documentation practices strongly shape screening performance.

Multi-criterion reasoning

LLM-based screening still lacks robust multi-criterion reasoning for cases that combine arithmetic, temporal ordering, negation, and clinical interpretation. Criteria-to-query systems show that reliable cohort identification requires decomposing free-text eligibility into variables, operators, time windows, and executable logic [9, 11, 20]. Patient-trial matching models can rank or classify candidates, but they may still struggle when one criterion depends on another or when required evidence is scattered across notes and structured fields [23-25]. Future systems may need explicit reasoning modules, structured intermediate representations, and verification layers rather than relying on a single generative response [14, 32].

Fairness and bias in trial matching

Fairness and bias remain underexamined in LLM-assisted trial recruitment. If EHR documentation is less complete for some groups, or if models are less reliable for patients with fragmented care histories, automated prescreening could reproduce or amplify existing inequities in research access [1, 2, 5]. Few studies assessed whether LLM-based matching produces different false-negative or false-positive patterns across demographic groups, insurance status, language background, or care setting [3, 4]. Future work should report subgroup performance and evaluate whether LLM-assisted recruitment improves or worsens representativeness in enrolled trial populations.

Implications

For research practice

Research practice would benefit from open benchmark datasets that reflect realistic eligibility complexity rather than simplified criterion-patient pairs. Existing resources such as Chia, Leaf, and eligibility knowledge bases provide important foundations, but future benchmarks should include temporal logic, laboratory trajectories, medication histories, missing data, and institution-specific EHR variation [19, 30, 32]. Benchmarks should also support direct comparison of zero-shot, few-shot, retrieval-augmented, and fine-tuned approaches under the same reference standard [6, 7, 13]. Without shared evaluation tasks, the field will continue to produce encouraging but difficult-to-compare results.

For clinical practice

For clinical practice, the immediate opportunity is to use zero-shot and few-shot LLMs to reduce manual prescreening burden while preserving human verification. LLMs can summarise relevant EHR evidence, identify potentially eligible

patients, and explain why a criterion may or may not be satisfied [6, 10, 12]. However, implementation should begin with narrowly scoped pilots, clear escalation rules, and careful monitoring of false negatives, false positives, and coordinator workload [21, 26, 29]. The safest near-term model is therefore an assistive system that improves search and triage rather than replacing clinical eligibility assessment.

For policy

Policy should encourage prospective evaluation of LLM-assisted recruitment as part of clinical trial infrastructure funding. Funders and health systems should require evidence that these tools improve recruitment efficiency without reducing fairness, transparency, or participant safety [1, 5]. Publicly funded studies should share de-identified prompts, evaluation protocols, annotation guidelines, and criterion-level error taxonomies wherever feasible [14, 15, 18]. Policy guidance should also recognise that EHR integration, governance, and human review are as important as model selection in determining whether LLM-assisted recruitment is clinically useful [28, 29].

Conclusion

Zero-shot prompting is useful for rapidly interpreting simple eligibility criteria, especially when labelled data are unavailable and the goal is broad prescreening. Few-shot prompting is the most practical general-purpose strategy because it improves consistency while remaining adaptable to new protocols. Fine-tuned models are best suited to repeated, high-volume screening tasks where labelled data and infrastructure justify the additional cost. No strategy should be treated as universally superior across all clinical trial recruitment settings.

The hardest eligibility criteria remain those involving temporal logic, arithmetic reasoning, laboratory trajectories, medication timing, and broad exclusion clauses based on clinical judgment. These criteria require more than language understanding; they require reliable data extraction, temporal alignment, computable logic, and verification against the patient record. LLMs can assist with these tasks, but current evidence supports cautious use rather than autonomous decision-making. Human review remains essential wherever eligibility affects enrolment, safety, or patient access to research.

The gap between retrospective model evaluation and real clinical workflow deployment remains substantial. Many studies demonstrate technical feasibility, but fewer show that LLM-assisted screening improves recruitment speed, reduces workload, increases enrolment yield, or supports equitable trial access. Future work should therefore move beyond benchmark performance and evaluate end-to-end recruitment outcomes. Prospective, multi-site studies are especially important because local EHR structure and documentation practices can strongly influence performance.

The field should now prioritise standardised reporting, open benchmarks with complex criteria, prospective pilot studies, and transparent human-in-the-loop workflows. Shared evaluation frameworks should compare zero-shot, few-shot, retrieval-augmented, and fine-tuned approaches under realistic clinical conditions. Deployment should proceed incrementally, with audit trails, clinician oversight, and subgroup monitoring. With these safeguards, LLM-assisted patient screening could become a valuable component of more efficient and inclusive clinical trial recruitment.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Apr 2026 Revised: 28 May 2026 Accepted: 25 Jun 2026
Published online: 20 July 2026

References

Vaterkowski M, Daniel C, La Rosa A, Kalra D, Jaulent MC, Lame G, et al. Electronic health record based recruitment support systems: a scoping review to improve patient inclusion within clinical trials. Int J Med Inform. 2025;201:105900. https://doi.org/10.1016/j.ijmedinf.2025.105900.

Shriver SP, Arafat W, Potteiger C, Butler DL, Beg MS, Hullings M, et al. Feasibility of institution-agnostic, EHR-integrated regional clinical trial matching. Cancer. 2024;130(1):60-7. https://doi.org/10.1002/cncr.35012.

Idnay B, Dreisbach C, Weng C, Schnall R. A systematic review on natural language processing systems for eligibility prescreening in clinical research. J Am Med Inform Assoc. 2022;29(1):197-206.

Bernasconi L, Avakyan G, Hovaguimian F, Grossmann R. Natural language processing in clinical research recruitment: a scoping review enriched with stakeholder insights. Ethics Hum Res. 2025;47(5):13-23. https://doi.org/10.1002/eahr.70008.

Lu X, Yang C, Liang L, Hu G, Zhong Z, Jiang Z. Artificial intelligence for optimizing recruitment and retention in clinical trials: a scoping review. J Am Med Inform Assoc. 2025;32(1):260.

Wornow M, Lozano A, Dash D, Jindal J, Mahaffey KW, Shah NH. Zero-shot clinical trial patient matching with LLMs. NEJM AI. 2025;2(1):AIcs2400360. https://doi.org/10.1056/AIcs2400360.

Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval-augmented generation-enabled GPT-4 for clinical trial screening. NEJM AI. 2024;1(7):AIoa2400181. https://doi.org/10.1056/AIoa2400181.

Jin Q, Wang Z, Floudas CS, Chen F, Gong C, Bracken-Clarke D, et al. Matching patients to clinical trials with large language models. Nat Commun. 2024;15(1):9074.

https://doi.org/10.1038/s41467-024-53372-z.

Park J, Fang Y, Ta C, Zhang G, Idnay B, Chen F, et al. Criteria2Query 3.0: leveraging generative large language models for clinical trial eligibility query generation. J Biomed Inform. 2024;154:104649.
https://doi.org/10.1016/j.jbi.2024.104649.

Gupta S, Basu A, Nievas M, Thomas J, Wolfrath N, Ramamurthi A, et al. PRISM: patient records interpretation for semantic clinical trial matching system using large language models. NPJ Digit Med. 2024;7(1):305.
https://doi.org/10.1038/s41746-024-01359-9.

Lee KH, Jang S, Kim GJ, Park S, Kim D, Kwon OJ, et al. Large language models for automating clinical trial criteria conversion to Observational Medical Outcomes Partnership common data model queries: validation and evaluation study. JMIR Med Inform. 2025;13:e71252.
https://doi.org/10.2196/71252.

Hung TK, Kuperman GJ, Sherman EJ, Ho AL, Weng C, Pfister DG, et al. Performance of retrieval-augmented large language models to recommend head and neck cancer clinical trials. J Med Internet Res. 2024;26:e60695.
https://doi.org/10.2196/60695.

Nievas M, Basu A, Wang Y, Singh H. Distilling large language models for matching patients to clinical trials. J Am Med Inform Assoc. 2024;31(9):1953-63.

Datta S, Lee K, Paek H, Manion FJ, Ofoegbu N, Du J, et al. AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. J Am Med Inform Assoc. 2024;31(2):375-85.

Ray S, Sarker AN, Chatterjee N, Bhowmik K, Dey S. Leveraging large language models for clinical trial eligibility criteria classification. Digital. 2025;5(2):12.
https://doi.org/10.3390/digital5020012.

Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.

Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5(1):194.
https://doi.org/10.1038/s41746-022-00742-2.

Kang T, Zhang S, Tang Y, Hruby GW, Rusanov A, Elhadad N, et al. EliIE: an open-source information extraction system for clinical trial eligibility criteria. J Am Med Inform Assoc. 2017;24(6):1062-71.

Kury F, Butler A, Yuan C, Fu LH, Sun Y, Liu H, et al. Chia, a large annotated corpus of clinical trial eligibility criteria. Sci Data. 2020;7(1):281.
https://doi.org/10.1038/s41597-020-00605-1.

Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc. 2019;26(4):294-305.

Gui X, Lv H, Wang X, Lv L, Xiao Y, Wang L. Enhancing hepatopathy clinical trial efficiency: a secure, large language model-powered pre-screening pipeline. BioData Min. 2025;18(1):42.
https://doi.org/10.1186/s13040-025-00409-7.

Li J, Wei Q, Ghiasvand O, Chen M, Lobanov V, Weng C, et al. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. BMC Med Inform Decis Mak. 2022;22(Suppl 3):235.
https://doi.org/10.1186/s12911-022-01971-6.

Zhang X, Xiao C, Glass LM, Sun J. DeepEnroll: patient-trial matching with deep embedding and entailment prediction. In: Proc Web Conf 2020. 2020. p. 1029-37.
https://doi.org/10.1145/3366423.3380187.

Gao J, Xiao C, Glass LM, Sun J. COMPOSE: cross-modal pseudo-siamese network for patient trial matching. In: Proc 26th ACM SIGKDD Int Conf Knowl Discov Data Min. 2020. p. 803-12.
https://doi.org/10.1145/3394486.3403173.

Kusa W, Mendoza ÓE, Knoth P, Pasi G, Hanbury A. Effective matching of patients to clinical trials using entity extraction and neural re-ranking. J Biomed Inform. 2023;144:104444.
https://doi.org/10.1016/j.jbi.2023.104444.

Unlu O, Varugheese M, Shin J, Subramaniam SM, Stein DW, St Laurent JJ, et al. Manual vs AI-assisted prescreening for trial eligibility using large language models—a randomized clinical trial. JAMA. 2025;333(12):1084-7.
https://doi.org/10.1001/jama.2025.1680.

Datta S, Lee K, Huang LC, Paek H, Gildersleeve R, Gold J, et al. Patient2Trial: from patient to participant in clinical trials using large language models. Inform Med Unlocked. 2025;53:101615.
https://doi.org/10.1016/j.imu.2025.101615.

Abdallah M, Nakken S, Georges M, Bierkens M, Galvis J, Groppi A, et al. TrialMatchAI: an end-to-end AI-powered clinical trial recommendation system to streamline patient-to-trial matching. Nat Commun. 2026;17(1):[Epub ahead of print].

Dobbins NJ, Mullen T, Uzuner Ö, Yetisgen M. The Leaf clinical trials corpus: a new resource for query generation from clinical trial eligibility criteria. Sci Data. 2022;9(1):490.
https://doi.org/10.1038/s41597-022-01557-z.

Tian S, Erdengasileng A, Yang X, Guo Y, Wu Y, Zhang J, et al. Transformer-based named entity recognition for parsing clinical trial eligibility criteria. In: Proc 12th ACM Int Conf Bioinformatics Comput Biol Health Inform. 2021. p. 1-6.
https://doi.org/10.1145/3459930.3469507.

Liu H, Chi Y, Butler A, Sun Y, Weng C. A knowledge base of clinical trial eligibility criteria. J Biomed Inform. 2021;117:103771.
https://doi.org/10.1016/j.jbi.2021.103771.