Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Large Language Models Across the Clinical Trial Lifecycle: A Scoping Review of Applications, Methodologies, and Validation Gaps

Huixue Zhou
University of Minnesota
Xinyu Zhou
Northwestern University
Shuang Zhou
University of Minnesota
Meijia Song
University of Minnesota
Xiaoyi Chen
University of Minnesota
Yifan Wu
University of Minnesota
Yiyu Chen
University of Minnesota
Kai Yu
University of Minnesota
Yongkang Xiao
University of Minnesota
Yu Hou
University of Minnesota
Zaifu Zhan
University of Minnesota
Xiangming Zhan
Arizona State University
Han Yang
University of Minnesota
Shufan Ming
University of Illinois Urbana-Champaign
Halil Kilicoglu
University of Illinois Urbana-Champaign
Yuan Luo

University of Washington

Rui Zhang

ruizhang@umn.edu

University of Minnesota

Article

Keywords: Large Language Models, Clinical Trials, Generative AI, Scoping Review, Natural Language Processing

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11343754/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Large Language Models Across the Clinical Trial Lifecycle: A Scoping Review of Applications, Methodologies, and Validation Gaps

Huixue Zhou [†1], Xinyu Zhou [†2], Shuang Zhou[1], Meijia Song[1], Xiaoyi Chen[1], Yifan Wu[1], Yiyu Chen[1], Kai Yu[1], Yongkang Xiao[1], Yu Hou[1], Zaifu Zhan[1], Xiangming Zhan[3], Han Yang[1], Shufan Ming[4], Halil Kilicoglu[4], Yuan Luo[2,5], Rui Zhang[1*]

[1]University of Minnesota .
[2]Northwestern University .
[3]Arizona State University .
[4]University of Illinois Urbana-Champaign .
[5]University of Washington, .

*Corresponding author(s). E-mail(s): ruizhang@umn.edu;

Abstract

Large language models (LLMs) are increasingly being incorporated into clinical trial research, yet evidence of their utility is fragmented across isolated tasks, trial stages, model architectures, and evaluation strategies. This fragmentation makes it difficult to determine where LLMs provide meaningful value, where evidence remains preliminary, and what barriers limit their translation into routine clinical research. We therefore systematically characterized generative LLM applications across the clinical trial lifecycle, defined here as the interconnected stages of trial design, participant recruitment, trial conduct, and data analysis. We reviewed 71 studies published between January 2023 and August 2025 and examined their applications, technical approaches, and validation strategies. LLM applications were concentrated in Analysis (40.8%) and Recruitment (31.0%), followed by Design (18.3%) and Conduct (5.6%). Among all studies, 34 (47.9%) reported better performance than traditional machine-learning or previous language-model comparators, whereas 32 (45.1%) did not include such comparisons. Human-based evaluation was reported in 25 studies (35.2%), but prospective and operational validation remained limited. By mapping applications and evidentiary maturity across the full clinical trial lifecycle, this review identifies a substantial gap between rapidly expanding technical capabilities and

evidence supporting real-world implementation. These findings provide a structured evidence base for prioritizing rigorous validation, transparent governance, and clinically meaningful evaluation as LLMs transition from experimental tools toward operational components of clinical research.

Keywords: Large Language Models, Clinical Trials, Generative AI, Scoping Review, Natural Language Processing

1 Introduction

Clinical trials, the cornerstone of evidence-based medicine, remain the gold standard for evaluating medical interventions [1]. These studies require substantial financial investment, multidisciplinary expertise, and long-term coordination across sites and stakeholders. Despite their scientific rigor, the labor- and cost-intensive nature of clinical trials continues to pose challenges to scalability and accessibility. Limited funding, complex regulatory requirements, and high personnel demands can constrain trial initiation and completion, particularly in resource-limited settings [2, 3].

Artificial intelligence (AI) has long been explored as a means of improving the efficiency of clinical research. More recently, the emergence of large language models (LLMs) has expanded the range of tasks that can be addressed using AI across clinical trial workflows. Unlike conventional machine learning approaches designed for specific prediction or classification tasks, LLMs can interpret, generate, and transform natural language across diverse biomedical contexts and can perform new tasks through prompting without task-specific model training [4]. These capabilities make LLMs particularly relevant to the text-intensive components of clinical research, including protocol drafting, eligibility criteria processing, patient-to-trial matching, participant communication, and analysis of trial-related documents [5].

The growing application of generative LLMs in clinical research has produced a rapidly expanding but fragmented literature spanning different stages of the clinical trial lifecycle [6]. Here, the clinical trial lifecycle refers to the interconnected stages of trial design, participant recruitment, trial conduct, and data analysis and reporting. Despite rapid growth in this field, existing studies are distributed across disparate clinical research tasks and vary substantially in their data sources, model implementation, evaluation strategies, and degree of human-based validation [6, 7]. This heterogeneity makes it difficult to determine where LLMs are being most extensively applied, which stages remain underexplored, whether findings are comparable across studies, and whether reported technical performance translates into evidence supporting real-world clinical research.

Existing reviews have provided important but complementary perspectives on this landscape, including AI-based trial outcome prediction [8], LLM-assisted clinical trial recruitment and patient-to-trial matching [9, 10], LLM-based assessment of clinical trial reporting quality [11], clinical trials evaluating LLM-based interventions [12], and broader applications, technical advances, and implementation challenges of LLMs across clinical trial workflows [6]. However, a systematic characterization that jointly

examines how LLMs are distributed and applied across the full clinical trial lifecycle, together with the technologies and validation strategies supporting these applications, remains limited. Such a lifecycle-wide perspective is important because examining individual applications in isolation makes it difficult to determine whether technological development is progressing consistently across the clinical research workflow or whether substantial gaps persist between technical capability and real-world validation [6, 13]. Systematically mapping these dimensions across the full lifecycle can therefore identify areas of concentrated development, underexplored trial stages, and methodological and validation gaps, providing a more integrated evidence base for guiding future research and the responsible integration of LLMs into clinical trials.

This scoping review provides a comprehensive analysis of LLM applications across the clinical trial lifecycle. Figure 1 provides an overview of the clinical trial stages, data characteristics, privacy considerations, LLM families, and methodological approaches examined in this review. We synthesize the available evidence to address three research questions:

1. **Utility:** How are LLMs applied across the clinical trial lifecycle, including trial design, participant recruitment, trial conduct, and data analysis?
2. **Technology:** What LLM architectures and enhancement strategies are used to support these applications?
3. **Validation:** To what extent have LLM-based approaches been evaluated against traditional computational methods and through human-based evaluation?

In parallel, we characterize the data sources, accessibility, sample sizes, and privacy and governance practices of the included studies to contextualize the environments in which these systems were developed and evaluated. By jointly examining utility, technology, data context, and validation within a common lifecycle framework, this review provides an integrated evidence map of the field, identifies underexplored clinical trial functions and methodological and validation gaps, and highlights priorities for the rigorous evaluation and responsible integration of LLMs into clinical research workflows.

2 Results

The database search yielded a total of 2,682 records from sources including PubMed, Web of Science, Cochrane Library, EMBASE, Scopus, IEEE Xplore, and the ACM Digital Library. After removing 913 duplicates, 1,769 unique titles and abstracts were screened. Following the exclusion of 1,577 records during title and abstract screening, 192 reports were sought for retrieval. Six reports could not be retrieved, leaving 186 full-text reports assessed for eligibility. Of these, 115 were excluded, and 71 studies were included in the review (Figure 5).

2.1 Applications of LLMs Across the Clinical Trial Lifecycle

Across the 71 eligible studies, the study was used as the primary unit of analysis, with each study assigned to a single clinical trial phase based on its primary application. Studies that explicitly addressed more than one phase were classified as Multiple rather

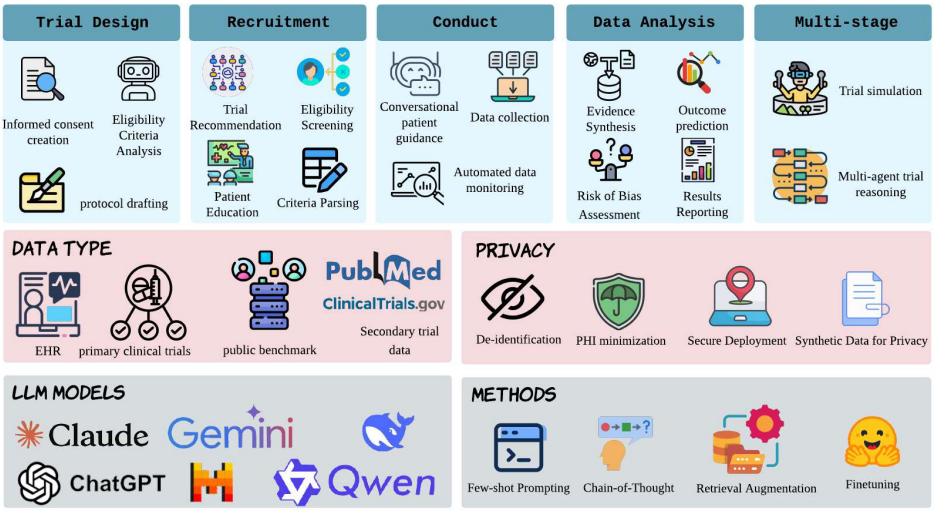


Fig. 1 Overview of the clinical trial lifecycle and key dimensions used to characterize LLM applications in this review. The upper panel summarizes representative LLM applications across trial Design, Recruitment, Conduct, Analysis, and Multi-stage workflows. The lower panels illustrate the data sources, privacy and governance considerations, LLM families, and methodological approaches examined across the included studies.

than counted separately within each phase. As shown in Figure 3, Analysis was the most frequently represented phase ($n = 29$, 40.8%), followed by Recruitment ($n = 22$, 31.0%), Design ($n = 13$, 18.3%), Conduct ($n = 4$, 5.6%), and Multiple ($n = 3$, 4.2%).

The distribution of included studies by publication year and clinical trial phase is shown in Figure 3b. Although the predefined eligibility window began in 2022, no studies published in 2022 met the final inclusion criteria. Included studies therefore spanned 2023 to 2025. Studies published in 2023 were relatively limited and primarily involved analysis-related applications. The number and diversity of applications increased during 2024, with studies increasingly represented in Recruitment, Design, and Conduct in addition to Analysis. Studies published in 2025 continued to span multiple clinical trial phases.

As illustrated in Figure 2, LLM applications spanned multiple stages of the clinical trial lifecycle and a diverse range of functional tasks and clinical specialties. Analysis-related studies were primarily associated with clinical feature extraction and other data-analysis tasks, whereas recruitment applications included eligibility criteria structuring and patient-related screening and support functions. Design-stage applications included protocol and trial design generation, eligibility criteria structuring, and other documentation-related tasks. Conduct-stage and multi-phase applications were less frequently represented. Across these trial phases, applications were evaluated in oncology, psychiatry, cardiology, neurology, surgery, and other clinical specialties, although many studies were not restricted to a specific specialty. The following sections describe these applications in greater detail by clinical trial phase.

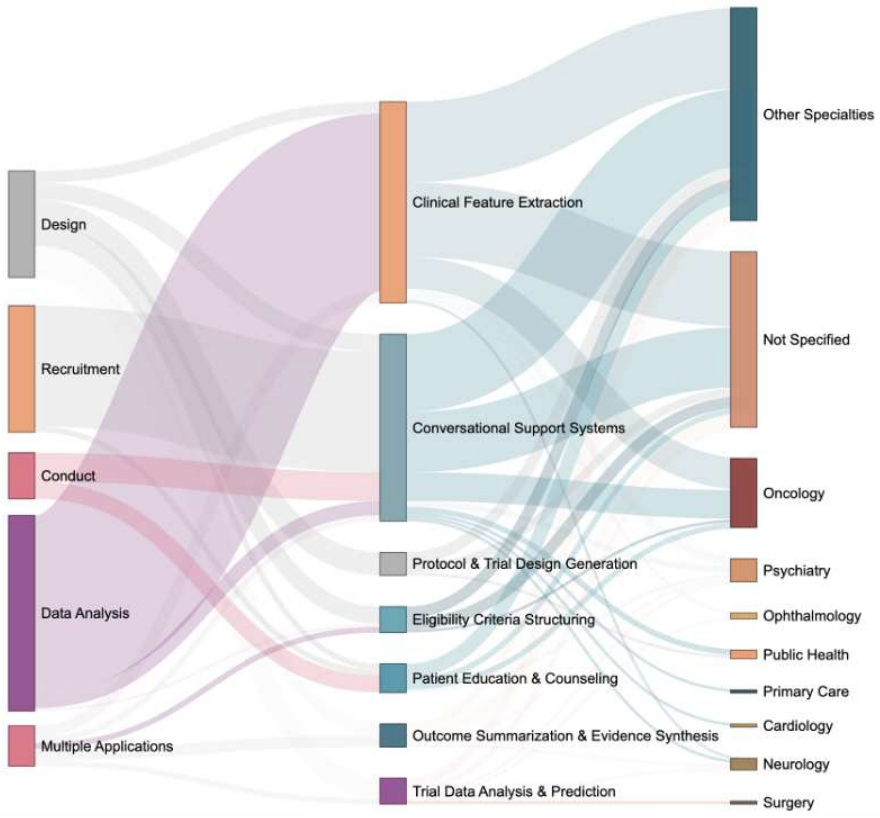


Fig. 2 Sankey diagram mapping LLM applications from trial phases to functional tasks and clinical specialties, illustrating the breadth and distribution of use cases across the clinical trial workflow.

2.1.1 LLM Applications in Trial Design

LLM applications in trial Design were identified in 13 studies ($n = 13$, 18.3%) and primarily involved protocol authoring and documentation, eligibility criteria engineering, and ethical or participant-facing documentation. The primary applications identified in this phase are summarized in Table 1.

Protocol Authoring and Documentation Support. LLMs were used to generate and organize trial-related documents, including study objectives, eligibility sections, and schedules of events. Several studies evaluated LLM-generated protocol content or structured representations of trial information [14, 16, 17]. Related work applied LLM-generated annotations to protocol digitization, including the identification of structured tables within trial documents [15].

Eligibility Criteria Engineering. LLMs were used to extract, generate, and refine eligibility criteria during trial design. Applications included extraction of clinical concepts, value constraints, and temporal qualifiers from unstructured trial protocols

Table 1 Applications of LLMs in clinical trial design.

Design Category	Representative Studies	Primary Utility
Protocol Authoring and Documentation Support	[14–17]	Supports generation and organization of trial protocols and related documents, including protocol sections and structured trial information.
Eligibility Criteria Engineering	[17–20]	Supports extraction, generation, and refinement of eligibility criteria, including inclusion and exclusion definitions and structured representations of eligibility requirements.
Participant-Facing and Regulatory Documentation	[21–23]	Supports preparation and evaluation of informed consent materials and generation of structured regulatory and Institutional Review Board (IRB) documentation.

[17]. Generative approaches were also used to draft inclusion and exclusion criteria from trial descriptions. In AutoTrial, LLM-generated criteria were evaluated by human reviewers for clinical relevance and quality [18]. Other studies examined automated processing and refinement of eligibility criteria for downstream trial applications [19, 20].

Participant-Facing and Regulatory Documentation. LLMs were also evaluated for preparing informed consent and regulatory documents. Shi et al. [21] evaluated LLM-generated informed consent materials with respect to measures including accuracy, completeness, and readability. Godwin et al. [22] developed an LLM-based system that converted study information into structured, editable documents for IRB submission. Maniaci et al. [23] examined the use of LLMs for summarizing clinical guidance and information relevant to trial-related decision making.

2.1.2 LLM Applications in Recruitment and Eligibility Screening

LLM applications in participant Recruitment were identified in 22 studies ($n = 22$, 31.0%) and primarily involved eligibility criteria processing, patient-to-trial matching, and recruitment feasibility or population assessment. These Recruitment applications are summarized in Table 2.

Criteria Structuring and Query Generation. LLMs were used to convert narrative eligibility criteria into structured representations for downstream screening and retrieval. Criteria2Query 3.0 generated structured representations from narrative eligibility criteria, while other studies used LLMs to generate retrieval queries from clinical summaries or classify eligibility statements [24–26].

LLM-Assisted Patient-Trial Matching. LLMs were applied to patient-to-trial matching using diverse approaches. Depending on the task and scale of the candidate trial set, systems could incorporate trial retrieval to narrow the search space and assess eligibility either at the level of individual inclusion and exclusion criteria or at the overall trial level. System outputs also varied, including eligibility classifications, explanatory rationales, and scores for prioritizing or ranking candidate trials. Frameworks such as PRISM, TrialGPT, and Patient2Trial illustrate different

Table 2 Applications of LLMs in clinical trial recruitment and eligibility screening.

Application Category	Representative Studies	Primary Utility
Eligibility Criteria Structuring and Query Translation	[24–26]	Converts narrative inclusion and exclusion criteria into structured or computable representations and generates queries or classifications that support automated eligibility screening.
LLM-Assisted Patient-Trial Matching	[27–31]	Matches patient information with trial eligibility requirements using approaches that may include retrieval, criterion- or trial-level eligibility assessment, classification, and candidate prioritization to support prescreening and recruitment.
Recruitment Feasibility and Population Assessment	[32, 33]	Uses trial and population information to assess recruitment feasibility, estimate enrollment-related outcomes, and identify candidate populations or biomarker-defined cohorts.

combinations of these functions [27–29]. Across oncology, hepatology, lung cancer, and general EHR-based screening, LLM-based approaches were evaluated against expert-annotated reference standards, existing matching systems, or physician recommendations [30, 31, 34–38]. These systems primarily support prescreening and recruitment rather than fully autonomous matching, as final eligibility determination may require additional clinical information and review by trial personnel.

Where error direction could be assessed, performance profiles differed across systems. Hung et al. reported 100% recall and 63.0% precision for retrieval-augmented GPT-4 relative to multidisciplinary physician recommendations, with errors primarily involving additional trial recommendations not selected by physicians [38]. In contrast, Wong et al. reported higher precision but lower recall for GPT-4 than for an oncology-specific expert system (87.6% vs. 84.3% precision; 67.3% vs. 97.4% recall), indicating more missed eligible matches [34]. Specialized applications included biomarker extraction for precision oncology [33], matching frameworks designed to assess or mitigate disparities across patient subgroups [39], and retrieval-augmented approaches for patient-to-trial matching [40, 41].

Recruitment Feasibility and Population Assessment. Several studies used LLMs to support recruitment-related assessment. Applications included prediction of trial enrollment success [32], patient-to-trial retrieval and prioritization [29], and extraction and structuring of genomic biomarkers from trial descriptions to support biomarker-based oncology trial matching [33].

Table 3 Utility-focused applications of LLMs in clinical trial conduct.

Application Category	Representative Studies	Primary Utility
Participant Interaction and Intake	[42, 43]	Supports participant-facing interactions, including initial assessment and counseling, by providing structured conversational interfaces for information collection and delivery.
Data Acquisition and Analysis	[44, 45]	Supports collection and analysis of participant-generated or trial-related data, including patient-reported outcomes and unstructured clinical interactions, to derive structured measures and outcome-related information.

2.1.3 LLM Applications in Trial Conduct

LLM applications in trial Conduct were identified in four studies ($n = 4$, 5.6%) and primarily involved participant interaction and the collection or analysis of participant-generated data. Table 3 summarizes the identified Conduct-phase applications and their primary utilities.

Automated Patient Interaction and Intake. Two studies evaluated LLMs as participant-facing conversational systems. Rosteck et al. [42] developed a hybrid LLM-driven system for initial mental health assessments and evaluated it in a randomized user study involving 72 healthy participants. The LLM-assisted system received lower usability scores than the form-based control, although qualitative feedback identified positive aspects related to reflection and empathetic interaction. Chung et al. [43] evaluated ChatGPT-assisted pre-vasectomy counseling in a prospective randomized study involving 18 patients, reporting shorter consultation duration and higher provider-rated patient understanding in the ChatGPT-assisted group. Thus, both participant-facing Conduct studies used prospective randomized designs, although Rosteck et al. evaluated healthy participants rather than patients.

Data Acquisition and Analysis. The other two Conduct-phase studies used LLMs to derive information from conversational or participant-generated data. Shen and Yin [44] evaluated a chatbot-based framework for patient-reported outcome collection and Bayesian mediation analysis using simulated patient-chatbot interactions. Dougherty et al. [45] retrospectively applied a zero-shot LLM to transcripts from psychological support sessions collected during a randomized psilocybin trial. Language-derived features, including sentiment and emotional breakthrough measures, were used to predict subsequent treatment response, with reported prediction accuracies of 85–88%.

Table 4 Utility-focused applications of LLMs in clinical trial data analysis.

Application Category	Representative Studies	Primary Utility
Evidence Extraction and Structuring	[46–49]	Extracts structured information from trial publications and related documents, including interventions, outcomes, numerical results, and other trial characteristics, to support downstream evidence analysis.
Evidence Synthesis and Summarization	[50–53]	Summarizes and synthesizes information across clinical trial documents to support evidence review, comparison, and interpretation.
Reporting Quality and Compliance	[54–57]	Assesses trial reports against reporting guidelines and extracts or identifies reporting elements to support evaluation of completeness and compliance.
Risk-of-Bias and Trustworthiness Assessment	[58–61]	Supports assessment of risk of bias and related methodological characteristics in randomized trials by extracting relevant evidence and generating domain-level assessments.

2.1.4 LLM Applications in Data Analysis

LLM applications in trial Analysis were identified in 29 studies ($n = 29$, 40.8%) and primarily involved evidence extraction and structuring, evidence synthesis and summarization, reporting-quality assessment, and risk-of-bias or trustworthiness assessment. The major Analysis applications and their primary utilities are summarized in Table 4.

Evidence Extraction and Structuring. LLMs were used to extract and structure information from clinical trial publications, registries, and clinical data. Applications included extraction of interventions, comparators, outcomes, numerical results, and other trial-related information [46, 47, 49–51, 62]. Other studies evaluated approaches incorporating grammar-constrained decoding and biomedical representations for structured information extraction [48, 63]. LLM-based extraction was also applied to clinical data, including the identification of left ventricular ejection fraction values and hospitalization-related outcomes from hospital records [64].

Evidence Synthesis and Summarization. LLMs were also evaluated for summarizing and synthesizing information across clinical trial documents. Applications included multi-document summarization, generation of treatment-related evidence summaries, and synthesis or simplification of findings from clinical trial reports [50–53]. These studies evaluated generated outputs using measures including factual accuracy, completeness, readability, and comparison with human-generated or expert-reviewed summaries.

Reporting Quality and Compliance. Several studies applied LLMs to assess or support clinical trial reporting. Applications included generation of reporting checklists and identification of reporting elements in trial publications [54, 55]. Other studies

used LLM-based approaches to assess adherence to CONSORT reporting requirements and identify missing or incomplete reporting elements in trial abstracts and publications [56].

Risk-of-Bias and Trustworthiness Assessment. LLMs were evaluated for risk-of-bias and related methodological assessments of randomized trials. Applications included identification of methodological limitations, generation of domain-level risk-of-bias assessments, and evaluation of trial trustworthiness [58–61]. Comparative evaluations assessed agreement between LLM-generated judgments and human or reference assessments, with performance varying across models and assessment tasks. RoBGuard further incorporated retrieval-augmented generation to support risk-of-bias assessment [59].

2.1.5 LLM Applications Across Multiple Trial Phases

Three studies ($n = 3$, 4.2%) applied LLMs across more than one clinical trial phase. These applications included simulated trial analysis, multi-agent systems for trial-related prediction, and digital-twin approaches for modeling participant trajectories.

Goldenholz et al. [65] evaluated LLM-based inductive reasoning in a simulated randomized controlled trial for epilepsy. The study used unstructured clinical notes to assess whether LLMs could identify treatment effects and symptom patterns and compared the resulting analyses with those of human experts.

Yue et al. [66] developed *ClinicalAgent*, a GPT-4-powered multi-agent system for clinical trial prediction. The framework incorporated ReAct and Least-to-Most reasoning and was evaluated across multiple trial-related tasks, including prediction of trial outcomes, participant enrollment, adverse events, and study duration.

Wang et al. [67] developed *TWIN-GPT*, an LLM-based framework for generating digital twins from longitudinal clinical trial data. The framework modeled individual participant trajectories and was evaluated for prediction of clinical outcomes and simulation of participant-level responses.

2.2 Data Provenance, Governance, and Ethical Compliance

The included studies used diverse data sources with varying levels of accessibility. As shown in Figure 3c, 77.6% of the identified data sources were publicly available, whereas 22.4% were non-public or institutionally controlled. Publicly available sources included clinical trial registries, published biomedical literature, and public benchmark datasets, whereas non-public sources included institutionally controlled EHR data, participant-level clinical trial data, and other restricted clinical datasets.

Figure 3d shows the distribution of dataset sizes across the included studies. The most common data size range was 101–1,000 ($n = 20$, 28.2%), followed by ≤ 100 ($n = 14$, 19.7%). Studies with 1,001–10,000 and 10,001–100,000 data instances each accounted for 13 studies (18.3%), while 3 studies (4.2%) included more than 100,000 data instances. Dataset size was not reported for 8 studies (11.3%).

Data-source distributions also differed across clinical trial phases (Figure 3e). Analysis-focused studies most frequently used secondary trial data, followed by public benchmark datasets. Recruitment studies used several data-source categories, with

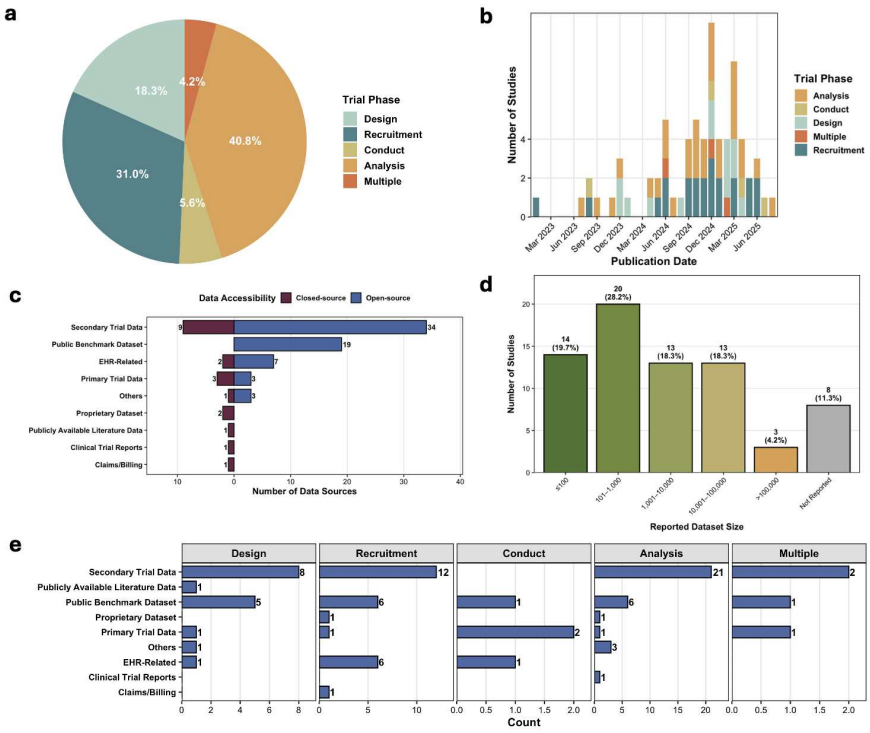


Fig. 3 Metadata Landscape of LLM Applications in Clinical Trial Research. (a) Distribution of LLM applications by trial phase; **(b)** Monthly trends in large language model applications across clinical trial phases; **(c)** Distribution of data sources used in included studies; **(d)** Distribution of Studies by Dataset Size; **(e)** Distribution of data types across trial phases.

secondary trial data, EHR-related data, and public benchmark datasets being the most frequently represented. Design studies used both secondary and public benchmark data, together with smaller numbers of EHR-related, primary trial, and other data sources. Conduct-phase studies more frequently included primary trial data, with additional use of public benchmark, EHR-related, and proprietary datasets. Studies spanning multiple trial phases primarily used public benchmark and secondary trial data.

We identified six studies (8.5%) that used non-public or institutionally controlled clinical data, including EHR-derived data, prospectively collected trial data, participant-level clinical trial records, and psychotherapy transcripts [27, 31, 34, 43, 45, 64]. Among these studies, two explicitly reported de-identification or anonymization of patient data before or during LLM-based processing [31, 34], while one explicitly reported ethics committee approval [31]. Several studies described additional safeguards related to model deployment or data handling, including the use of a secure private endpoint, private model-hosting infrastructure, or restricted processing of de-identified or non-PHI patient representations [27, 31, 34, 64].

Reporting of data-governance practices varied across these studies. Information on de-identification, ethics oversight, model-hosting environments, external data transmission, and access controls was not consistently provided. Four of the six studies did not explicitly report de-identification of the patient-level data used in the LLM workflow [27, 43, 45, 64], and five did not explicitly report IRB or ethics oversight in the publication [27, 34, 43, 45, 64]. The absence of such reporting does not indicate that these safeguards were absent, but limits assessment of the privacy and governance procedures implemented in individual studies.

2.3 LLM Methods, Optimization, and Evaluation

The included studies used a range of LLM architectures, with proprietary models accounting for the majority of applications. As shown in Figure 4a, approximately 83.1% of studies used proprietary LLMs, most commonly OpenAI GPT models, including GPT-3.5 and GPT-4.

Studies used several strategies to adapt LLMs to clinical trial-related tasks (Figure 4e). Operational definitions of the prompting and adaptation methods used in this review are provided in Table 5. Prompt engineering was the most frequently reported approach, used in 69.0% of studies. Zero-shot prompting was used in 38.8% of studies, while few-shot prompting was used in 14.3%. Chain-of-thought (CoT) prompting and other multi-step reasoning approaches were reported in 6.1%. Fine-tuning was used in 16.9% of studies, including approaches incorporating biomedical or clinical trial-related data from sources such as PubMed, ClinicalTrials.gov, and institutional trial registries. Retrieval-augmented generation (RAG) was used in 14.1% of studies and incorporated external information sources, including biomedical literature, clinical trial registries, and study-specific document collections. These adaptation methods were applied across tasks including information extraction, question answering, patient-to-trial matching, and other trial-related applications.

Table 5 Definitions of LLM Prompting and Adaptation Methods Used in Included Studies

Method	Operational Definition
Zero-shot prompting	The model performs a task using natural-language instructions without task-specific examples.
Few-shot prompting	The model is provided with a small number of example input-output pairs to guide task-specific behavior.
Chain-of-Thought (CoT)	A prompting strategy that elicits intermediate reasoning steps before generation of a final response.
Retrieval-Augmented Generation (RAG)	An approach in which an external retrieval system retrieves relevant documents or data that are provided as context to the LLM during generation.
Fine-tuning	Additional training of a pretrained language model using task-specific or domain-specific data to adapt model parameters to the target application.

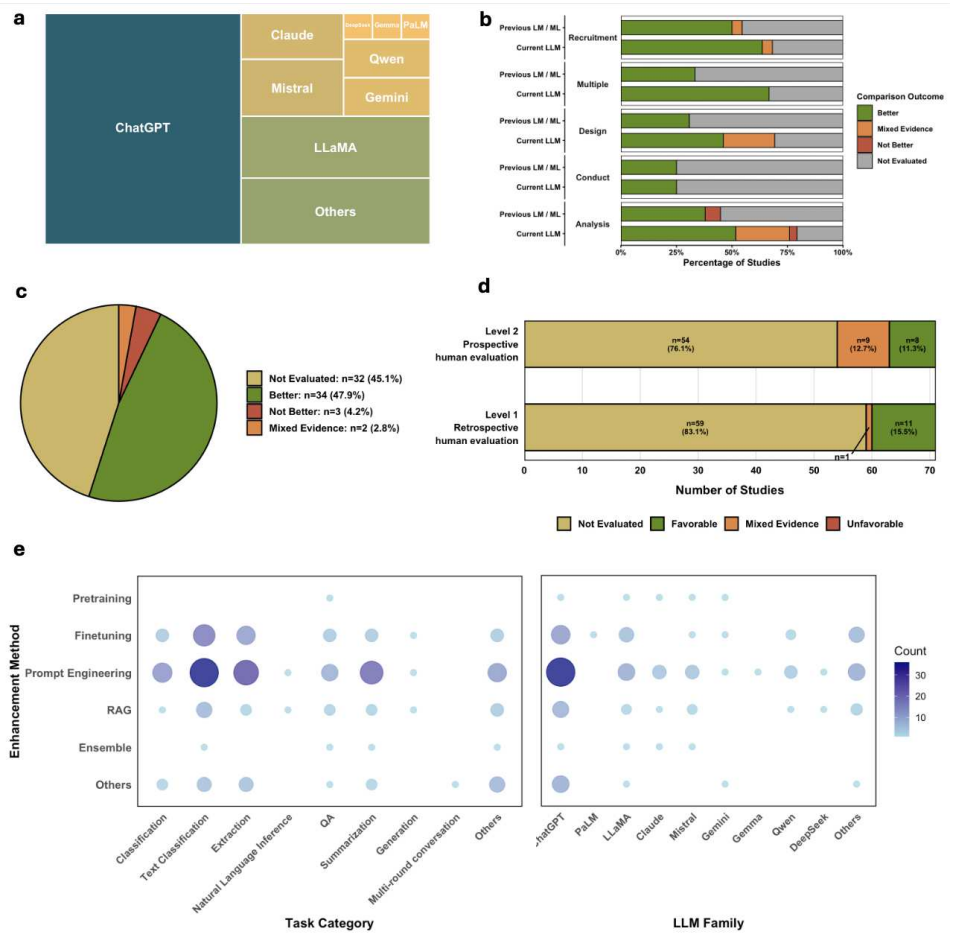


Fig. 4 Landscape of LLM model usage, comparative performance, human evaluation, and enhancement methods across clinical trial applications. **(a)** Distribution of large language models used across included studies. **(b)** Comparison outcomes across clinical trial phases, stratified by comparator type. Comparisons with other current LLMs were distinguished from comparisons with previous language models or machine learning methods. **(c)** Overall comparison outcomes for LLM-based approaches versus previous language models or machine-learning methods across all included studies. **(d)** Outcomes of retrospective human-reference comparisons and prospective human evaluations of LLM-generated outputs. **(e)** Enhancement methods used across studies, mapped to machine-learning task categories and LLM families. Bubble sizes indicate the number of studies using each enhancement method.

Evaluation approaches varied across the included studies. Quantitative performance metrics, including accuracy and F1-score, were reported in 85.9% of studies. Figure 4b–d summarizes computational comparisons and human-based evaluations. To distinguish different types of computational comparators, we separately examined comparisons with other current LLMs and comparisons with previous language models or machine-learning methods. Figure 4b shows the outcomes of these two comparator types across clinical trial phases.

Across all 71 studies, 34 (47.9%) reported better performance in at least one comparison with a previous language model or machine-learning method, 3 (4.2%) reported that the LLM-based approach did not perform better, and 2 (2.8%) reported mixed results. The remaining 32 studies (45.1%) did not report a comparison with a previous language model or machine-learning method (Figure 4c).

Human-based evaluation was identified in 25 of 71 studies (35.2%) and was classified into two levels according to evaluation design (Figure 4d). Level 1 represented retrospective human-reference comparisons, in which LLM outputs were compared with pre-existing human annotations or decisions. Twelve studies (16.9%) included this type of evaluation; 11 reported favorable findings and one reported mixed findings. Level 2 represented prospective human evaluation of LLM-generated outputs, in which evaluators assessed outputs generated as part of the study rather than relying solely on pre-existing human labels. Seventeen studies (23.9%) included this type of evaluation; 8 reported favorable findings and 9 reported mixed findings. Four studies included both Level 1 and Level 2 evaluations; therefore, the two categories were not mutually exclusive.

Human-based evaluations covered tasks including eligibility screening, patient-to-trial matching, information extraction, informed consent generation, evidence synthesis, and guideline-related applications. For example, PRISM was evaluated using human assessment for patient-to-trial matching [27]. In informed consent generation, documents generated using Mistral 8×22B were assessed by human reviewers on dimensions including readability and actionability [21]. In another study evaluating ChatGPT for CRSwNP guideline synthesis, human evaluation identified limitations in response accuracy and potentially unsafe recommendations [23].

The design of prospective human evaluations differed across studies. In many studies, evaluators rated LLM-generated outputs according to criteria such as accuracy, completeness, readability, usefulness, or clinical appropriateness. These evaluations generally involved human assessment of LLM-generated outputs rather than paired designs in which humans and LLMs independently performed the same task under comparable conditions.

3 Discussion

This scoping review of 71 studies maps the application of LLMs across the clinical trial lifecycle. LLM applications spanned trial Design, Recruitment, Conduct, and Analysis, with uses ranging from protocol and eligibility-related tasks to patient-to-trial matching, participant-facing interaction, information extraction, evidence synthesis, and methodological assessment. This expansion is consistent with broader trends in healthcare AI toward more proactive and integrated applications across clinical workflows [68]. At the same time, the evidence base reveals substantial differences in adoption, evaluation, and implementation across clinical trial phases.

Applications were concentrated in Analysis (40.8%) and Recruitment (31.0%), followed by Design (18.3%) and Conduct (5.6%). One possible explanation for this distribution is the difference in operational risk across trial phases. In Design and Analysis, many LLM outputs can undergo human review before affecting participants

or trial operations. In contrast, applications during Conduct may involve participant interaction, collection of participant-generated information, or other activities occurring during trial execution, increasing requirements for reliability, safety, and oversight. Consistent with this distinction, Conduct-related applications identified in this review were relatively constrained in scope and included structured assessment, participant counseling, conversational data collection, and analysis of participant-generated data [42–45]. The observed phase distribution should also be interpreted in light of potential publication and survivorship bias. Unsuccessful or discontinued LLM applications, including attempts to integrate LLMs into trial Design or Conduct, may be less likely to appear in the published literature. The distribution observed in this review may therefore reflect both differences in the feasibility of applying LLMs across trial phases and differences in which applications progress to formal evaluation and publication.

Compared with the broader use of AI for physiological monitoring and multimodal data reported in previous clinical trial research [69], the studies included in this review showed limited use of LLMs for direct interpretation of biosignals such as electrocardiography, accelerometry, or medical imaging. Conduct-related applications instead focused predominantly on language-mediated interaction and data collection. Rosteck et al. evaluated an LLM-assisted conversational system for initial mental health assessment, while Chung et al. evaluated ChatGPT-based counseling in a procedural setting [42, 43]. Shen and Yin used LLM-mediated dialogue for collection of patient-reported information in a simulated setting [44]. Within the scope of this review, LLM applications therefore remained predominantly language-centered, with limited integration of physiological, imaging, or sensor-derived data. This distinguishes current generative LLM applications from the broader use of multimodal AI in clinical research. A future direction is the integration of generative language models with multimodal foundation models capable of processing imaging, physiological signals, and wearable-sensor data. Such integration could extend LLM-supported trial workflows beyond text-based interaction and documentation toward multimodal participant monitoring and analysis. Future studies should determine whether these systems provide clinically meaningful information beyond language-only approaches and establish appropriate validation and governance frameworks for their use in prospective trials.

Computational evaluation showed frequent favorable performance relative to previous language models or machine-learning methods, but comparator use was not consistent across studies. Of the 71 included studies, 34 (47.9%) reported better performance in at least one comparison with a previous language model or machine-learning method, whereas 32 (45.1%) did not include this type of comparison. The remaining studies reported either mixed results or no improvement over the comparator. These findings should be interpreted in the context of substantial variation in tasks, datasets, comparator selection, and evaluation metrics. Comparisons with previous language models or machine-learning methods also address a different question from human-based evaluation. They establish relative computational performance within a specific experimental setting but do not by themselves demonstrate clinical utility, workflow benefit, or equivalence to expert performance.

Human-based evaluation was identified in 25 of 71 studies (35.2%), but the design and interpretation of these evaluations varied substantially. Twelve studies (16.9%)

included retrospective human-reference comparisons, of which 11 reported favorable findings and one reported mixed findings. Seventeen studies (23.9%) included prospective human evaluation of LLM-generated outputs, with eight reporting favorable findings and nine reporting mixed findings; four studies included both forms of evaluation. The less uniform findings in prospective human evaluation may partly reflect differences in tasks, evaluators, assessment criteria, and study designs. Importantly, most prospective evaluations involved experts rating LLM-generated outputs for criteria such as accuracy, completeness, readability, usefulness, or clinical appropriateness rather than paired comparisons in which humans and LLMs independently performed the same task under equivalent conditions. Favorable expert ratings therefore provide evidence regarding the perceived quality or appropriateness of model outputs but should not be interpreted as demonstrating equivalence or superiority to human experts. For example, human evaluations of informed consent generation assessed dimensions including readability and actionability, whereas evaluation of ChatGPT for CRSwNP guideline synthesis identified limitations in accuracy and potentially unsafe recommendations [21, 23]. Future evaluations would benefit from clinically relevant endpoints that extend beyond aggregate performance metrics to include error severity, workflow impact, safety, and the consequences of incorrect model outputs.

Patient-to-trial matching further illustrates why the direction and consequences of model errors should be considered alongside aggregate performance. In studies that allowed error direction to be examined, different systems showed different precision-recall profiles. Hung et al. reported 100% recall and 63.0% precision for a retrieval-augmented GPT-4 approach relative to multidisciplinary physician recommendations, with additional trial recommendations accounting for an important component of the disagreement [38]. In contrast, Wong et al. reported higher precision but lower recall for GPT-4 than for an oncology-specific expert system (87.6% vs. 84.3% precision; 67.3% vs. 97.4% recall), indicating more missed potentially eligible matches [34]. These error profiles have different implications for recruitment workflows. Additional recommendations may increase downstream review burden, whereas missed matches may prevent potentially eligible patients from being considered for enrollment. Evaluations of recruitment systems should therefore report error-specific measures such as precision, recall, sensitivity, and specificity when applicable, rather than relying only on aggregate accuracy or F1-score.

Publicly available data played a prominent role in the development and evaluation of LLM applications across the included studies. Common resources included public trial registries such as ClinicalTrials.gov, published biomedical literature, and publicly available benchmark datasets. Shared and accessible resources can facilitate reproducibility and methodological comparison by allowing multiple research groups to develop and evaluate methods using common sources. At the same time, repeated evaluation on a limited set of widely used public resources may not capture the diversity of patient populations, institutional practices, documentation patterns, and workflow conditions encountered in clinical trials. Validation across independent datasets and institutional settings will therefore be important for determining whether performance observed on publicly available resources generalizes to operational clinical trial environments.

Data governance becomes particularly important when LLM applications rely on non-public or institutionally controlled clinical data. Studies in this review used electronic health records, prospectively collected trial data, participant-level clinical trial records, and psychotherapy transcripts [27, 31, 34, 43, 45, 64]. The level of detail provided about governance practices varied considerably. Some studies described safeguards such as de-identification, institutional review, secure endpoints, or local model deployment, whereas others provided limited information about how sensitive clinical information was managed. Lack of documentation should not be interpreted as evidence that such safeguards were absent; rather, it limits assessment of data protection and governance practices. Studies using non-public clinical data should, at minimum, document data provenance, de-identification procedures, model hosting arrangements, transmission to external model providers, access controls, and institutional or regulatory oversight. More consistent documentation of these elements would facilitate assessment of privacy, reproducibility, and implementation feasibility across studies.

Several limitations should be considered when interpreting this review. First, the rapid evolution of generative LLMs means that model architectures, deployment practices, and evaluation approaches may change faster than the peer-reviewed literature can capture. Second, classification of studies into clinical trial phases required judgment when an application could plausibly support more than one stage, although studies that substantively spanned multiple phases were classified separately. Third, the review focused specifically on generative language models and therefore does not represent the broader literature on AI methods for imaging, physiological signals, wearable sensors, or other non-language modalities. Fourth, human-based evaluations varied substantially in study design, reference standards, evaluator expertise, and outcome measures, limiting direct comparison across studies. Finally, the evidence base may be affected by publication and survivorship bias because unsuccessful or discontinued LLM applications may be less likely to be reported.

LLM applications now span the clinical trial lifecycle, but the maturity of the evidence differs substantially across phases and use cases. Current applications are concentrated in Recruitment and Analysis, while relatively few studies have evaluated LLMs for activities occurring during trial Conduct. Computational comparisons frequently reported favorable performance, but human-based evaluation remains limited and heterogeneous, and prospective human evaluation often consists of expert assessment of model outputs rather than direct comparison with human performance. Progress toward operational use will therefore depend not only on further improvements in computational performance but also on prospective evaluation, clinically meaningful error analysis, transparent data governance, and validation across diverse real-world trial settings.

4 Methods

4.1 Search Strategy

A comprehensive literature search was conducted across seven major bibliographic databases: *PubMed*, *Web of Science*, *Embase*, *Scopus*, *IEEE Xplore*, *Cochrane*, and the *ACM Digital Library*. Searches were performed between August 1, 2024, and July

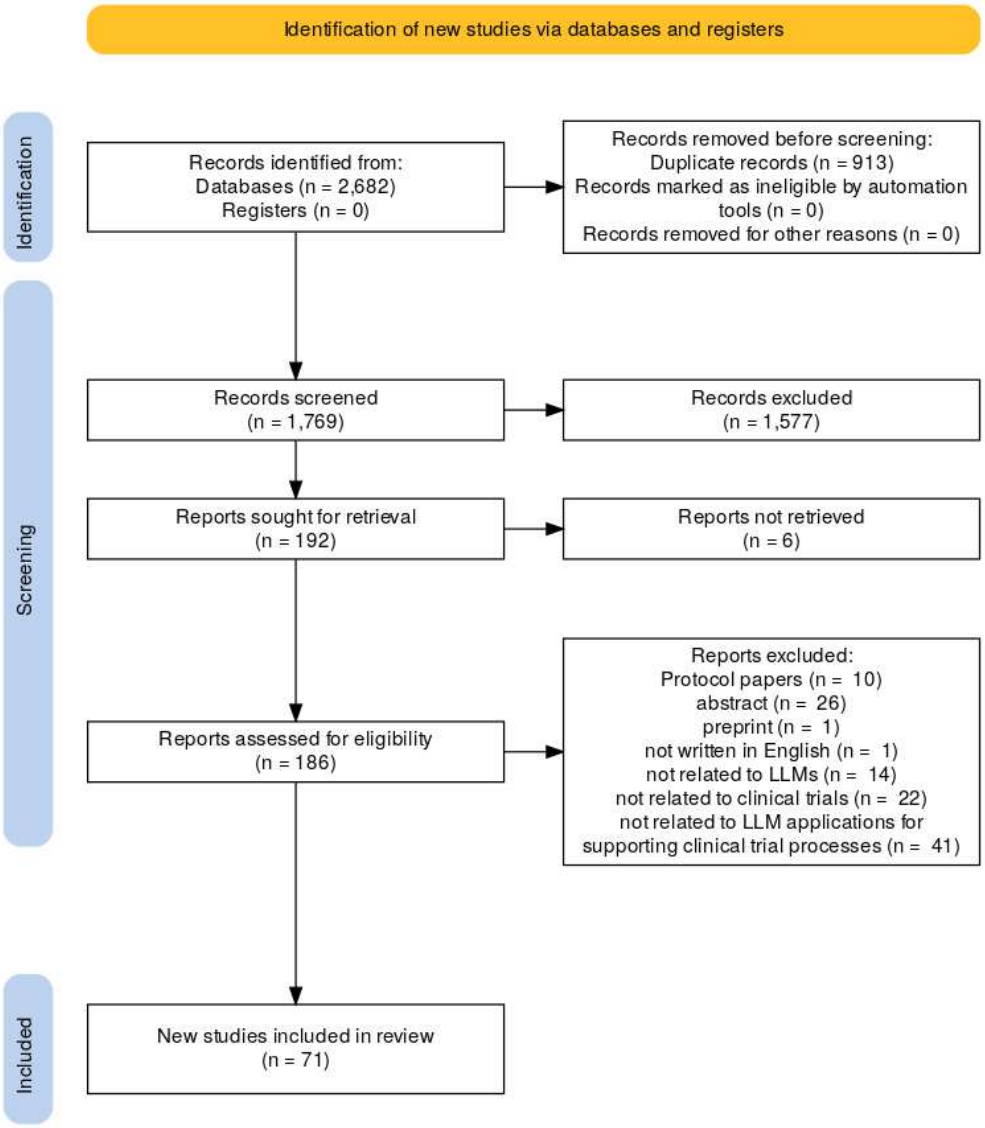


Fig. 5 PRISMA flow diagram summarizing the identification, screening, eligibility, and inclusion of studies.

31, 2025. This scoping review was reported in accordance with the PRISMA extension for scoping reviews (PRISMA-ScR).

The search strategy was designed to identify studies that explicitly integrated LLMs within the context of clinical trial design, conduct, analysis, or reporting.

Boolean queries combined clinical trial–related terms (`"clinical trial*"`, `"randomized trial*"`, `"randomized controlled trial*"`, `"RCT*"`) with LLM-related terms (`"large language model*"`, `"LLM"`, `"foundation model*"`, `"generative model"`, `"GPT"`, `"LLaMA"`, `"Mixtral"`, `"Claude"`, `"Gemini"`, `"Bard"`, `"BLOOMZ"`, `"Qwen"`, `"DeepSeek"`, `"PaLM"`, `"Falcon"`, `"Vicuna"`, `"Alpaca"`, `"Gemma"`).

All results were exported in RIS format and consolidated in *EndNote 21* for deduplication. Duplicates were removed prior to screening.

4.2 Eligibility Criteria and Screening

Inclusion criteria required that studies (1) explicitly applied or evaluated an LLM within any phase of the clinical trial lifecycle, including Design, Recruitment, Conduct, or Analysis; (2) were published in peer-reviewed journals or conference proceedings from 2022 through 2025; and (3) were available in English. Studies were excluded if they (a) focused exclusively on AI models outside the LLM class (e.g., BERT), (b) presented non-trial biomedical applications, or (c) were review articles, abstracts without full text, or non-peer-reviewed preprints. Two independent reviewers screened titles and abstracts, followed by full-text review.

To support consistent coding of included studies, we adopted a four-stage classification framework for LLM applications across the clinical trial lifecycle: Design, Recruitment, Conduct, and Analysis. The framework was informed by prior descriptions of AI and machine-learning applications across clinical trial workflows [69, 70] and was operationalized specifically for LLM applications. The operational definitions for each stage are provided in Table 6. The study was the primary unit of analysis for study-level summaries, with each included study contributing one observation regardless of the number of LLMs, tasks, or datasets evaluated. Each study was assigned to one mutually exclusive clinical trial phase based on its primary application. Studies that substantively spanned more than one phase were classified as Multiple.

Evaluation approaches were coded according to comparator type and human-evaluation design. Computational comparisons with other contemporary generative LLMs were distinguished from comparisons with previous language models or conventional machine-learning/NLP methods. Human-based evaluations were classified as retrospective human-reference comparisons when LLM outputs were compared with pre-existing human annotations or decisions, and as prospective human evaluations when human evaluators assessed LLM-generated outputs produced as part of the study. Studies could contribute to both human-evaluation categories when both designs were used. Detailed definitions of the extracted evaluation variables and coding criteria are provided in Appendix Table A1.

A standardized data extraction form was developed to characterize the included studies across study characteristics, clinical trial applications, LLM technologies, data characteristics, privacy and governance, and evaluation strategies. Extracted variables included publication characteristics and medical domain; clinical trial stage and LLM function; model family and version; data source, accessibility, and sample size; model enhancement strategies; privacy and governance practices; and evaluation approaches, including comparisons with human evaluators and conventional machine learning or

Table 6 Operational Definitions of Clinical Trial Stages for LLM Applications

Stage	Operational Definition
Design	LLM applications used before participant enrollment to support trial planning and optimization, including protocol development, eligibility-criteria refinement, and early predictive modeling.
Recruitment	LLM applications used for identifying, screening, and onboarding eligible participants into clinical trials, including automated eligibility structuring and patient–trial matching.
Conduct	LLM applications targeting activities that occur during trial execution, including participant interaction, assessment, conversational support, adherence-related functions, and participant-generated data collection or monitoring.
Analysis	LLM applications targeting analysis, interpretation, synthesis, or reporting of clinical trial data and evidence, including evidence extraction, summarization, methodological assessment, and reporting-quality evaluation.
Multiple	LLM applications that operate across two or more clinical trial phases, typically integrating simulation, predictive reasoning, or end-to-end workflow support. Examples include digital-twin models, multi-agent frameworks, and inductive reasoning systems capable of linking design, recruitment, conduct, and analysis tasks.

natural language processing methods. The complete data extraction framework is provided in Appendix A.

Included studies were categorized according to the predefined clinical trial lifecycle framework as Design, Recruitment, Conduct, Analysis and Reporting, or Multi-stage. LLM applications were further characterized according to their specific functions within each stage. Model use and enhancement strategies, including prompting, retrieval-augmented generation, fine-tuning, and other approaches, were summarized across studies. Data characteristics were analyzed according to source, accessibility, sample size, and use of patient-level data, with additional extraction of reported privacy and governance practices for studies involving sensitive clinical data.

Evaluation approaches were characterized according to the reported evaluation metrics and the presence of comparisons with human evaluators or conventional computational methods. Study characteristics and extracted variables were synthesized descriptively using counts and proportions. Because of substantial heterogeneity in study objectives, datasets, LLM applications, and evaluation methods, no quantitative meta-analysis was performed.

Data Availability

The data supporting the findings of this review were extracted from publicly available publications. The aggregated data extraction tables are available in the Supplementary Information.

Code Availability

The R scripts used for descriptive analyses and figure generation are available from the corresponding author upon reasonable request.

Author Contributions

HZ, XZ, SZ, YL, HK, and RZ contributed to the study conception and methodology. HZ, XZ, SZ, MS, XC, YW, YC, KY, YX, YH, ZZ, and SM contributed to literature screening, data extraction, and analysis. HZ, XZ, and SZ drafted the manuscript. All authors contributed to manuscript review and revision and approved the final version.

Competing Interests

The authors declare no competing interests.

Ethics Approval

Ethics approval was not required because this scoping review analyzed previously published studies and did not involve the collection of individual-level participant data.

Funding

This study was supported by the National Center for Complementary and Integrative Health (R01AT009457 and U01AT012871) and the National Institute on Aging (R01AG078154).

References

[1] Frieden, T.R.: Evidence for health decision making—beyond randomized, controlled trials. New England Journal of Medicine **377**(5), 465–475 (2017)

[2] Djurisic, S., Rath, A., Gaber, S., Garattini, S., Bertele, V., Ngwabyt, S.-N., *et al.*: Barriers to the conduct of randomised clinical trials within all disease areas. Trials **18**(1), 360 (2017)

[3] Emanuel, E.J., Schnipper, L.E., Kamin, D.Y., Levinson, J., Lichter, A.S.: The costs of conducting clinical research. Journal of Clinical Oncology **21**(22), 4145–4150 (2003)

[4] Thirunavukarasu, A.J., Ting, D.S.W., Elangovan, K., Gutierrez, L., Tan, T., Ting, D.: Large language models in medicine. Nature Medicine **29**(8), 1930–1940 (2023)

[5] Chopra, H., Annu, Shin, D.K., Munjal, K., Priyanka, Dhama, K., Emran, T.B.: Revolutionizing clinical trials: the role of ai in accelerating medical breakthroughs. International Journal of Surgery **109**(12), 4211–4220 (2023) https://doi.org/10.1097/JS9.0000000000000705

[6] Lin, A., Wang, Z., Jiang, A., Chen, L., Qi, C., Zhu, L., Mou, W., Gan, W., Zeng, D., Xiao, M., Chu, G., Peng, S., Wong, H.Z.H., Zhang, L., Zhang, H., Deng, X., Wang, Y., Zhang, J., Cheng, Q., Tang, B., Luo, P.: Large language models in clinical trials: applications, technical advances, and future directions. BMC Medicine **23**(1), 563 (2025) https://doi.org/10.1186/s12916-025-04348-9 . Accessed 2025-12-29

[7] Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., Tara, M.: A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Medical Informatics and Decision Making **25**(1), 117 (2025) https://doi.org/10.1186/s12911-025-02954-4 . Accessed 2025-12-01

[8] Qian, L., Lu, X., Haris, P., Zhu, J., Li, S., Yang, Y.: Enhancing clinical trial outcome prediction with artificial intelligence: a systematic review. Drug Discovery Today **30**(4), 104332 (2025) https://doi.org/10.1016/j.drudis.2025.104332 . Accessed 2025-12-29

[9] Ghosh, S., Schneider, M., Reinicke, C., Eickhoff, C.: Cohort Discovery: A Survey on LLM-Assisted Clinical Trial Recruitment (2025). https://arxiv.org/abs/2506.15301

[10] Chen, H., Li, X., He, X., *et al.*: Enhancing patient-trial matching with large language models: A scoping review of emerging applications and approaches. JCO Clinical Cancer Informatics **9**, 2500071 (2025) https://doi.org/10.1200/CCI-25-00071

[11] Srinivasan, A., Berkowitz, J., Friedrich, N.A., Kivelson, S., Tatonetti, N.P.: Large Language Model Analysis of Reporting Quality of Randomized Clinical Trial Articles: A Systematic Review. JAMA Network Open **8**(8), 2529418 (2025) https://doi.org/10.1001/jamanetworkopen.2025.29418 . Accessed 2025-12-29

[12] Omar, M., Nadkarni, G.N., Klang, E., Glicksberg, B.S.: Large language models in medicine: A review of current clinical trials across healthcare applications. PLOS Digital Health **3**(11), 0000662 (2024) https://doi.org/10.1371/journal.pdig.0000662 . Accessed 2025-12-01

[13] He, Z., Yang, L., Liang, Z., Li, X., Du, J.: Clinical outcomes and reporting quality of large language model interventions in practice: a systematic evidence map. npj Digital Medicine **9**, 672 (2026) https://doi.org/10.1038/s41746-026-02837-6

[14] Markey, N., El-Mansouri, I., Rensonnet, G., Van Langen, C., Meier, C.: From RAGs to riches: Utilizing large language models to write documents for clinical trials. Clinical Trials **22**(5), 626–631 (2025) https://doi.org/10.1177/17407745251320806 . Accessed 2025-11-10

[15] Kumar, B., Amar, J., Yang, E., Li, N., Jia, Y.: Selective Fine-tuning on LLM-labeled Data May Reduce Reliance on Human Annotation: A Case Study Using

[16] Reinisch, M., He, J., Liao, C., Siddiqui, S., Xiao, B.: CTP-LLM: Clinical Trial Phase Transition Prediction Using Large Language Models. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 3667–3672. IEEE, Lisbon, Portugal (2024). https://doi.org/10.1109/BIBM62325.2024.10822746 . https://ieeexplore.ieee.org/document/10822746/ Accessed 2025-11-11

[17] Bornet, A., Khlebnikov, P., Meer, F., Haas, Q., Yazdani, A., Zhang, B., Amini, P., Teodoro, D.: 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 (2025) https://doi.org/10.1093/jamia/ocae311 . Accessed 2025-11-10

[18] Wang, Z., Xiao, C., Sun, J.: AutoTrial: Prompting Language Models for Clinical Trial Design. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 12461–12472. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.766 . https://aclanthology.org/2023.emnlp-main.766 Accessed 2025-11-11

[19] Datta, S., Lee, K., Paek, H., Manion, F.J., Ofoegbu, N., Du, J., Li, Y., Huang, L.-C., Wang, J., Lin, B., Xu, H., Wang, X.: 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 (2024) https://doi.org/10.1093/jamia/ocad218 . Accessed 2025-11-11

[20] Curran, C., Neehal, N., Murugesan, K., Bennett, K.P.: Examining Trustworthiness of LLM-as-a-Judge Systems in a Clinical Trial Design Benchmark. In: 2024 IEEE International Conference on Big Data (BigData), pp. 4627–4631. IEEE, Washington, DC, USA (2024). https://doi.org/10.1109/BigData62323.2024.10825592 . https://ieeexplore.ieee.org/document/10825592/ Accessed 2025-11-11

[21] Shi, Q., Luzuriaga, K., Allison, J.J., Oztekin, A., Faro, J.M., Lee, J.L., Hafer, N., McManus, M., Zai, A.H.: Transforming Informed Consent Generation Using Large Language Models: Mixed Methods Study. JMIR Medical Informatics **13**, 68139–68139 (2025) https://doi.org/10.2196/68139 . Accessed 2025-11-10

[22] Godwin, R.C., Bryant, A.S., Wagener, B.M., Ness, T.J., DeBerry, J.J., Horn, L.L., Graves, S.H., Archer, A.C., Melvin, R.L.: IRB-draft-generator: A generative AI tool to streamline the creation of institutional review board applications. SoftwareX **25**, 101601 (2024) https://doi.org/10.1016/j.softx.2023.101601 . Accessed 2025-11-11

[23] Maniaci, A., Saibene, A.M., Calvo-Henriquez, C., Vaira, L., Radulesco, T., Michel, J., Chiesa-Estomba, C., Sowerby, L., Lobo Duro, D., Mayo-Yanez, M., Maza-Solano, J., Lechien, J.R., La Mantia, I., Cocuzza, S.: Is generative pre-trained

transformer artificial intelligence (Chat-GPT) a reliable tool for guidelines synthesis? A preliminary evaluation for biologic CRSwNP therapy. European Archives of Oto-Rhino-Laryngology **281**(4), 2167–2173 (2024) https://doi.org/10.1007/s00405-024-08464-9 . Accessed 2025-11-11

[24] Park, J., Fang, Y., Ta, C., Zhang, G., Idnay, B., Chen, F., Feng, D., Shyu, R., Gordon, E.R., Spotnitz, M., Weng, C.: Criteria2Query 3.0: Leveraging generative large language models for clinical trial eligibility query generation. Journal of Biomedical Informatics **154**, 104649 (2024) https://doi.org/10.1016/j.jbi.2024.104649 . Accessed 2025-11-11

[25] Peikos, G., Kasela, P., Pasi, G.: 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), pp. 367–372. IEEE, Bangkok, Thailand (2024). https://doi.org/10.1109/WI-IAT62293.2024.00058 . https://ieeexplore.ieee.org/document/10973512/ Accessed 2025-11-11

[26] Ray, S., Sarker, A.N., Chatterjee, N., Bhowmik, K., Dey, S.: Leveraging Large Language Models for Clinical Trial Eligibility Criteria Classification. Digital **5**(2), 12 (2025) https://doi.org/10.3390/digital5020012 . Accessed 2025-11-11

[27] Gupta, S., Basu, A., Nievas, M., Thomas, J., Wolfrath, N., Ramamurthi, A., Taylor, B., Kothari, A.N., Schwind, R., Miller, T.M., Nadaf-Rahrov, S., Wang, Y., Singh, H.: PRISM: Patient Records Interpretation for Semantic clinical trial Matching system using large language models. npj Digital Medicine **7**(1), 305 (2024) https://doi.org/10.1038/s41746-024-01274-7 . Accessed 2025-11-11

[28] Jin, Q., Wang, Z., Floudas, C.S., Chen, F., Gong, C., Bracken-Clarke, D., Xue, E., Yang, Y., Sun, J., Lu, Z.: Matching patients to clinical trials with large language models. Nature Communications **15**(1), 9074 (2024) https://doi.org/10.1038/s41467-024-53081-z . Accessed 2025-11-11

[29] Datta, S., Lee, K., Huang, L.-C., Paek, H., Gildersleeve, R., Gold, J., Pillai, D., Wang, J., Higashi, M.K., Shi, L., Gulko, P.S., Xu, H., Weng, C., Wang, X.: Patient2Trial: From patient to participant in clinical trials using large language models. Informatics in Medicine Unlocked **53**, 101615 (2025) https://doi.org/10.1016/j.imu.2025.101615 . Accessed 2025-11-11

[30] Nievas, M., Basu, A., Wang, Y., Singh, H.: Distilling large language models for matching patients to clinical trials. Journal of the American Medical Informatics Association **31**(9), 1953–1963 (2024) https://doi.org/10.1093/jamia/ocae073 . Accessed 2025-11-11

[31] Gui, X., Lv, H., Wang, X., Lv, L., Xiao, Y., Wang, L.: Enhancing hepatopathy clinical trial efficiency: a secure, large language model-powered prescreening pipeline. BioData Mining **18**(1), 42 (2025) https://doi.org/10.1186/

s13040-025-00458-5 . Accessed 2025-11-11

[32] Yue, L., Xing, S., Chen, J., Fu, T.: TrialEnroll: Predicting Clinical Trial Enrollment Success with Deep & Cross Network and Large Language Models. In: Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 1–9. ACM, Shenzhen China (2024). https://doi.org/10.1145/3698587.3701375 . https://dl.acm.org/doi/10.1145/3698587.3701375 Accessed 2025-11-11

[33] Alkhoury, N., Shaik, M., Wurmus, R., Akalin, A.: Enhancing biomarker based oncology trial matching using large language models. npj Digital Medicine **8**(1), 250 (2025) https://doi.org/10.1038/s41746-025-01673-4 . Accessed 2025-11-11

[34] Wong, C., Zhang, S., Gu, Y., Moung, C., Abel, J., Usuyama, N., Weerasinghe, R., Piening, B., Naumann, T., Bifulco, C., Poon, H.: Scaling clinical trial matching using large language models: A case study in oncology. In: Deshpande, K., Fiterau, M., Joshi, S., Lipton, Z., Ranganath, R., Urteaga, I., Yeung, S. (eds.) Proceedings of the 8th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 219, pp. 846–862. PMLR, ??? (2023). https://proceedings.mlr.press/v219/wong23a.html

[35] Rybinski, M., Kusa, W., Karimi, S., Hanbury, A.: Learning to match patients to clinical trials using large language models. Journal of Biomedical Informatics **159**, 104734 (2024) https://doi.org/10.1016/j.jbi.2024.104734 . Accessed 2025-11-11

[36] Devi, A., Uttrani, S., Singla, A., Jha, S., Dasgupta, N., Natarajan, S., S. Punekar, R., A. Pickett, L., Dutt, V.: Automating Clinical Trial Eligibility Screening: Quantitative Analysis of GPT Models versus Human Expertise. In: Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments, pp. 626–632. ACM, Crete Greece (2024). https://doi.org/10.1145/3652037.3663922 . https://dl.acm.org/doi/10.1145/3652037.3663922 Accessed 2025-11-11

[37] Beattie, J., Neufeld, S., Yang, D., Chukwuma, C., Gul, A., Desai, N., Jiang, S., Dohopolski, M.: Utilizing Large Language Models for Enhanced Clinical Trial Matching: A Study on Automation in Patient Screening. Cureus (2024) https://doi.org/10.7759/cureus.60044 . Accessed 2025-11-11

[38] Hung, T.K.W., Kuperman, G.J., Sherman, E.J., Ho, A.L., Weng, C., Pfister, D.G., Mao, J.J.: Performance of Retrieval-Augmented Large Language Models to Recommend Head and Neck Cancer Clinical Trials. Journal of Medical Internet Research **26**, 60695 (2024) https://doi.org/10.2196/60695 . Accessed 2025-11-11

[39] Ji, Y., Ma, W., Sivarajkumar, S., Zhang, H., Sadhu, E.M., Li, Z., Wu, X., Visweswaran, S., Wang, Y.: Mitigating the risk of health inequity exacerbated by large language models. npj Digital Medicine **8**(1), 246 (2025) https://doi.org/10.1038/s41746-025-01576-4 . Accessed 2025-11-11

[40] Saeidi, M.: Streamlining Clinical Trial Recruitment: A Two-Stage Zero-Shot LLM Approach with Advanced Prompting. In: Hegselmann, S., Zhou, H., Healey, E., Chang, T., Ellington, C., Mhasawade, V., Tonekaboni, S., Argaw, P., Zhang, H. (eds.) Proceedings of the 4th Machine Learning for Health Symposium. Proceedings of Machine Learning Research, vol. 259, pp. 886–896. PMLR, ??? (2025). https://proceedings.mlr.press/v259/saeidi25a.html

[41] Yuan, J., Tang, R., Jiang, X., Hu, X.: Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial Matching. AMIA ... Annual Symposium proceedings. AMIA Symposium **2023**, 1324–1333 (2023)

[42] Rosteck, N., Striegl, J., Loitsch, C.: Bridging the Treatment Gap: A Novel LLM-Driven System for Scalable Initial Patient Assessments in Mental Healthcare. In: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1–8. ACM, Yokohama Japan (2025). https://doi.org/10.1145/3706599.3720043 . https://dl.acm.org/doi/10.1145/3706599.3720043 Accessed 2025-11-11

[43] Chung, D., Sidhom, K., Dhillon, H., Bal, D.S., Fidel, M.G., Jawanda, G., Patel, P.: Real-world utility of ChatGPT in pre-vasectomy counselling, a safe and efficient practice: a prospective single-centre clinical study. World Journal of Urology **43**(1), 32 (2024) https://doi.org/10.1007/s00345-024-05385-4 . Accessed 2025-11-11

[44] Shen, S., Yin, J.: Bayesian mediation analysis using patient-reported outcomes from AI chatbots to infer causal pathways in clinical trials. PLOS One **20**(7), 0326517 (2025) https://doi.org/10.1371/journal.pone.0326517 . Accessed 2025-11-11

[45] Dougherty, R.F., Clarke, P., Atli, M., Kuc, J., Schlosser, D., Dunlop, B.W., Hellerstein, D.J., Aaronson, S.T., Zisook, S., Young, A.H., Carhart-Harris, R., Goodwin, G.M., Ryslik, G.A.: Psilocybin therapy for treatment resistant depression: prediction of clinical outcome by natural language processing. Psychopharmacology **242**(7), 1553–1561 (2025) https://doi.org/10.1007/s00213-023-06432-5 . Accessed 2025-11-11

[46] Lee, K., Paek, H., Huang, L.-C., Hilton, C.B., Datta, S., Higashi, J., Ofoegbu, N., Wang, J., Rubinstein, S.M., Cowan, A.J., Kwok, M., Warner, J.L., Xu, H., Wang, X.: SEETrials: Leveraging large language models for safety and efficacy extraction in oncology clinical trials. Informatics in Medicine Unlocked **50**, 101589 (2024) https://doi.org/10.1016/j.imu.2024.101589

[47] Wadhwa, S., DeYoung, J., Nye, B., Amir, S., Wallace, B.C.: Jointly Extracting Interventions, Outcomes, and Findings from RCT Reports with LLMs. In: Deshpande, K., Fiterau, M., Joshi, S., Lipton, Z., Ranganath, R., Urteaga, I., Yeung, S. (eds.) Proceedings of the 8th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 219, pp. 754–771. PMLR, ???

(2023). https://proceedings.mlr.press/v219/wadhwa23a.html

[48] Schmidt, D.M., Cimiano, P.: Grammar-constrained decoding for structured information extraction with fine-tuned generative models applied to clinical trial abstracts. Frontiers in Artificial Intelligence **7**, 1406857 (2025) https://doi.org/10.3389/frai.2024.1406857 . Accessed 2025-11-11

[49] Liu, J., Lai, H., Zhao, W., Huang, J., Xia, D., Liu, H., Luo, X., Wang, B., Pan, B., Hou, L., Chen, Y., Ge, L.: AI-driven evidence synthesis: data extraction of randomized controlled trials with large language models. International Journal of Surgery **111**(3), 2722–2726 (2025) https://doi.org/10.1097/JS9.0000000000002215 . Accessed 2025-11-11

[50] Rubinstein, S., Mohsin, A., Banerjee, R., Ma, W., Mishra, S., Kwok, M., Yang, P., Warner, J.L., Cowan, A.J.: Summarizing clinical evidence utilizing large language models for cancer treatments: a blinded comparative analysis. Frontiers in Digital Health **7**, 1569554 (2025) https://doi.org/10.3389/fdgth.2025.1569554 . Accessed 2025-11-11

[51] DeYoung, J., Martinez, S.C., Marshall, I.J., Wallace, B.C.: Do Multi-Document Summarization Models *Synthesize* ? Transactions of the Association for Computational Linguistics **12**, 1043–1062 (2024) https://doi.org/10.1162/tacl_a_00687 . Accessed 2025-11-11

[52] White, R., Peng, T., Sripitak, P., Rosenberg Johansen, A., Snyder, M.: 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, pp. 396–402. ACM, Lisbon Portugal (2023). https://doi.org/10.1145/3582515.3609559 . https://dl.acm.org/doi/10.1145/3582515.3609559 Accessed 2025-11-11

[53] Shaib, C., Li, M., Joseph, S., Marshall, I., Li, J.J., Wallace, B.: Summarizing, Simplifying, and Synthesizing Medical Evidence using GPT-3 (with Varying Success). In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 1387–1407. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-short.119 . https://aclanthology.org/2023.acl-short.119 Accessed 2025-11-11

[54] Li, Z., Luo, X., Yang, Z., Zhang, H., Wang, B., Ge, L., Bian, Z., Zou, J., Chen, Y., Zhang, L., ADVANCED Working Group, Wang, Y., Li, H., Zhu, D., Yao, Y., Peng, D., Lai, H., Zhang, J., Wang, F., Duan, M., Li, Y., Tang, S.: RAPID: Reliable and efficient Automatic generation of submission rePortIng checklists with large language moDels. Journal of the American Medical Informatics Association **32**(8), 1340–1349 (2025) https://doi.org/10.1093/jamia/ocaf093 . Accessed 2025-11-11

[55] Wrightson, J.G., Blazey, P., Moher, D., Khan, K.M., Ardern, C.L.: GPT for

RCTs? Using AI to determine adherence to clinical trial reporting guidelines. BMJ Open **15**(3), 088735 (2025) https://doi.org/10.1136/bmjopen-2024-088735 . Accessed 2025-11-11

[56] Al Qurashi, A.A., Hajja, A., Alabdul Razzak, G.F., Alkuwaity, D.W., Chaudhri, E.N., Alharbi, R.I., Al Dwehji, A.M.O., Almusa, H.A., Asaad, A.H., Alobaidi, H.A., Halawani, I.R., Gelidan, A.G.: Evaluating Compliance of Randomized Controlled Trial Abstracts in Plastic Surgery Journals with CONSORT Guidelines Using GPT-4 AI. Plastic and Reconstructive Surgery - Global Open **12**(10), 6227 (2024) https://doi.org/10.1097/GOX.0000000000006227 . Accessed 2025-11-11

[57] Jiang, L., Lan, M., Menke, J.D., Vorland, C.J., Kilicoglu, H.: Text classification models for assessing the completeness of randomized controlled trial publications based on CONSORT reporting guidelines. Scientific Reports **14**(1), 21721 (2024) https://doi.org/10.1038/s41598-024-72130-7 . Accessed 2025-11-11

[58] Lan, M., Cheng, M., Hoang, L., Ter Riet, G., Kilicoglu, H.: Automatic categorization of self-acknowledged limitations in randomized controlled trial publications. Journal of Biomedical Informatics **152**, 104628 (2024) https://doi.org/10.1016/j.jbi.2024.104628 . Accessed 2025-11-11

[59] Ji, C., Zhao, B., Wang, Z., Wang, Y., Zhang, Y., Cheng, Y., Feng, R., Zhang, X.: RoBGuard: Enhancing LLMs to Assess Risk of Bias in Clinical Trial Documents. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics, pp. 1258–1277. Association for Computational Linguistics, Abu Dhabi, UAE (2025). https://aclanthology.org/2025.coling-main.84/

[60] Au, L.S., Qu, L., Nielsen, J., Ge, Z., Gurrin, L.C., Mol, B.W., Wang, R.: Using artificial intelligence to semi-automate trustworthiness assessment of randomized controlled trials: a case study. Journal of Clinical Epidemiology **180**, 111672 (2025) https://doi.org/10.1016/j.jclinepi.2025.111672 . Accessed 2025-11-11

[61] Šuster, S., Baldwin, T., Verspoor, K.: Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. Research Synthesis Methods **15**(6), 988–1000 (2024) https://doi.org/10.1002/jrsm.1749 . Accessed 2025-11-11

[62] Shah-Mohammadi, F., Finkelstein, J.: Addressing Semantic Variability in Clinical Outcome Reporting Using Large Language Models. BioMedInformatics **4**(4), 2173–2185 (2024) https://doi.org/10.3390/biomedinformatics4040116 . Accessed 2025-11-11

[63] Windisch, P., Dennstädt, F., Koechli, C., Schröder, C., Aebersold, D.M., Förster, R., Zwahlen, D.R.: The Impact of Temperature on Extracting Information From Clinical Trial Publications Using Large Language Models. Cureus (2024) https://doi.org/10.7759/cureus.75748 . Accessed 2025-11-11

[64] Pabon, M.A., Vaduganathan, M., Claggett, B.L., Chatur, S., Siqueira, S., Marti-Castellote, P., De Boer, R.A., Hernandez, A.F., Inzucchi, S.E., Kosiborod, M.N., Lam, C.S.P., Martinez, F., Shah, S.J., Desai, A.S., Jhund, P.S., McMurray, J.J.V., Solomon, S.D., Vardeny, O.: In-hospital course of patients with heart failure with improved ejection fraction in the DELIVER trial. European Journal of Heart Failure **26**(12), 2532–2540 (2024) https://doi.org/10.1002/ejhf.3410 . Accessed 2025-11-11

[65] Goldenholz, D.M., Goldenholz, S.R., Habib, S., Westover, M.B.: Inductive reasoning with large language models: A simulated randomized controlled trial for epilepsy. Epilepsy Research **211**, 107532 (2025) https://doi.org/10.1016/j.eplepsyres.2025.107532

[66] Yue, L., Xing, S., Chen, J., Fu, T.: 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. BCB '24. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3698587.3701359 . https://doi.org/10.1145/3698587.3701359

[67] Wang, Y., Fu, T., Xu, Y., Ma, Z., Xu, H., Du, B., Lu, Y., Gao, H., Wu, J., Chen, J.: TWIN-GPT : Digital Twins for Clinical Trials via Large Language Model. ACM Transactions on Multimedia Computing, Communications, and Applications, 3674838 (2024) https://doi.org/10.1145/3674838 . Accessed 2025-11-11

[68] Luo, Y., Wunderink, R.G., Lloyd-Jones, D.: Proactive vs reactive machine learning in health care: Lessons from the COVID-19 pandemic. JAMA **327**(7), 623–624 (2022) https://doi.org/10.1001/jama.2021.24991

[69] Collado-Mesa, F., Alvarez, E., Arheart, K.: Artificial intelligence applied to clinical trials: opportunities and challenges. Health and Technology **13**(2), 203–213 (2023) https://doi.org/10.1007/s12553-023-00738-2

[70] Miller, M.I., Shih, L.C., Kolachalama, V.B.: Machine learning in clinical trials: A primer with applications to neurology. Neurotherapeutics **20**(4), 1066–1080 (2023) https://doi.org/10.1007/s13311-023-01384-2

Appendix A. Data Extraction Form

Table A1: Data extraction variables and operational definitions

Concept	Definition / Extracted Information
Basic Study Information	
Study ID	Unique identifier (Rayyan ID).

Continued on next page

Table A1 continued

Concept	Definition / Extracted Information
Title	Full title of the article.
Authors / Journal	Journal or conference venue; publication status (full paper vs. abstract).
Publication Year / Month	Year and month the study was published.
Medical Domain	Clinical specialty or disease area.
Geographic Focus	Location of data sources or first author affiliation.
Eligibility and Scope	
Study Type	Randomized trial, observational study, simulation, benchmarking, etc.
Eligibility for Review	If excluded, reasons documented (e.g., non-LLM AI, not related to clinical trials).
Trial Stage	Design, Recruitment, Conduct, Analysis, or Multiple.
LLM Application Characteristics	
LLM Utility / Function	Main operational role (e.g., eligibility parsing, patient matching, summarization, document drafting, PRO capture, simulation).
Real-world Deployment	Whether used prospectively in an ongoing trial or retrospectively on existing data.
LLM Family	GPT, LLaMA, Qwen, Claude, Gemini, etc.
LLM Version	Specific model (e.g., GPT–4o, LLaMA–2–70B).
Data Characteristics	
Data Sources	EHR, primary trial data, benchmark datasets, registries, literature, etc.
Data Accessibility	Publicly available / Non-public or institutionally controlled / Mixed.
Dataset Size	Number of patient records, documents, or trials included.
Patient Data Passed to LLM	Yes/No; direct vs. de-identified vs. synthetic.
Privacy and Governance	
IRB Approval Reported	Whether ethical approval was disclosed.
HIPAA or PHI Handling	Whether HIPAA or PHI protections were mentioned.
De-identification Procedures	Description of anonymization pipeline if provided.
Hosting Environment	Local deployment, on-premises secure endpoint, commercial API, or hybrid.
Privacy-Preserving Methods	E.g., de-identification, PHI minimization, synthetic data, sandboxing.
Machine Learning Methods	

Continued on next page

Table A1 continued

Concept	Definition / Extracted Information
ML Task Category	Classification, extraction, summarization, retrieval, simulation, etc.
Enhancement Techniques	Zero-shot, few-shot, CoT, RAG, finetuning, ensembles, etc.
Evaluation and Validation	
Evaluation Purpose	Primary evaluation approach, including technical benchmarking, comparison with computational methods, human-reference comparison, or human evaluation of LLM-generated outputs.
Evaluation Metrics	Reported quantitative or qualitative evaluation measures, including accuracy, precision, recall, F1-score, AUROC, BLEU/ROUGE, agreement measures, expert ratings, and task-specific metrics.
Comparison With Current LLMs	Whether the proposed approach was compared with other contemporary generative LLMs; comparative outcome coded as Better, Not Better, Mixed, or Not Evaluated.
Comparison With Previous Language Models / ML	Whether the proposed approach was compared with previous language models or machine-learning/NLP methods; comparative outcome coded as Better, Not Better, Mixed, or Not Evaluated.
Retrospective Human-Reference Comparison	Whether LLM outputs were retrospectively compared with pre-existing human annotations, decisions, or other human-derived reference standards; evaluation outcome coded as Favorable, Mixed, or Not Favorable where applicable.
Prospective Human Evaluation	Whether human evaluators assessed LLM-generated outputs as part of the study, including ratings of accuracy, completeness, readability, usefulness, actionability, or clinical appropriateness; evaluation outcome coded as Favorable, Mixed, or Not Favorable where applicable.
Human Evaluation Design	Type of human involvement, including expert rating of LLM outputs, comparison with human-authored outputs, or paired evaluation in which humans and LLMs performed comparable tasks.
Code Availability	Whether source code or implementation materials were reported as publicly available.