Panacea: A foundation model for clinical trial search, summarization, design, and recruitment

Jiacheng Lin[1], Hanwen Xu[2], Zifeng Wang[1], Sheng Wang[2#], Jimeng Sun[1#]
[1] Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL
[2] Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA
[#]Corresponding authors. Emails: swang@cs.washington.edu, jimeng@illinois.edu

Abstract

Clinical trials are fundamental in developing new drugs, medical devices, and treatments. However, they are often time-consuming and have low success rates. Although there have been initial attempts to create large language models (LLMs) for clinical trial design and patient-trial matching, these models remain task-specific and not adaptable to diverse clinical trial tasks. To address this challenge, we propose a clinical trial foundation model named Panacea, designed to handle multiple tasks, including trial search, trial summarization, trial design, and patient-trial matching. We also assemble a large-scale dataset, named TrialAlign, of 793,279 trial documents and 1,113,207 trial-related scientific papers, to infuse clinical knowledge into the model by pre-training. We further curate TrialInstruct, which has 200,866 of instruction data for fine-tuning. These resources enable Panacea to be widely applicable for a range of clinical trial tasks based on user requirements.

We evaluated Panacea on a new benchmark, named TrialPanorama, which covers eight clinical trial tasks. Our method performed the best on seven of the eight tasks compared to six cutting-edge generic or medicine-specific LLMs. Specifically, Panacea showed great potential to collaborate with human experts in crafting the design of eligibility criteria, study arms, and outcome measures, in multi-round conversations. In addition, Panacea achieved 14.42% improvement in patient-trial matching, 41.78% to 52.02% improvement in trial search, and consistently ranked at the top for five aspects of trial summarization. Our approach demonstrates the effectiveness of Panacea in clinical trials and establishes a comprehensive resource, including training data, model, and benchmark, for developing clinical trial foundation models, paving the path for AI-based clinical trial development.

Introduction

Clinical trials are research studies conducted on humans to evaluate the safety and efficacy of new medical treatments, interventions, or devices before they are approved for widespread use. They form the foundation of modern medicine.[1–5] The challenges in clinical trials are three-fold. First, a clinical trial involves several interconnected design components, including trial descriptions, eligibility criteria, study arms, outcome metrics, and more, that need to be collectively designed to ensure optimal patient recruitment and outcome assessment. Second, clinical trial data are usually highly sensitive and private, hence often not amenable to pubic cloud-based tools (e.g., GPT-4[6]) for processing and analysis. Third, clinical trial development requires multiple tasks, such as eligibility criteria design and patient recruitment, which require substantial domain expertise.

Machine learning models have shown promise in improving clinical trial development.[7–12] However, current models are often specialized for specific tasks, leading to challenges in managing the resulting models and utilizing training data effectively across interconnected clinical trial activities. Recently, foundation models have been highlighted as the generalist AI that can solve multiple tasks in many biomedical domains.[13–19] For example, GPT-4 was used to assist clinical trial design and trial-patient matching.[7,20–22] We thus hypothesize that a small but specialized clinical trial foundation model could be a Swiss Army Knife tool that simultaneously addresses multiple clinical trial tasks.

We present `Panacea`, a clinical trial foundation model that can address eight clinical trial tasks, including trial design, patient-trial matching, trial search, and trial summarization. The training of `Panacea` consists of an *alignment step* and an *instruction-tuning step*. During the alignment step, we train `Panacea` from a general-domain model using a large collection of trial documents and trial-related scientific papers. This step adapts `Panacea` to the vocabulary commonly used in clinical trials. To conduct the alignment, we create the `TrialAlign` dataset from diverse resources, covering a comprehensive set of indications and medications for any clinical trial. The instruction-tuning step further enables `Panacea` to comprehend the user explanation of the task definition and the output requirement. By leveraging our curated `TrialInstruct` dataset, `Panacea` can handle multiple clinical trial tasks without needing to re-train.

We compared `Panacea` to six cutting-edge large language models on a new clinical trial benchmark `TrialPanorama`. This benchmark covers eight tasks spanning trial design, patient-trial matching, trial search, and trial summarization. Our experiments showed that `Panacea` can facilitate experts through conversations, leading to superior design of eligibility criteria, study arms, and outcome measures. Especially on patient-trial matching, we found that our method achieved, on average, 14.42% F1 improvement on two datasets. On trial search, `Panacea` obtained a 41.78% improvement in query generation and a 52.02% improvement in query expansion. Finally, we propose evaluating trial summaries based on the alignment of their trial goals, conclusions, and keywords with reference summaries. We found that `Panacea` yield the best performance for the challenging multi-trial summarization tasks.

We have made all our training datasets (`TrialAlign` and `TrialInstruct`) and the evaluation benchmark (`TrialPanorama`) available for future research and benchmarking of clinical trial foundation models. Additionally, we have open-sourced the code and model weights of `Panacea`. `Panacea` can run on a single-GPU machine, making it easy to use within an organization. Fine-tuning `Panacea` on 200 thousand documents only takes seven hours using a standard cluster with 4 A-100 GPUs. This advantage allows for further customization of `Panacea` on local proprietary data using limited computational resources.

Table 1: We curate `TrialPanorama` benchmark to evaluate our trial foundation `Panacea` on eight clinical trial tasks spanning trial design, patient-trial matching, trial search, and trial summarization. Here is the summary of the clinical trial tasks, dataset sizes, and evaluation metrics.

Task type	Task name	Metric	Description	Data size		
				Train	Dev	Test
Trial search	Query generation	Jaccard index	Generate searchable queries based on specific clinical trial requirements for database retrieval.	1,837	324	925
	Query expansion	Jaccard index	Broaden search parameters to include related terms and conditions to enhance trial discovery.	43,350	7,650	2,500
Trial summarization	Single-trial summarization	ROUGE, LLM-based metric	Summarize key details and results of individual clinical trials.	4,250	7,50	1,000
	Multi-trial summarization	ROUGE, LLM-based metric	Compile and compare outcomes across multiple clinical trials for comprehensive insights.	1,725	304	252
Trial design	Criteria design	BLEU ROUGE	Define eligibility criteria for patient selection in clinical trials.	30,559	5,392	549
	Study arm design	Clinical relevance	Develop different intervention groups to assess the effects of treatments.	45516	8032	549
	Outcome measure design		Establish methods for measuring trial results and effectiveness of interventions.	38,088	6,721	549
Patient-trial matching	Patient-trial matching	F1, BACC, KAPPA	Match eligible patients with suitable clinical trials, 3-class classification problem	24,146	4,261	11,341

Results

Overview of `Panacea`

Our goal is to develop `Panacea`, a domain-specific foundation model for clinical trial tasks. Like previous works on developing domain-specific foundation models,[23,24] the biggest challenge for developing `Panacea` is to curate the high-quality fine-tuning data to align `Panacea` to clinical trial vocabulary and create the specific instruction data for clinical trial tasks. `Panacea` consists of two main steps: an alignment step, which adapts `Panacea` to the vocabulary used in clinical trials, and an instruction-tuning step, which instructs `Panacea` on each clinical trial task. We built two datasets `TrialAlign` and `TrialInstruct` for the alignment step and the instruction-tuning step, respectively.

`TrialAlign` consists of 793,279 de-identified trial documents collected from 14 diverse sources and 1,113,207 scientific papers related to clinical trials (see **Methods**), representing a large-scale collection of clinical trial documents. By classifying these trial documents to terms in the International Classification of Diseases (ICD-10) ontology, we found that at least 100 conditions have 10,000 documents (**Fig. 1a**), indicating the good coverage of our dataset. Likewise, by classifying trial-related scientific papers to Medical Subject Headings (MeSH) terms, we found that at least 119 terms have more than 10,000 papers and at least 1,921 terms have more than 1,000 papers (**Fig. 1b**). The scale and the coverage of `TrialAlign` enable `Panacea` to be generalized to various conditions and treatments.

`TrialInstruct` contains instruction-tuning data from eight diverse tasks, including criteria design, study arm design, outcome measure design, patient-trial matching, query generation, query expansion, single-trial summarization, and multi-trial summarization, instructing `Panacea` on solving these tasks (**Fig. 1c**). Each task contains at least 2,000 data points, where each data point contains an instruction, an input, and an output (**Fig. 1d**). Since these eight tasks are related, we jointly fine-tuned the model using instruction data from these eight tasks, transforming `Panacea` into an all-in-one tool for clinical trial applications (**Fig. 1e**).

To evaluate `Panacea`, we built the first large-scale benchmark `TrialPanorama` that covers eight specific tasks in clinical trials (**Table 1**). Since these tasks contain both classification and generation tasks, `TrialPanorama` allows us to evaluate `Panacea` in various machine learning settings. We made this benchmark fully open-source.

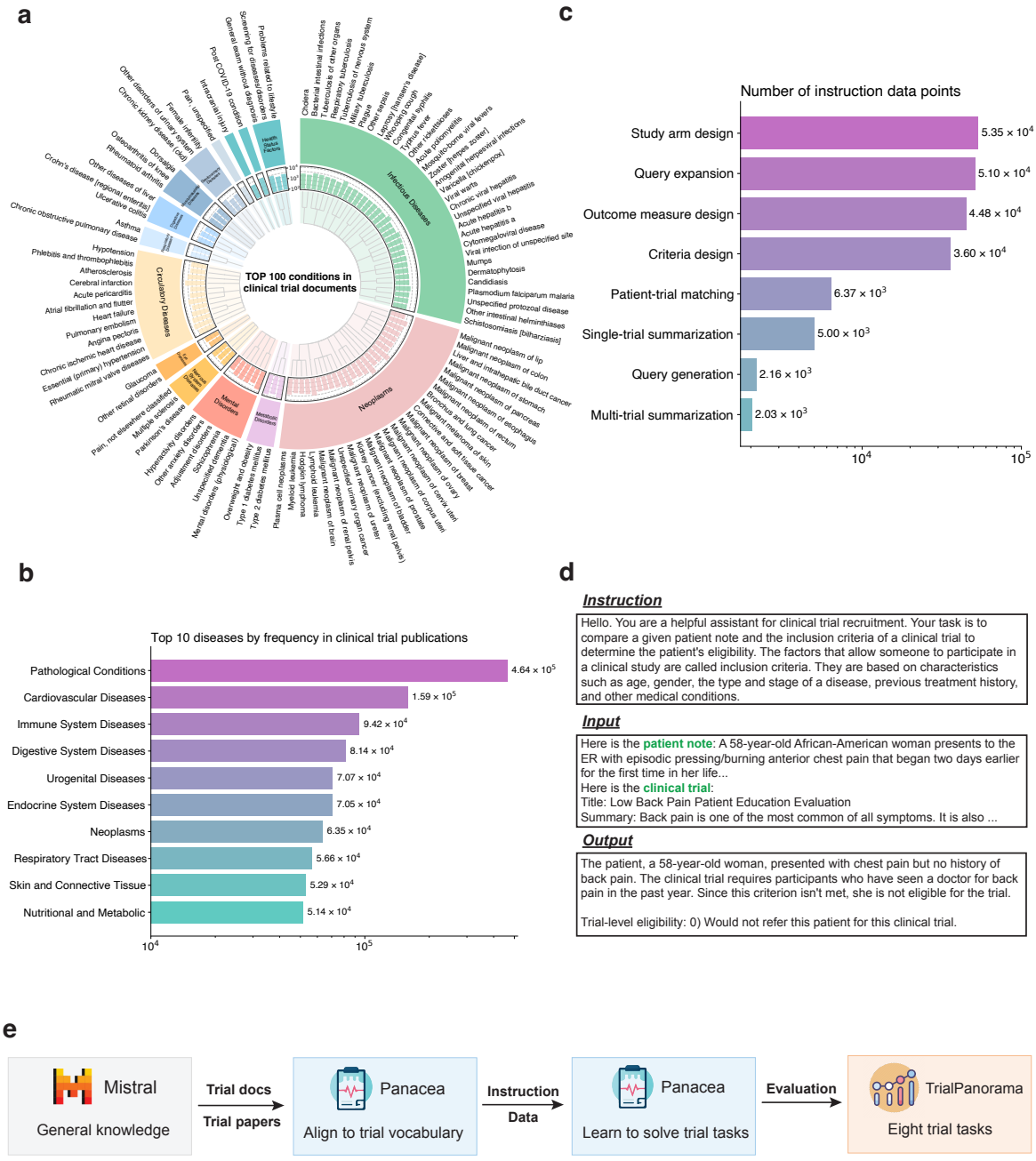

Figure 1: **Overview of `Panacea`. a,** Number of de-identified trial documents in each ICD-10 category. The top 100 conditions with the most number of trial documents are illustrated here. **b,** Bar plot showing the most frequent diseases in clinical trial publications according to the MeSH terms. **c,** Bar plot showing the number of instruction data points per clinical trial task in `TrialInstruct`. **d,** An example of an instruction data point in `TrialInstruct`. **e,** `Panacea` first uses `TrialAlign` to fine-tune Mistral, then uses `TrialInstruct` for instruction tuning. We create `TrialPanorama` benchmark to evaluate `Panacea` and other LLMs on trial tasks.

Accurate trial search through query generation and expansion

Clinical trial search is an important task for clinical trial design and research. Trial designers often need to study similar trials to ensure their design aligns with existing trials. The goal of the trial search is to find relevant trials based on user inputs, which serves as the foundation for designing and matching trials. The key to a successful trial search is to create comprehensive search terms. As a result, we evaluate *query generation*, which converts unstructured user input to a list of keywords (**Fig. 2a**), and *query expansion*, which further expands this keyword list to relevant terms (**Fig. 2b**). These two tasks assess the ability to derive high-quality queries based on user intent, which is crucial for a successful trial search.

We first evaluated query generation by formulating it as a text classification problem that classifies user inputs into specific diseases, interventions, phases, status, and study types. We found that `Panacea` substantially outperformed existing approaches regarding the Jaccard index (**Fig. 2d**). The improvement is larger on diseases and interventions, which are more challenging due to the large number of classes in these two categories (**Fig. 2c**), indicating that `Panacea` can accurately convert user inputs into the structured format that is compatible with downstream machine learning classifiers.

Next, we evaluated query expansion by formulating it as a text generation problem. We did not provide the candidate keywords to the models since real-world keywords might have never been seen in the training trials. Similar to our observations in the query generation, `Panacea` achieved the best results on query expansion in terms of Jaccard index (**Fig. 2e**). We attribute the inferior performance of existing models on query expansion to the lack of fine-tuning on trial-related datasets. In contrast, `Panacea` is fine-tuned on `TrialAlign`, adapting it to the vocabulary used in clinical trials. The promising results of `Panacea` on query expansion and generation demonstrate its ability to precisely understand user intent, providing an accurate tool for finding relevant clinical trials.

A novel metric to evaluate trial summarization

Once similar trials are identified, the next task is to understand those trials via summarization. We evaluated the performance of `Panacea` on trial summarization. We studied both single-trial summarization, which aims to provide a concise summary of a specific trial study (**Fig. 3a**), and multi-trial summarization, which aims to summarize multiple trial studies that study similar conditions and interventions (**Fig. 3b**).

Since it could be biased to evaluate summarization using lexical-based metrics, we propose a novel metric based on large language models (see **Methods**, **Supplementary Figures 1 and 2**). In particular, we provided the ground truth summarization and the model-generated summarization to Claude and asked if these summarizations studied the same problem and made the same conclusion. We found that `Panacea` and comparison approaches can correctly summarize the trial goal, while the summarization of the trial conclusion is less accurate (**Fig. 3c-d**). Moreover, summarizing multiple trials is more challenging than summarizing a single trial based on the proposed metric. Nevertheless, our method still outperformed comparison approaches in summarizing multiple trials, suggesting its potential to assist researchers in extracting key information from many related trial studies.

We further used query generation and query expansion to evaluate trial summarization by extracting diseases, and interventions, and expanding them (**Fig. 3c-d**) from each trial. We examined whether the generated summarization can derive the same keywords as the ground truth summarization. We found that `Panacea` achieved the best performance on three of the six keyword categories while achieving comparable on the other categories. Moreover, we calculated the ROUGE score, which is used as the metric for trial summarization in previous works,[25,26] and observed improved performance by `Panacea` as well on multi-trial

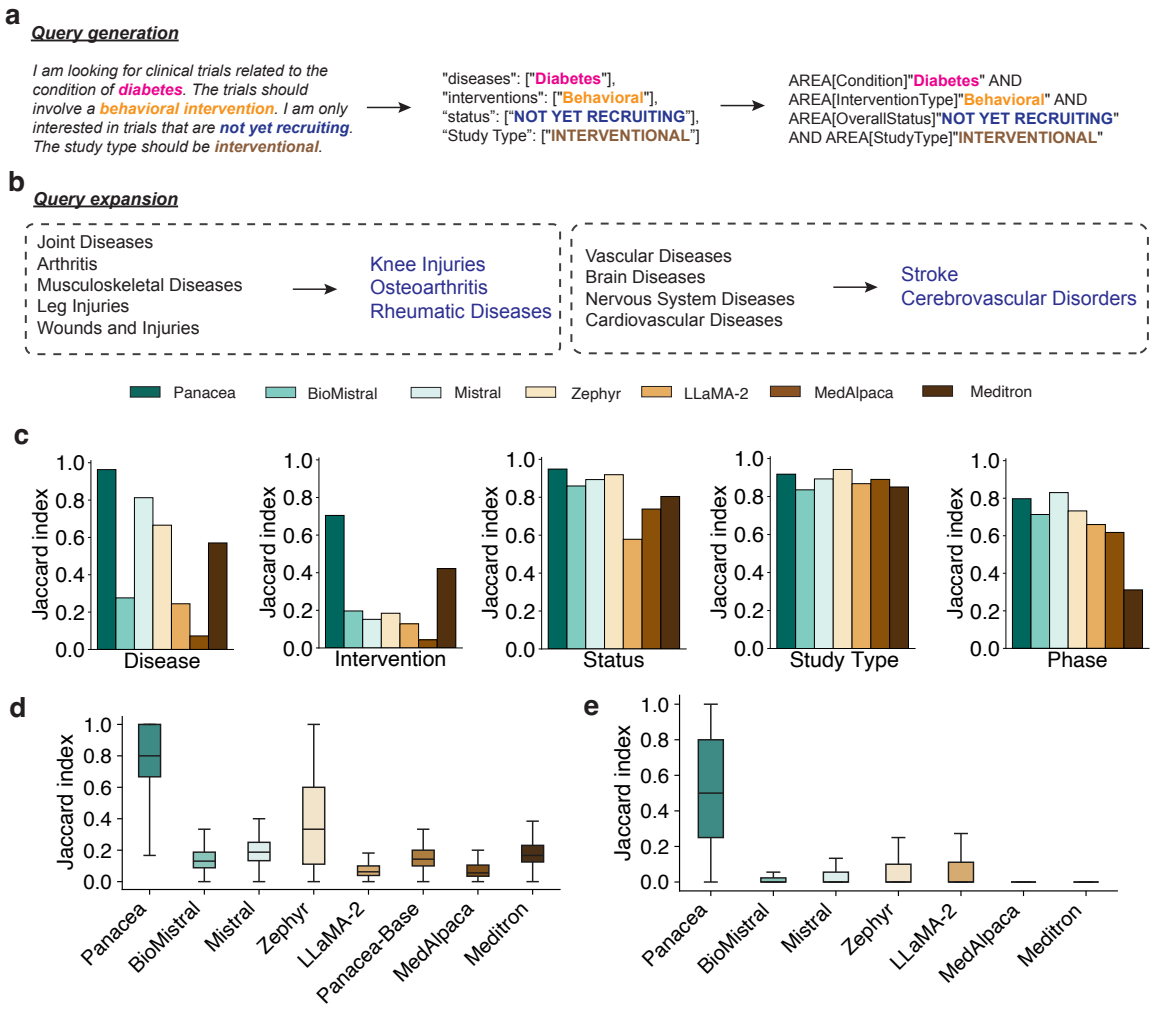

Figure 2: **Evaluation on trial search.** **a,** Query generation aims to convert free text user input into a structured query that contains five categories: disease, intervention, phase, status, and study type. **b,** Query expansion aims to expand a set of keywords. Candidate keywords are not provided. **c,** Comparison of query generation in five specific categories in terms of Jaccard index. **d,** Comparison of query generation in terms of Jaccard index. **e,** Comparison of query expansion in terms of Jaccard index.

summarization (**Fig. 3e**). Finally, we used a case study to show that `Panacea` can correctly summarize the goal and the conclusion for 11 trial studies, while comparison models failed to (**Fig. 3f**).

Improved performance on clinical trial design

The first step toward a successful trial execution is designing a detailed trial protocol synopsis. We evaluated `Panacea` on three tasks in trial design (See examples in **Fig. 4a**): **Criteria design** defines the eligibility criteria (i.e., the inclusion and exclusion criteria) for patient recruitment; **Study arm design** outlines the different treatment arms that will be applied to different patient subgroups; **Outcome measures design** specifies the metrics that are used to assess the trial success. We formulated these three tasks as a conditional text generation problem, which takes conditions, treatments, and the design of previous steps (e.g., reference criteria are used to generate study arms) as inputs to generate specific design text.

Because trials are described in plain text, we first exploited standard natural language processing metrics BLEU and ROUGE to evaluate the lexical similarity. We found that `Panacea` attained the best performance on all three clinical trial design tasks in terms of BLEU and ROUGE (**Fig. 4b**). First, we observed that `Panacea` substantially outperformed general-domain models, including our base model Mistral,[27] confirming the benefit of fine-tuning using clinical trial-related data. Second, we found that `Panacea` improved the study arm design more than the other two tasks. Compared to criteria and outcome measures, study arm descriptions are more customized according to the disease and the treatment. The larger improvement of `Panacea` on study arms design demonstrates `Panacea`'s strong generalization ability. Finally, BioMistral,[28] which is fine-tuned on general biomedical data, also outperformed Mistral, further demonstrating the value of domain-specific data. Nevertheless, `Panacea` still outperformed BioMistral by fine-tuning using our clinical trial-specific data `TrialAlign` and `TrialInstruct`, suggesting that data with improving domain specificity leads to better performance.

Lexical similarity metrics are widely used to evaluate text generation problems, but might not be clinically specific enough to evaluate the generations by `Panacea`. Recently, LLMs have been used to evaluate the generated text by exploiting their strong ability in text understanding. Here, we exploit Claude[29] to evaluate these three tasks by asking the model whether the generated task is clinically relevant (see **Methods**, **Supplementary Figures 3-5**). We found that `Panacea` outperforms all methods on criteria and study arms design, demonstrating the high quality of generation by `Panacea` (**Fig. 4b**).

Moreover, we examined a De Novo generation setting, using the generated output in the previous step as the input for the next step. For example, we used the generated criteria instead of the reference criteria as the input for generating study arms. De Novo generation frees users from providing any descriptions for the trial. We found that the performance of all methods dropped in this setting compared to the setting that utilizes reference input (**Fig. 4c**). Nevertheless, our method still outperforms all existing methods by a large margin, indicating its superior performance on this De Novo trial design. We further compared the generated text by three methods with the ground truth text on criteria design, where only `Panacea` can generate the correct criteria (**Fig. 4d**). Collectively, the promising performance of `Panacea` demonstrates its potential to automate clinical trial design.

Accurate patient-trial matching

We next evaluate the performance of `Panacea` on patient-trial matching. Given a patient note and a trial description, we aim to determine whether this patient is eligible for the trial by

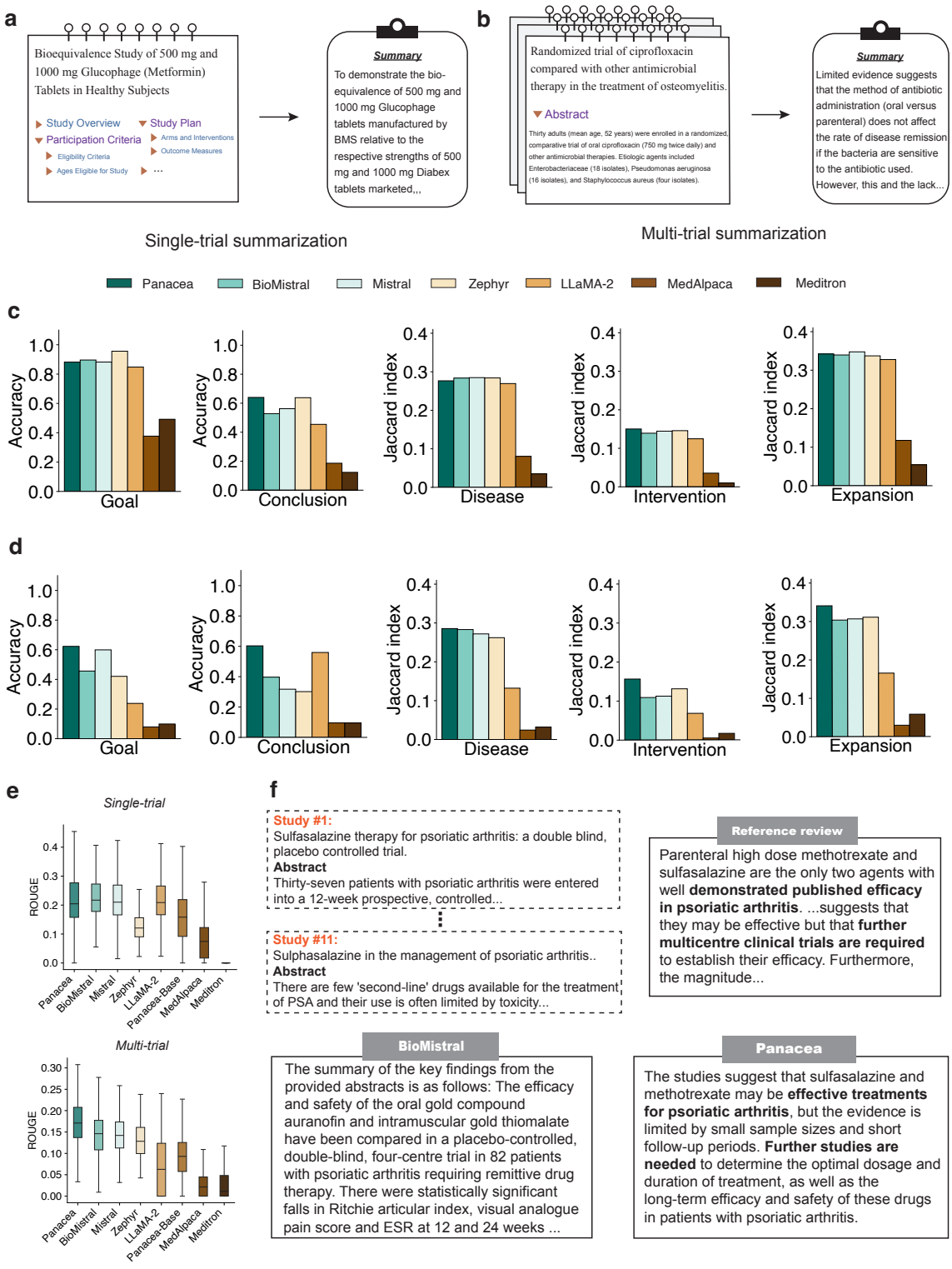

Figure 3: **Evaluating `Panacea` on trial summarization. a,b,** Trial summarization aims to provide a concise summary, including trial goal and conclusion, for a single trial (**a**) or multiple trials (**b**). **c,d,** Evaluation on single-trial summarization (**c**) and multiple-trial summarization (**d**) by using Claude-based metric and trial search-based metrics. **e,** Comparison on trial search in terms of ROUGE. **f,** A case study illustrating how `Panacea` successfully summarize multiple studies.

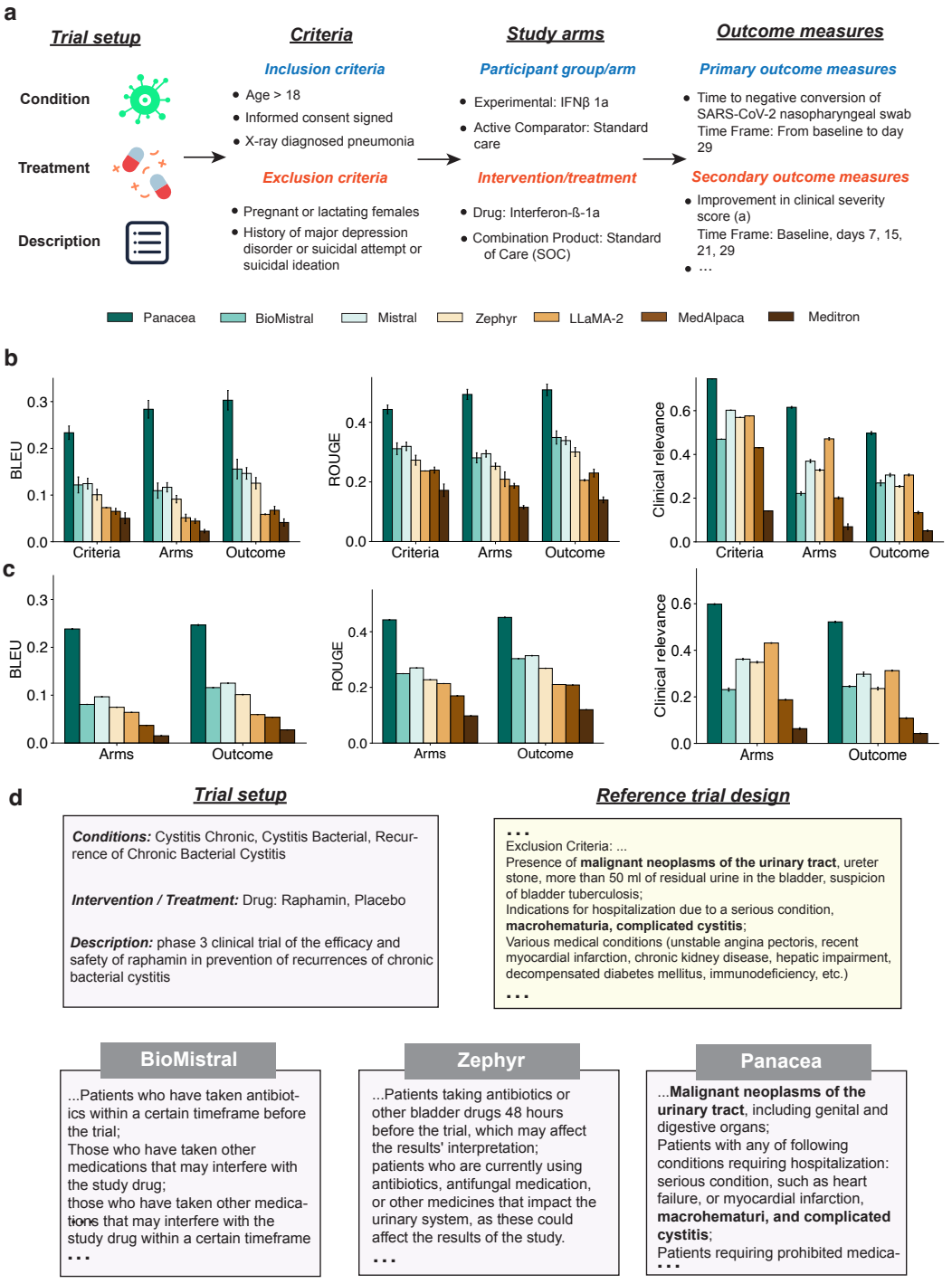

Figure 4: **Evaluation on clinical trial design. a,** Problem setting of clinical trial design, which aims to generate criteria, study arms, and outcome measures. Criteria are used as input to generate study arms. Criteria and study arms are used to generate the outcome measures. **b,** Evaluation on trial design in terms of BLEU, ROUGE, and clinical relevance, where the reference design in the previous step is given as the input to the next step. **c,** Evaluation on trial design in terms of BLEU, ROUGE, and clinical relevance, where the generated design in the previous step is given as the input to the next step. **d,** A case study comparing criteria generation by different methods. `Panacea` can generate criteria that match the reference trial design.

formulating this problem as a three-class classification task: eligible, excluded, or irrelevant (**Fig. 5a**).

We first evaluated our method on the TREC 2021 dataset,[30] which consists of a training set and a test set. We used the training set to construct instructions in `TrialAlign`, and then assessed the performance of `Panacea` on the test set. We found that `Panacea` outperformed all comparison approaches in terms of balanced accuracy (BACC), Cohen's KAPPA score, Recall, Precision, and F1, indicating the effectiveness of using `TrialInstruct` to fine-tune the model (**Fig. 5b-f**). To investigate the generalizability of our method, we further tested our method on the SIGIR dataset[31] where the entire dataset is used as the test set. We found that our method again attained the best performance on all three metrics, demonstrating the strong generalizability of our method.

As the eligible class is crucial for patient-trial recruitment, we further examined a binary classification setting. In this setting, we grouped "excluded" and "irrelevant" into one category, and "eligible" into the other in order to determine whether a patient is eligible for a trial. Our method outperformed all comparison approaches in terms of F1, precision, and recall, indicating its applicability to real-world trial recruitment (**Fig. 5g-i**). Finally, we used a case study to illustrate how our method successfully classified a patient as eligible by examining each criterion and coming to a conclusion based on their criteria (**Fig. 5j**). In contrast, LLaMA-2[32] made an incorrect conclusion by hallucinating an exclusion criterion not stated in the trial description.

Discussion

In this paper, we introduce a specialized foundation model called `Panacea` for use in clinical trials. We tested `Panacea` in eight different clinical trial tasks, including trial design, patient-trial matching, trial search, and trial summarization. In comparison to other general domain foundation models and biomedical foundation models, `Panacea` demonstrated state-of-the-art performance across all eight tasks. We believe that the impressive performance of `Panacea` can be attributed to the fine-tuning process using `TrialAlign` and `TrialInstruct`. `TrialAlign` comprises a large collection of trial documents and papers from various areas, allowing `Panacea` to be applied to different conditions and treatments. Meanwhile, `TrialInstruct` contains 200,866 instructions curated from existing databases, effectively guiding `Panacea` in each task. Furthermore, we have developed a clinical trial benchmark `TrialPanorama` and a language model-based metric for evaluating trial summarization. Together, these resources offer an end-to-end solution for AI-based clinical trial development.

The rapid development of large language models (LLMs) has enabled their potential as foundational models for medical tasks.[14] Current efforts predominantly follow two strategies: fine-tuning general domain LLMs with medical domain datasets,[33–35] and instructing a general domain LLM with a description of the target tasks and showing example inputs and outputs (referred to as "prompting").[36–38] The MedPaLM model is a prime example of the first approach, illustrating how fine-tuning a general domain model on medical datasets can markedly enhance its ability to answer medical questions.[34] This success has inspired further research into fine-tuning LLMs for specific clinical trial tasks, such as generating eligibility criteria.[7] Moreover, it has been demonstrated that generalist LLMs can be effectively adapted to medical tasks through strategic prompting.[38] In the direction of prompting, TrialGPT showcased that GPT-4 can be adapted to predict patient eligibility for clinical trials through prompting.[20] However, these approaches either do not address clinical trial tasks or focus on individual clinical trial-related tasks. In contrast, `Panacea` outlines a comprehensive range

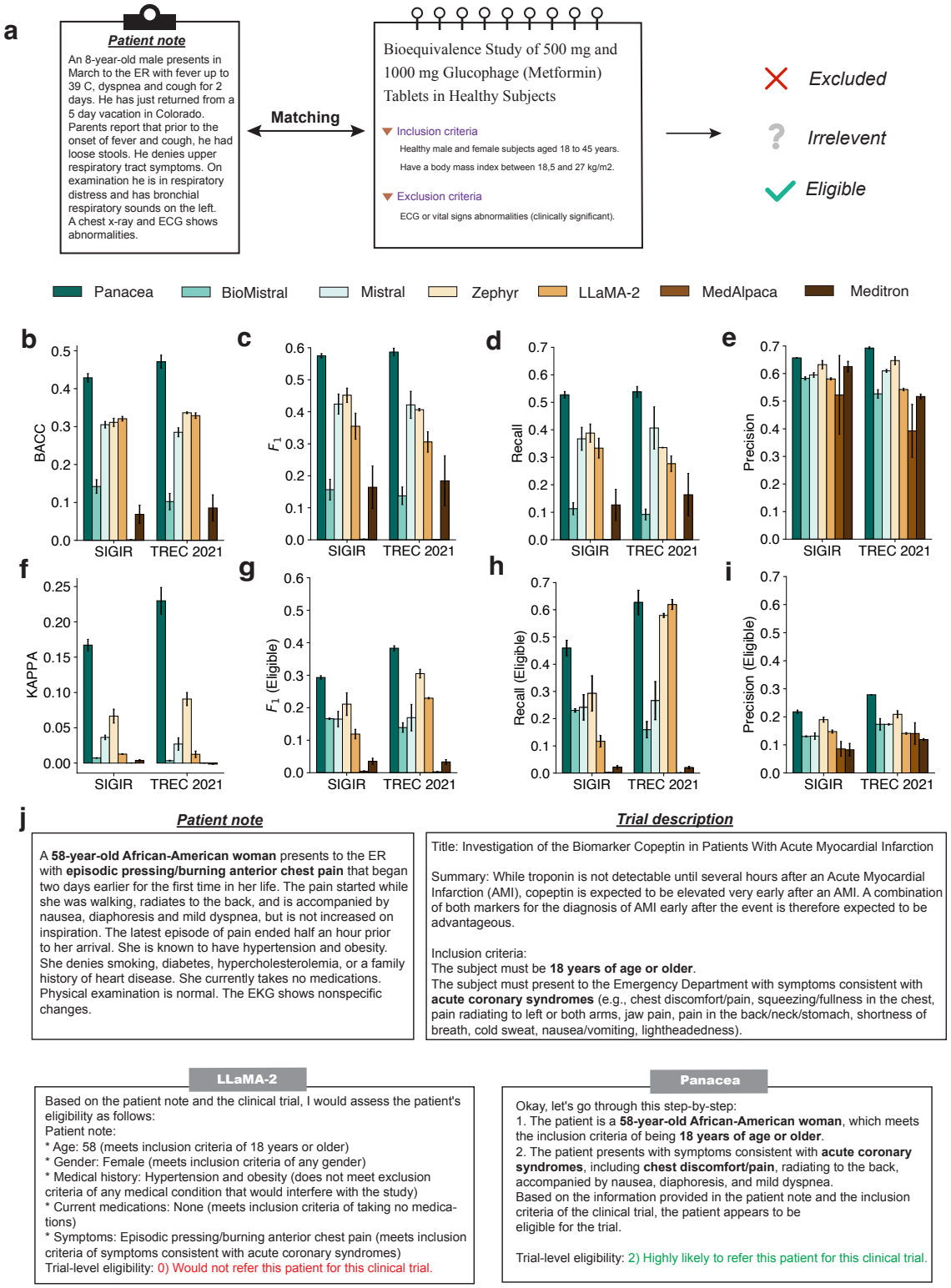

Figure 5: **Evaluation on patient-trial matching.** **a,** Problem setting of patient-trial matching, which classifies a patient into three categories based on the patient note and the trial description. **b-f** Comparison on two patient-trial matching datasets SIGIR and TREC 2021 in terms of balanced accuracy (BACC) (**b**), Cohen's KAPPA (**c**), recall (**d**), precision (**e**), and F1 (**f**). **g-i** Comparison on classifying patients into eligible and ineligible in terms of F1 (**g**), precision (**h**), and recall (**i**). **j,** A case study illustrating how `Panacea` successfully classifies the patient into eligible by examining each criterion.

of clinical trial tasks suitable for AI assistance, establishing the first versatile foundational model specifically designed for clinical trial applications.

This study has several limitations that we would like to address in the future. First, despite being fine-tuned on clinical trial instruction datasets, LLMs may still produce biased or low-quality outputs. Enhancing model alignment such as reinforcement learning from human feedback[39] is crucial future work before `Panacea` can be deployed in production settings. Second, for high-stakes applications such as clinical trials, it is essential to detect and regulate LLM hallucinations, which can occur, particularly in areas not well-covered by the LLM training data. It is worth exploring to enable LLMs to either reject an answer[40] or utilize external knowledge bases to correct its outputs.[41] Third, continually updating the model's knowledge is vital for maintaining relevance and accuracy in a rapidly evolving medical landscape. Therefore, it is worth exploring efficient knowledge updating techniques for `Panacea`[42] or enhancing it with retrieval-augmented generation.[43] Fourth, although `Panacea` demonstrates significant improvements across various benchmark datasets, there is a need to develop more evaluation metrics to comprehensively assess LLM performance in more clinical trial tasks. Additionally, conducting user studies could further demonstrate the benefits of `Panacea` in assisting experts with clinical development projects.

Method

Creating `TrialAlign` dataset

Data collection We first collected trial documents (English version) from 14 sources, as shown in **Supplementary Table 1**. Each clinical trial data consists of various parts that encapsulate the essence of the study. For instance, the "Study Overview" provides a general summary and a detailed description of the trial, along with its official title and the health conditions being targeted. The "Intervention/Treatment" section describes the medical approach or therapy being tested. The "Eligibility Criteria" outlines who can participate, detailing the eligibility requirements, age, and sex specifications, and whether healthy volunteers are accepted. The "Study Plan" delves into the methodology, explaining the design of the study, the types of interventions and arms involved, and the outcomes being measured, both primary and secondary. This structured approach ensures a comprehensive understanding of the trial's scope, methodology, and intended outcomes. We then collected trial papers in two databases, i.e., Embase and PubMed, from Cochrane Library's trial section.[44] These papers provide a rich foundation of medical knowledge and evidence-based findings beneficial to the model's learning.

Filtering For trial documents, we further conduct intra- and inter-source de-duplication and then remove the personally identifiable information (PII), finally obtaining 793k trial document data. Further, to avoid information leakage, we selected documents with registration dates before 2023-01-01 as the training corpus. The remaining is used for test data curation. For trial papers, we de-duplicated all the papers and the final 1.11M trial paper corpus consists of abstracts of all the papers and full text of 97k papers from PubMed Central (PMC). Similarly, to avoid information leakage, we choose papers published before 2023-01-01, which ensures the dates of related clinical trials of the selected papers are definitely before 2023-01-01.

Document/paper structure organization For trial documents, we follow the format shown in clinicaltrial.gov[45] to organize all the corpus for alignment. Each trial document is arranged into a markdown format passage. For trial documents from clinicaltrial.gov, each document contains section (1) "Public Title"; (2) "Study Overview" covering subsections "Brief Summary", "Detailed Description", "Official Title", "Conditions" and "Intervention/Treatment"; (3) Participation Criteria, including subsections "Eligibility Criteria", "Ages Eligibility for Study", "Sexes Eligibility for Study" and "Accepts Healthy Volunteers"; (4) "Study Plan", including subsection "How is the study designed?" that contains "Design Details" and "Arms and Interventions", subsection "What is the study measuring?" containing primary and secondary outcome measures; (5)Terms related to the study. For trial documents from other sources, each document contains "Public Title", "Scientific Title", "Study Type", "Study Design", "Intervention", "Inclusion Criteria", "Exclusion Criteria", "Primary Outcome Measures" and "Secondary Outcome Measures". For trial paper data, each paper contains "Title", "Abstract" and full text (if any).

Creating `TrialInstruct` dataset

The aim of constructing `TrialInstruct` is to provide `Panacea` with the ability to follow human instructions, especially in clinical trial domains.

Trial search Trial search includes query generation and query expansion. To construct instruction data for query generation, we leverage GPT-3.5 to generate 2,161 samples for training and 925 for the test. Specifically, we first manually construct 20 seed data about query generation customized for clinicaltrial.gov database API, and then leverage GPT-3.5 to generate the data. We will remove data similar to the original data and add them to the

seed dataset to repeat the above process (see prompt in **Supplementary Figure 6**). In the final stage, we send requests with these generated data to the clinicaltrial.gov database and remove those without any search results. For query expansion data curation, we turn to the mesh terms section in clinicaltrial.gov documents. Each document contains synonymous mesh terms. We keep five terms for each document as input and the others as output. For example, the input mesh terms are Gastroenteritis, Gastrointestinal Diseases, Digestive System Diseases, Colonic Diseases, Intestinal Diseases, Pathologic Processes, while the output terms are Inflammatory Bowel Diseases, Ulcer, Anti-Bacterial Agents, and Vancomycin. We select documents before 2023-01-01 for training and after 2023-01-01 for test. We finally obtained 50k training data and 2,500 test data.

Trial summarization Trial summarization contains single-trial and multi-trial summarization. To curate single-trial summarization data, we leverage clinicaltrial.gov documents. Specifically, the brief summary section serves as the output and the other parts serve as the input. We finally have 5k training data (before 2023-01-01) and 1k test data (after 2023-01-01). For the multi-trial summarization data curation, we derived our dataset from Cochrane dataset of systematic reviews,[46] i.e., we only selected data pairs containing clinical trial papers. Specifically, each multi-trial summarization data contains a PMID set and a review paper. The review is a high-level conclusion from papers in the PMID set. The data curation process started with the matching between the PMID sets and all the trial paper PMIDs in `TrialAlign`. We select those data pairs with at least three trial-related papers in the PMID set. We finally constructed 2,029 samples for training and 252 for test, derived from the Cochrane dataset's training and validation sets due to the missing test labels in the original Cochrane dataset.

Trial design We construct multi-turn conversation data for trial design due to the difficulty of one-turn design, even for frontier models like GPT-4.[6] Such conversation format data are more realistic and benefit users to get more accurate designs as conversations progress. To construct these conversation data, we focus on trial documents in clinicaltrial.gov and adopt a two-stage strategy to construct the conversation data. For criteria design, we first input criteria and trial setup, which contains title, conditions, drugs, and phase, to ask GPT-3.5 to output the reasons for designing those criteria one by one. In the second stage, we input the criteria, and reasons generated in the first stage, and trial setup, to ask GPT-3.5 to construct multi-turn conversation data (see **Supplementary Figure 7**). This can ensure that GPT-3.5 generated trial part data is actual. Likewise, for study arm design, we input study arms, criteria, and trial setup. In the second stage, we collect the generated conversation data given the study arms, reasons, criteria, and trial setup (see **Supplementary Figure 8**). For outcome measures, the input in the first stage is outcome measures, study arms, criteria, and trial setup, while the input in the second stage is outcome measures, reasons, study arms, criteria, and trial setup (see **Supplementary Figure 9**). We use trial documents from clinicaltrial.gov to construct these data, before 2023-01-01 for training and after 2023-01-01 for testing. We finally obtained 35,951 and 549 for the criteria design's training and test set, 53,548 and 549 for the study arm design, and 44,809 and 549 for the outcome measure design.

Patient-trial matching We converted existing representative patient-trial matching datasets into instruction format, i.e., SIGIR[31] and TREC 2021[30] cohorts. Each instruction data of patient-trial matching follows the structure: "Instruction", "One-shot demonstration", "Input patient notes", "Input Criteria" and "Output trial-level eligibility", as illustrated in **Supplementary Figure 10**. We split the TREC 2021 into the training (28,406 samples) and test sets (7,424 samples), and all SIGIR data serves as the test set (3,869 samples). Specifically, the patient-criteria pairs of 80% of patients in TREC 2021 formed into the training set, while those pairs of the remaining 20% of patients in TREC 2021 are test data. For evaluation, we trained our `Panacea` on the training set derived from TREC 2021 and evaluated on the test

set of TREC 2021 and all data in SIGIR.

Creating `TrialPanorama` benchmark

We built the first large-scale benchmark `TrialPanorama`, including eight tasks in clinical trials. The training and test data constructed in the previous section are viewed as the benchmark data. We evaluated the models on `TrialPanorama` to assess each model's performance across different clinical trial tasks.

Details of `Panacea` model

In this section, we detail the techniques in `Panacea`, including the alignment and instruction finetuning steps.

Alignment We built on the Mistral-7B-Base model[27] in this study. After parameter initialization, `Panacea` was trained on the 1.8M `TrialAlign` data. We trained the model using the AdamW optimizer[47] with a batch size 512 for one epoch. We adopted a cosine learning rate scheduler with a peak learning rate 2×10^{-6} and 10% warm-up steps. We set max sequence length as 8192 tokens. To improve training speed and optimize the memory, we adopted DeepSpeed ZeRO-3[48] and FlashAttention-2[49] strategies. After the alignment process, we obtain the `Panacea`-Base model. During the alignment step, `Panacea` was trained on 4 Nvidia A100 80G for four days.

Instruction tuning We further finetuned `Panacea`-Base on the `TrialInstruct` datasets, leading to the `Panacea` model. We trained our `Panacea` for one epoch with a batch size 256. Similar to the alignment step, we also leveraged a cosine learning rate scheduler with a peak learning rate as 2×10^{-5} and 10% warm-up steps. The max sequence length is set as 2048. Deep ZeRO-3 and FlashAttention-2 techniques are also adopted in the instruction tuning phase.

Details of experiments on trial search

In the trial search experiments, we focused on optimizing `Panacea` for two tasks: query generation and query expansion (see **Supplementary Figure 11**). These two tasks are pivotal for enhancing the efficiency and precision of searches within large clinical trial databases.

Query generation in this context essentially functions as a Named Entity Recognition (NER) task where the model identifies and categorizes key pieces of information from the trial descriptions relevant to user queries. To facilitate the generation of structured queries in a JSON format, we employed a specialized tool called JsonFormer.[50] This tool is instrumental in guiding the model to generate content for each key in the JSON structure sequentially.

Once the JSON format is generated, it is automatically converted into a Search Expression using a rule-based system. The conversion rules are straightforward: within the same key, terms are combined using the OR operator, and between different keys, the terms are combined using the AND operator. This structured approach ensures that the generated queries are precise and align well with the syntactical requirements of the search engines used in clinical trial databases.

For the query expansion task, this process enhances the original query by adding semantically related terms, thereby broadening the search scope to include relevant trials that may not use the exact phrasing of the original query terms. `Panacea` was trained to suggest additional keywords based on the initial input terms. The model learned to recognize and predict related terms that could be associated with the initial query, expanding the search breadth effectively.

Details of experiments on trial summarization

The experiments on trial summarization were designed to test `Panacea`'s capabilities in condensing complex clinical trial information into succinct summaries. This component of our research focused on two specific tasks: single-trial summarization and multi-trial summarization (see **Supplementary Figure 12**).

To evaluate summarization tasks, we propose a novel metric based on Claude 3. We use Claude 3 to decide whether the model-generated summarization and the ground truth summarization studied the same problem and made the same conclusion, following prompts in **Supplementary Figure 1 and 2**. Specifically, Claude 3 directly outputs the goal alignment results for each test sample. For conclusion consistency, we first use Claude to evaluate model-generated summaries and ground truth summaries, respectively. Then, we calculate the matching accuracy between the model-generated summarization and ground truth summarization.

Details of experiments on clinical trial design

In our experimental setup for evaluating the `Panacea` model's capabilities in clinical trial design, we utilized a multi-turn conversation format for the test data. This format consists of sequential (user, chatbot) pairs, reflecting a realistic interaction scenario where the model, acting as a chatbot, responds to user queries about designing a trial. The initial three rounds usually provide essential background information related to the trial design, such as the trial's objectives, target population, and key endpoints. These initial conversations set the stage for the more complex interactions that follow. Starting from the fourth round of conversation, the model is tasked with predicting the chatbot's responses based on the cumulative conversation history, which tests the model's ability to maintain context and continuity over successive interactions.

To ensure the reliability of the experimental results and prevent the propagation of errors through the conversation chain, a teaching forcing strategy was implemented: regardless of the model's output in any given round, the subsequent round's input incorporates the groundtruth from the previous rounds rather than the model-generated responses. This method allows the model to be evaluated on its ability to adhere closely to a scientifically valid trial design path without being influenced by potential errors in its previous outputs.

To assess the relevance between models' designed trials and ground truth, we employ Claude 3 to calculate clinical relevance. Specifically, we input each model's output and the ground truth into Claude 3 to determine the relevance of the information generated by the model compared to the ground truth. The inputs to Claude 3 for clinical relevance evaluation are detailed in **Supplementary Figures 3, 4, and 5**, respectively. When a model's outputs are relevant to the ground truth, Claude will output a 1; otherwise, it outputs a 0. We then calculate the clinical relevance using the following formula:

$$\text{Clinical relevance} = \frac{\sum(\text{Relevance scores})}{N} \tag{1}$$

Here, "Relevance scores" refer to the series of 1s and 0s output by Claude 3 for each comparison between a model's output and the ground truth. N is the total number of outputs evaluated. This proportion reflects the percentage of times the model's output was deemed clinically accurate relative to the ground truth, quantifying the frequency at which the model produces clinically relevant information.

Details of experiments on patient-trial matching

In the patient-trial matching experiments, we employed a distinctive approach to training the `Panacea` model, focusing not on utilizing the entirety of the training data but rather on a selected subset. Initially, all available training data was subjected to a filtering process with Claude 3 Haiku. This involved predicting responses for each instance in the training set. Only those instances where Claude 3 Haiku's predictions were accurate were retained for further processing. The rationale was to ensure that the model was learning from correctly reasoned examples and that the training data was high quality. The responses generated by Claude 3 Haiku, which correctly matched the groundtruth data, were then used as the new training corpus for `Panacea`. This step was crucial because the standard training datasets for patient-trial matching typically include labels indicating eligible or excluded but lack a detailed reasoning process for these outcomes. By incorporating Claude 3 Haiku's responses, which involve step-by-step reasoning based on the input data, we injected reasoning capabilities into `Panacea` during the training process. Through this innovative training approach, `Panacea` showed superior performance in patient-trial matching tasks. The ability to reason and logically process eligibility criteria translated into higher accuracy and reliability in matching patients to appropriate trials. The evaluation prompt for patient-trial matching can be seen in **Supplementary Figure 10**.

The patient-trial matching is a three-class classification task for both SIGIR and TREC2021 datasets. Three classes for SIGIR are: 0) Would not refer this patient for this clinical trial; 1) Would consider referring this patient to this clinical trial upon further investigation; and 2) Highly likely to refer this patient for this clinical trial, while TREC2021 has: 0) Excluded (patient meets inclusion criteria, but is excluded on the grounds of the trial's exclusion criteria); 1) Not relevant (patient does not have sufficient information to qualify for the trial); and 2) Eligible (patient meets inclusion criteria and exclusion criteria do not apply).

Code and data availability

The `TrialAlign` data for the alignment step, the `TrialInstruct` data for the instruction tuning step, and the `TrialPanorama` benchmark data are available at https://figshare.com/articles/dataset/TrialAlign/25989403, https://doi.org/10.6084/m9.figshare.25990090.v1, and https://doi.org/10.6084/m9.figshare.25990075, respectively. Panacea code is available at https://github.com/linjc16/Panacea.

Claude 3 evaluation for single-trial summarization

Please evaluate the following generated summaries based on the given criteria. Assign a score from 0 to 1 for each criterion, where 0 indicates the lowest performance and 1 indicates a higher level of performance.

[Groundtruth Summary]
{groundtruth}
[End of Groundtruth Summary]
[Generated Summary]
{input}
[End of Generated Summary]

Evaluation Criteria:
Goal Alignment:
Score 0: The goals described in both summaries are completely different.
Score 1: The goals described have partial overlap or similarity.

After assessing each criterion, provide a brief explanation for each score. Finally, summarize your scores in the following format for clarity:
Goal Alignment: 0

Summary: {input}
Based on this summary, is this trial study effective or not.
If effective, output 1, otherwise output 0.
Directly output the number.
Output:

Supplementary Figure 1: Prompt for evaluation metrics on single-trial summarization.

Claude 3 evaluation for multi-trial summarization

Please evaluate the following generated summaries based on the given criteria. Assign a score 0 or 1 for each criterion, where 0 indicates the lowest performance and 1 indicates a higher level of performance.

[Generated Summary]
{input}
[End of Generated Summary]

[Groundtruth Summary]
{groundtruth}
[End of Groundtruth Summary]

Evaluation Criteria:
Topic Alignment:
Whether the topic in generated summary is similar to the groundtruth summary. If similar, output 1; otherwise, output 0.

After assessing each criterion for the score. Finally, summarize your scores in the following format for clarity:
Topic Alignment: 0. Only output the number without explanation.

Summary: {input}
Based on this input summary, is there enough evidence or not.
If enough, output 1, otherwise output 0.
Directly output the number.
Output:

Supplementary Figure 2: Prompt for evaluation metrics on multi-trial summarization.

Input to Claude 3 for clinical relevance evaluation (Criteria Design)

Act as an impartial judge and evaluate whether the criteria mentioned in a model's output are present in the full list of the groundtruth criteria.
Output '1' or '0', where '1' means the criteria mentioned in the model's output are fully included in the groundtruth criteria list, and '0' means the criteria from the model's output are not included in the groundtruth.
You should provide an explanation for the evaluation.

Example:
[Model Output]
Excellent! Moving on to the third criterion, I propose "Ability to provide written informed consent." Informed consent is a fundamental ethical requirement in clinical research. Participants must fully understand the trial and voluntarily agree to participate.
[End of Model Output]
[Groundtruth Criteria list]
Inclusion Criteria:~Age between 18 and 120 years at time of consent~Ability to speak and understand English~Clinical stage I, II or IIIa NSCLC~Candidate for RTS segmentectomy, as determined by the operating surgeon~Exclusion Criteria:~Anticoagulation with inability to cease anticoagulant therapy prior to surgery~Incurable coagulopathy~Systemic vascular disease or vasculitis~Not a candidate for RTS segmentectomy
[End of Groundtruth Criteria]
Match prediction: 0

Now, evaluate the following model output and groundtruth criteria list.
You should first output the 'match prediction' at the beginning of the response by `Match prediction: `, e.g., `Match prediction: 1`. Then, Provide an explanation for your evaluation.

[Model Output]
{model_output}
[End of Model Output]

[Groundtruth Criteria list]
{groundtruth}
[End of Groundtruth Criteria]

Supplementary Figure 3: Prompt used to calculate clinical relevance for criteria design.

Input to Claude 3 for clinical relevance evaluation (Study Arms)

Act as an impartial judge and evaluate whether the study arms mentioned in a model's output are present in the full table of groundtruth study arms.
Output '1' or '0', where '1' means the study arms mentioned in the model's output are fully included in the groundtruth study arm table, and '0' means the study arms from the model's output are not included in the groundtruth.
You should provide an explanation for the evaluation.

Example:
[Model Output]
The placebo comparator arm, which we'll call "Control: Placebo," will also include obese subjects with Type 2 Diabetes at risk of Nonalcoholic Steatohepatitis. Participants in this arm will receive a placebo, which will be designed to mimic the appearance of the active treatment but will not contain any active drug. The primary purpose of this arm is to compare the safety and efficacy of HU6 to the placebo, to determine if any observed effects are due to the active treatment or could be attributed to other factors.
[End of Model Output]
[Groundtruth Study Arm]
Participant Group/Arm	Intervention/Treatment
Experimental: Active Treatment: HU6 Planned doses of HU6 	Drug: HU6 * HU6 is being evaluated for its efficacy in improving liver fat content in obese subjects with Type 2 Diabetes at risk of Nonalcoholic Steatohepatitis (NASH)
Placebo Comparator: Placebo Comparator Non-active study drug 	Other: Placebo * Placebo
[End of Groundtruth Study Arm]
Match prediction: 1

Now, evaluate the following model output and groundtruth study arm table. You should first output the 'match prediction' at the beginning of the response by Match prediction: , e.g., Match prediction: 1. Then, provide an explanation for your evaluation.

[Model Output]
{model_output}
[End of Model Output]

[Groundtruth Study Arm]
{groundtruth}
[End of Groundtruth Study Arm]

Supplementary Figure 4: Prompt used to calculate clinical relevance for study arms.

Input to Claude 3 for clinical relevance evaluation (Outcome Measures)

Act as an impartial judge and evaluate whether the outcome measures mentioned in a model's output are present in the full table of groundtruth outcome measures.
Output '1' or '0', where '1' means the outcome measures mentioned in the model's output are fully included in the groundtruth outcome measures table, and '0' means the outcome measures from the model's output are not included in the groundtruth.
You should provide an explanation for the evaluation.

Example:
[Model Output]
Absolutely. To measure the recruitment rate, we can track the number of participants who enroll in the study within a specified time frame. For this trial, we can monitor the recruitment rate up to 8 weeks after recruitment first opens. The goal is to achieve a recruitment rate of at least 70% to ensure the feasibility of conducting the full-scale trial.
[End of Model Output]

[Groundtruth Primary Outcome Measures]
Outcome Measure	Measure Description	Time Frame
Feasibility and safety	No adverse impacts of the study procedures on participants	Up to 3 weeks post-surgery
Recruitment	Recruitment rate of at least 70%	Up to 8 weeks after recruitment first opens
Randomization	Ability to randomize patients to one of two groups	Baseline
Data collection of stapler reload model	Ability to collect the type of stapler reloads used	Up to 3 weeks post-surgery
Data collection of stapler quantities	Ability to collect the number of stapler reloads used	Up to 3 weeks post-surgery
Data collection of energy sealing data	Ability to collect the sealing time in seconds	Up to 3 weeks post-surgery
Data collection of energy device data	Ability to collect the generator setting of the energy device	Up to 3 weeks post-surgery
[End of Groundtruth Primary Outcome Measures]

[Groundtruth Secondary Outcome Measures]
Outcome Measure	Measure Description	Time Frame
Adverse events (AEs) and complications	Short-term clinical outcomes, as measured by postoperative AEs and complications, will be recorded during patient follow-ups.	3 weeks post-surgery
Intraoperative costs of stapler or energy device use	Surgical device (stapler or energy) costs per surgery will be collected and evaluated in Canadian dollars.	Up to 3 weeks following hospital discharge
Hospitalization costs based on length of hospital stay	Inpatient hospitalization costs per day following surgery will be collected in Canadian dollars.	From admission to discharge, up to 14 days
[End of Groundtruth Secondary Outcome Measures]

Match prediction: 1

Now, evaluate the following model output and groundtruth outcome measures table.
You should first output the 'match prediction' at the beginning of the response by `Match prediction: `, e.g., `Match prediction: 1`.
Then, provide an explanation for your evaluation.

[Model Output]
{model_output}
[End of Model Output]

[Groundtruth Primary Outcome Measures]
{prim_out_meas}
[End of Groundtruth Primary Outcome Measures]

[Groundtruth Secondary Outcome Measures]
{sec_out_meas}
[End of Groundtruth Secondary Outcome Measures]

Supplementary Figure 5: Prompt used to calculate clinical relevance for outcome measures.

GPT 3.5 prompt for generating query generation data

Your task is to create some search expressions. Each search expression is a string that contains one or more of the following fields: Condition, InterventionType, LeadSponsorName, OverallStatus, Phase, StudyType, ResponsiblePartyInvestigatorFullName, NCTId, LocationCountry, StartDate, and CompletionDate. The search expression may contain multiple fields and each field may contain one or more values.

Example: AREA[Condition]"Diabetes" AND AREA[InterventionType]"Behavioral" AND AREA[OverallStatus]"NOT YET RECRUITING" AND AREA[StudyType]"INTERVENTIONAL".

List of 10 generated search expressions:

You are asked to generated a text-formatted query given a search expression used for searching clinical trials in a database.
The search expression is a string that contains one or more of the following fields: Condition, InterventionType, LeadSponsorName, OverallStatus, Phase, StudyType, ResponsiblePartyInvestigatorFullName, NCTId, LocationCountry, StartDate, and CompletionDate. The search expression may contain multiple fields and each field may contain multiple values.
The generated query should contain the information from the search expression in a human-readable format. The query can be converted back to the search expression. The generated query should imitate user's natural language and be as informative as possible. The generated query style should be diverse.

Search expression: {search_expression}
Generated query:

Supplementary Figure 6: Prompt used to construct query generation task data with GPT-3.5.

GPT 3.5 prompt for generating criteria design data

Given the information below about a clinical trial, please analyze and provide reasons for the design of each criterion listed under the "Criteria" section. For each criterion (both inclusion and exclusion criteria), explain why it is reasonable and necessary for the goals and structure of this trial.

Title: {brief_title}
Official Title: {official_title}
Conditions: {conditions}
Intervention / Treatment: {interventions}
Study Type: {study_type}
Phase: {phase}
Brief Summary: {brief_summary}
Criteria: {eligibility_criteria}

Given the information below about a clinical trial, please generate multi-turn conversation data used for training models. The generated conversation should revolve around criteria design, including the inclusion and exclusion criteria. Moreover, the generated conversation should contain interactions between users and chatbots: most of the time, chatbot gives advice on criteria design; when there is something needed to be clarified, users can provide some ideas to the chatbot. Also, somethimes when chatbot asks the user for ideas, the user may have no idea and then chatbot should give some suggestions. In such way, they complete the design of all of the criteria one by one and step by step.

Below is the information about the clinical trial:
Title: {brief_title}
Official Title: {official_title}
Conditions: {conditions}
Intervention / Treatment: {interventions}
Study Type: {study_type}
Phase: {phase}
Brief Summary: {brief_summary}
Criteria: {eligibility_criteria}
Reasons for the design of each criterion: {reasons generated above}

Now generate the conversation data for the design of the criteria. The information the user should implicitly provide includes the following: Title, Conditions, Intervention / Treatment, Study Type, Phase. In the final part of the conversation, the conversation should output the full criteria provided above. Note that all the information in output full criteria can be exactly found from the conversation. Note that you should fully leverage the reasons provided for the design of each criterion in some smart way to generate the conversation data. The role in the generated conversation should be "User" and "Chatbot".

Supplementary Figure 7: Prompt for generating criteria design conversation data.

GPT 3.5 prompt for generating study arm design data

Given the information below about a clinical trial, please analyze and provide reasons for the design of each study arms listed under the "Study Arms" section. For each study arm, focus on Participant Group/Arm, Intervention/Treatment, and so on. Explain why they are reasonable and necessary for the goals and structure of this trial.

Title: {brief_title}
Official Title: {official_title}
Conditions: {conditions}
Intervention / Treatment: {interventions}
Study Type: {study_type}
Phase: {phase}
Brief Summary: {brief_summary}
Criteria: {eligibility_criteria}
Study Arms: {arms_and_interventions}

Given the information below about a clinical trial, please generate multi-turn conversation data used for training models. The generated conversation should revolve around study arm design, including the participant group/arm and intervention/treatment. Moreover, the generated conversation should contain interactions between users and chatbots: most of the time, chatbot gives advice on study arm design; when there is something needed to be clarified, users can provide some ideas to the chatbot. Also, somethimes when chatbot asks the user for ideas, the user may have no idea and then chatbot should give some suggestions. In such way, they complete the design of all of the study arms one by one and step by step.

Below is the information about the clinical trial:
Title: {brief_title}
Official Title: {official_title}
Conditions: {conditions}
Intervention / Treatment: {interventions}
Study Type: {study_type}
Phase: {phase}
Brief Summary: {brief_summary}
Criteria: {eligibility_criteria}
Study Arms: {arms_and_interventions}
Reasons for the design of each study arm: {reasons generated above}

Now generate the conversation data for the design of the study arms. The information the user should provide at the beginning of the conversation includes the following: Title, Conditions, Intervention / Treatment, Study Type, Phase, Criteria, Design Details and so on.In the final part of the conversation, the conversation should output the full study arms provided above. Note that all the information in output full study arms can be exactly found from the conversation. Note that you should fully leverage the reasons provided for the design of each study arm in some smart way to generate the conversation data. Note that the user\'s aim is to design the study arms, and the chatbot should provide some advice and suggestions. The role in the generated conversation should be "User" and "Chatbot".

Supplementary Figure 8: Prompt for generating study arm design conversation data.

GPT 3.5 prompt for generating ourcome measure design data

Given the information below about a clinical trial, please analyze and provide reasons for the design of each outcome measure listed under the "Primary Outcome Measure" and "Second Outcome Measure" sections. For each outcome measure, focus on Outcome Measure, Measure Description, Time Frame, and so on. Explain why they are reasonable and necessary for the goals and structure of this trial.

Title: {brief_title}
Official Title: {official_title}
Conditions: {conditions}
Intervention / Treatment: {interventions}
Study Type: {study_type}
Phase: {phase}
Brief Summary: {brief_summary}
Criteria: {eligibility_criteria}
Study Arms: {arms_and_interventions}
Design Details: {design_details}
Primary Outcome Measure: {primary_outcome_measures}
Second Outcome Measure: {secondary_outcome_measures}

Given the information below about a clinical trial, please generate multi-turn conversation data used for training models. The generated conversation should revolve around outcome measure design, including the primary and secondary outcome measures. For each outcome measure, focus on Outcome Measure, Measure Description, Time Frame, and so on. Moreover, the generated conversation should contain interactions between users and chatbots: most of the time, chatbot gives advice on outcome measure design; when there is something needed to be clarified, users can provide some ideas to the chatbot. Also, somethimes when chatbot asks the user for ideas, the user may have no idea and then chatbot should give some suggestions. In such way, they complete the design of all of the outcome measures one by one and step by step.

Below is the information about the clinical trial:
Title: {brief_title}
Official Title: {official_title}
Conditions: {conditions}
Intervention / Treatment: {interventions}
Study Type: {study_type}
Phase: {phase}
Brief Summary: {brief_summary}
Criteria: {eligibility_criteria}
Design Details: {design_details}
Study Arms: {arms_and_interventions}
Primary Outcome Measure: {primary_outcome_measures}
Second Outcome Measure: {secondary_outcome_measures}
Reasons for the design of each outcome measure: {reasons generated above}

Now generate the conversation data for the design of the outcome measures. The information the user should implicitly provide includes the following: Title, Conditions, Intervention / Treatment, Study Type, Phase, Criteria, Design Details, Study Arms and so on. In the final part of the conversation, the conversation should output full outcome measures provided above, including primary outcome measures and secondary outcome measures. Note that all the information in output full outcome measures can be exactly found from the conversation. Note that you should fully leverage the reasons provided for the design of each outcome measure in some smart way to generate the conversation data. Note that the user\'s aim is to design the outcome measures, and the chatbot should provide some advice and suggestions.The role in the generated conversation should be "User" and "Chatbot".

Supplementary Figure 9: Prompt for generating outcome measure design conversation data.

Evaluation prompt for patient-trial matching

Hello. You are a helpful assistant for clinical trial recruitment. Your task is to compare a given patient note and the inclusion criteria of a clinical trial to determine the patient's eligibility. The factors that allow someone to participate in a clinical study are called inclusion criteria. They are based on characteristics such as age, gender, the type and stage of a disease, previous treatment history, and other medical conditions.

The assessment of eligibility has a three-point scale: 0) Excluded (patient meets inclusion criteria, but is excluded on the grounds of the trial's exclusion criteria); 1) Not relevant (patient does not have sufficient information to qualify for the trial); and 2) Eligible (patient meets inclusion criteria and exclusion criteria do not apply).
You should make a trial-level eligibility on each patient for the clinical trial, i.e., output the scale for the assessment of eligibility.

Here is an example patient note: Patient is a 45-year-old man with a history of anaplastic astrocytoma of the spine complicated by severe lower extremity weakness and urinary retention s/p Foley catheter, high-dose steroids, hypertension, and chronic pain. The tumor is located in the T-L spine...
Here is an example clinical trial:
Title: Is the Severity of Urinary Disorders Related to Falls in People With Multiple Sclerosis
Target diseases: Fall, Multiple Sclerosis, Lower Urinary Tract Symptoms
Interventions: Clinical tests
Summary: Falls are a common problem in people with multiple sclerosis (PwMS) and can lead to severe consequences (trauma, fear of falling, reduction of social activities). Prevention of falls is one of the priority targets of rehabilitation for PwMS and walking difficulties, which can result of different factors (motor impairment, ataxia, sensitive disorders, fatigability…). Urinary incontinence has been evoked as predictive of falls. But lower urinary tract symptoms (LUTSs) are frequent in PwMS, the prevalence of LUTSs is high (32-96.8%) and increases with MS duration and severity of ...
Inclusion criteria: inclusion criteria: age ≥ 18 years, Multiple sclerosis (MS) diagnosis, Lower urinary tract symptoms with or without treatment, Expanded Disability Status Scale score between 1 and 6.5
Example trial-level eligibility: 0) Would not refer this patient for this clinical trial.

Here is the patient note: *{patient_note}*

Here is the clinical trial: *{clinical_trial}*

Let's think step by step.
Finally, you should always repeat Trial-level eligibility in the last line by `Trial-level eligibility: `, e.g., `Trial-level eligibility: 2) Highly likely to refer this patient for this clinical trial.`.

Supplementary Figure 10: Prompt for evaluation on patient-trial matching.

Evaluation prompt for query generation

Given a query used for searching clinical trials in a database, conduct exact extracttion of related entities from the query and then generate a JSON object that can be used to query the database. If a field is not provided, leave it empty fiiled with 'N/A'.

Query: {query}

Output:

Evaluation prompt for query expansion

Given MeSH Terms used for searching clinical trials in a database, expand the input MeSH terms and then generate a JSON object that contains the expanded MeSH terms. Don't include the original MeSH terms in the expanded MeSH terms.

Input MeSH Terms: {mesh term list}

Expanded MeSH Terms:

Supplementary Figure 11: Prompt for evaluation on trial search.

Evaluation prompt for single-trial summarization

Your task is to create a clear, concise, and accurate summary of the provided clinical trial document. The summary should capture the key aspects of the trial.
The output should only be the summarization of the given trial. Do not explain how you summarize it.

Input Text: {text}
Summary:

Evaluation prompt for multi-trial summarization

Your task is to synthesize the key findings from a collection of study abstracts related to a specific clinical trial related research question.
Combine the insights from the provided abstracts into a cohesive summary. Your summary should integrate the findings rather than listing them separately. It's crucial to maintain the scientific integrity of the original studies while ensuring the summary is accessible and informative.
The output should only be the summary. Do not explain how you summarize it.

Study Abstracts: {titles and abstracts of multiple trial papers}
Summary:

Supplementary Figure 12: Prompt for evaluation on trial summarization.

Supplementary Table 1: Statistics of `TrialAlign`.

Source	# Total	# Train (< 2023)	# Test (≥ 2023)
ClinicalTrials.gov[45]	467,944	432,676	31,023
ChiCTR (China)[51]	76,186	65,181	11,005
EUCTR (EU)[52]	43,599	43,315	284
JRCT (Japan)[53]	64,650	60,645	4,005
ANZCTR (Australian New Zealand)[54]	24,657	23,374	1,283
ISRCTN.org[55]	24,174	22,966	1,208
ReBEC (Brazil)[56]	6,735	5,889	846
CRIS (Korea)[57]	8,953	8,428	525
DRKS (German)[58]	15,693	13,789	1,904
IRCT (Iran)[59]	37,782	34,097	3,685
TCTR (Thailand)[60]	8,649	7,443	1,206
LTR (Netherland)[61]	9,768	9,768	0
PACTR (Africa)[62]	4,047	3,848	199
SLCTR (Sri Lanka)[63]	442	421	21
Trial Papers (Embase[64] + PubMed[65])	1,113,207	1,113,207	-

References

[1] Ling, A. L. *et al.* Clinical trial links oncolytic immunoactivation to survival in glioblastoma. *Nature* **623**, 157–166 (2023).

[2] Heitmann, J. S. *et al.* A covid-19 peptide vaccine for the induction of sars-cov-2 t cell immunity. *Nature* **601**, 617–622 (2022).

[3] Hammond, T. C. *et al.* A phase 1/2 clinical trial of invariant natural killer t cell therapy in moderate-severe acute respiratory distress syndrome. *Nature Communications* **15**, 974 (2024).

[4] Giamarellos-Bourboulis, E. J. *et al.* Activate: randomized clinical trial of bcg vaccination against infection in the elderly. *Cell* **183**, 315–323 (2020).

[5] Gilbert, P. B. *et al.* Immune correlates analysis of the mrna-1273 covid-19 vaccine efficacy clinical trial. *Science* **375**, 43–50 (2022).

[6] Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[7] Wang, Z., Xiao, C. & Sun, J. Autotrial: Prompting language models for clinical trial design. In Bouamor, H., Pino, J. & Bali, K. (eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 12461–12472 (Association for Computational Linguistics, 2023).

[8] Gao, J., Xiao, C., Glass, L. M. & Sun, J. Compose: Cross-modal pseudo-siamese network for patient trial matching. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 803–812 (2020).

[9] Wang, Z. & Sun, J. Trial2vec: Zero-shot clinical trial document similarity search using self-supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 6377–6390 (2022).

[10] Gligorijevic, J. *et al.* Optimizing clinical trials recruitment via deep learning. *Journal of the American Medical Informatics Association* **26**, 1195–1202 (2019).

[11] Zhang, X., Xiao, C., Glass, L. M. & Sun, J. Deepenroll: patient-trial matching with deep embedding and entailment prediction. In *Proceedings of the web conference 2020*, 1029–1037 (2020).

[12] Kim, J. H. *et al.* Towards clinical data-driven eligibility criteria optimization for interventional covid-19 clinical trials. *Journal of the American Medical Informatics Association* **28**, 14–22 (2021).

[13] Tu, T. *et al.* Towards generalist biomedical AI. *CoRR* **abs/2307.14334** (2023).

[14] Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).

[15] Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863–874 (2024).

[16] Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **30**, 850–862 (2024).

[17] Cui, H. *et al.* scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods* 1–11 (2024).

[18] Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine* **29**, 2307–2316 (2023).

[19] Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* 1–8 (2024).

[20] Jin, Q. *et al.* Matching patients to clinical trials with large language models. *ArXiv* (2023).

[21] Yuan, J., Tang, R., Jiang, X. & Hu, X. Large language models for healthcare data augmentation: An example on patient-trial matching. *arXiv preprint arXiv:2303.16756* (2023).

[22] Wong, C. *et al.* Scaling clinical trial matching using large language models: A case study in oncology. *CoRR* **abs/2308.02180** (2023).

[23] Li, C. *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* **36** (2024).

[24] Chaves, J. M. Z. *et al.* Training small multimodal models to bridge biomedical competency gap: A case study in radiology imaging. *arXiv preprint arXiv:2403.08002* (2024).

[25] DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B. & Wang, L. L. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486* (2021).

[26] Jiang, P. *et al.* Trisum: Learning summarization ability from large language models with structured rationale. *arXiv preprint arXiv:2403.10351* (2024).

[27] Jiang, A. Q. *et al.* Mistral 7b. *arXiv preprint arXiv:2310.06825* (2023).

[28] Labrak, Y. *et al.* Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373* (2024).

[29] Anthropic, A. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card* (2024).

[30] Roberts, K., Demner-Fushman, D., Voorhees, E. M., Bedrick, S. & Hersh, W. R. Overview of the trec 2021 clinical trials track. In *Proceedings of the thirtieth text retrieval conference (TREC 2021)* (2021).

[31] Koopman, B. & Zuccon, G. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 669–672 (2016).

[32] Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[33] Luo, R. *et al.* Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinform.* **23** (2022).

[34] Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

[35] Chen, Z. *et al.* Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079* (2023).

[36] Van Veen, D. *et al.* Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine* 1–9 (2024).

[37] Tayebi Arasteh, S. *et al.* Large language models streamline automated machine learning for clinical studies. *Nature Communications* **15**, 1603 (2024).

[38] Nori, H. *et al.* Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *CoRR* **abs/2311.16452** (2023).

[39] Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35**, 27730–27744 (2022).

[40] Lin, Z., Trivedi, S. & Sun, J. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187* (2023).

[41] Semnani, S., Yao, V., Zhang, H. & Lam, M. WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2387–2413 (2023).

[42] Hu, E. J. *et al.* LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations* (2021).

[43] Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* **33**, 9459–9474 (2020).

[44] Collaboration, C. *et al.* Cochrane central register of controlled trials (central) (2014).

[45] Bergeris, A., Ide, N. C. & Tse, T. Clinicaltrials. gov (2005).

[46] Wallace, B. C., Saha, S., Soboczenski, F. & Marshall, I. J. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings* **2021**, 605 (2021).

[47] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019* (OpenReview.net, 2019). URL https://openreview.net/forum?id=Bkg6RiCqY7.

[48] Rajbhandari, S., Rasley, J., Ruwase, O. & He, Y. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16 (IEEE, 2020).

[49] Dao, T. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691* (2023).

[50] 1rgs. Jsonformer: A bulletproof way to generate structured json from language models (2023).

[51] Wu, T. *et al.* Chinese clinical trial registry: mission, responsibility and operation. *Journal of evidence-based medicine* **4**, 165–167 (2011).

[52] Egger, G. F. *et al.* European union clinical trials register: on the way to more transparency of clinical trial data. *Expert Review of Clinical Pharmacology* **6**, 457–459 (2013).

[53] Shiokawa, T. Background, introduction and activity of the japan primary registries network. *Journal of Evidence-Based Medicine* **2**, 41–43 (2009).

[54] Askie, L. M. Australian new zealand clinical trials registry: history and growth. *Journal of Evidence-Based Medicine* **4**, 185–187 (2011).

[55] Faure, H. & Hrynaszkiewicz, I. The isrctn register: achievements and challenges 8 years on. *Journal of evidence-based medicine* **4**, 188–192 (2011).

[56] Laguardia, J. *et al.* Brazilian clinical trials registry and the challenges for clinical research governance. *Journal of Evidence-Based Medicine* **4**, 156–160 (2011).

[57] Park, H.-Y. Primary registry of the who international clinical trial registry platform: Clinical research information service (cris). *Journal of the Korean Medical Association* **54**, 92–97 (2011).

[58] Hasselblatt, H., Dreier, G., Antes, G. & Schumacher, M. The german clinical trials register: challenges and chances of implementing a bilingual registry. *Journal of Evidence-Based Medicine* **2**, 36–40 (2009).

[59] Solaymani-Dodaran, M., Ostovar, A., Khalili, D. & Vasei, M. Iranian registry of clinical trials: path and challenges from conception to a world health organization primary register. *Journal of Evidence-Based Medicine* **2**, 32–35 (2009).

[60] Tulvatana, W., Kulvichit, K., Thinkhamrop, B. & Tatsanavivat, P. Thai clinical trials registry. *Journal of Evidence-Based Medicine* **4**, 182–184 (2011).

[61] Driessen, M. *et al.* The dutch nationwide trauma registry: the value of capturing all acute trauma admissions. *Injury* **51**, 2553–2559 (2020).

[62] Abrams, A. & Siegfried, N. The pan african clinical trials registry: year one data analysis of the only african member of the world health organization network of primary registries. *Journal of Evidence-Based Medicine* **3**, 195–200 (2010).

[63] Ranawaka, U. K. & Goonaratna, C. The sri lanka clinical trials registry–moving forward. *Journal of Evidence-Based Medicine* **4**, 179–181 (2011).

[64] Elsevier Science. Embase [electronic database]. Electronic Database (1974). Produced by Elsevier Science, Amsterdam, The Netherlands.

[65] Canese, K. & Weis, S. Pubmed: the bibliographic database. *The NCBI handbook* **2** (2013).