

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

Jiacheng Lin¹, Hanwen Xu², Zifeng Wang¹, 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.

24 Introduction

25 Clinical trials are research studies conducted on humans to evaluate the safety and efficacy
26 of new medical treatments, interventions, or devices before they are approved for widespread
27 use. They form the foundation of modern medicine.¹⁻⁵ The challenges in clinical trials are
28 three-fold. First, a clinical trial involves several interconnected design components, including
29 trial descriptions, eligibility criteria, study arms, outcome metrics, and more, that need to be
30 collectively designed to ensure optimal patient recruitment and outcome assessment. Second,
31 clinical trial data are usually highly sensitive and private, hence often not amenable to public
32 cloud-based tools (e.g., GPT-4⁶) for processing and analysis. Third, clinical trial develop-
33 ment requires multiple tasks, such as eligibility criteria design and patient recruitment, which
34 require substantial domain expertise.

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

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

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

64 We have made all our training datasets (**TrialAlign** and **TrialInstruct**) and the evalu-
65 ation benchmark (**TrialPanorama**) available for future research and benchmarking of clinical
66 trial foundation models. Additionally, we have open-sourced the code and model weights
67 of **Panacea**. **Panacea** can run on a single-GPU machine, making it easy to use within an
68 organization. Fine-tuning **Panacea** on 200 thousand documents only takes seven hours using
69 a standard cluster with 4 A-100 GPUs. This advantage allows for further customization of
70 **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	750	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.	45,516	8,032	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

71 Results

72 Overview of Panacea

73 Our goal is to develop **Panacea**, a domain-specific foundation model for clinical trial tasks.
74 Like previous works on developing domain-specific foundation models,^{23,24} the biggest chal-
75 lenge for developing **Panacea** is to curate the high-quality fine-tuning data to align **Panacea**
76 to clinical trial vocabulary and create the specific instruction data for clinical trial tasks.
77 **Panacea** consists of two main steps: an alignment step, which adapts **Panacea** to the vocab-
78 ulary used in clinical trials, and an instruction-tuning step, which instructs **Panacea** on each
79 clinical trial task. We built two datasets **TrialAlign** and **TrialInstruct** for the alignment
80 step and the instruction-tuning step, respectively.

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

90 **TrialInstruct** contains instruction-tuning data from eight diverse tasks, including crite-
91 ria design, study arm design, outcome measure design, patient-trial matching, query genera-
92 tion, query expansion, single-trial summarization, and multi-trial summarization, instructing
93 **Panacea** on solving these tasks (**Fig. 1c**). Each task contains at least 2,000 data points,
94 where each data point contains an instruction, an input, and an output (**Fig. 1d**). Since
95 these eight tasks are related, we jointly fine-tuned the model using instruction data from these
96 eight tasks, transforming **Panacea** into an all-in-one tool for clinical trial applications (**Fig.**
97 **1e**).

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

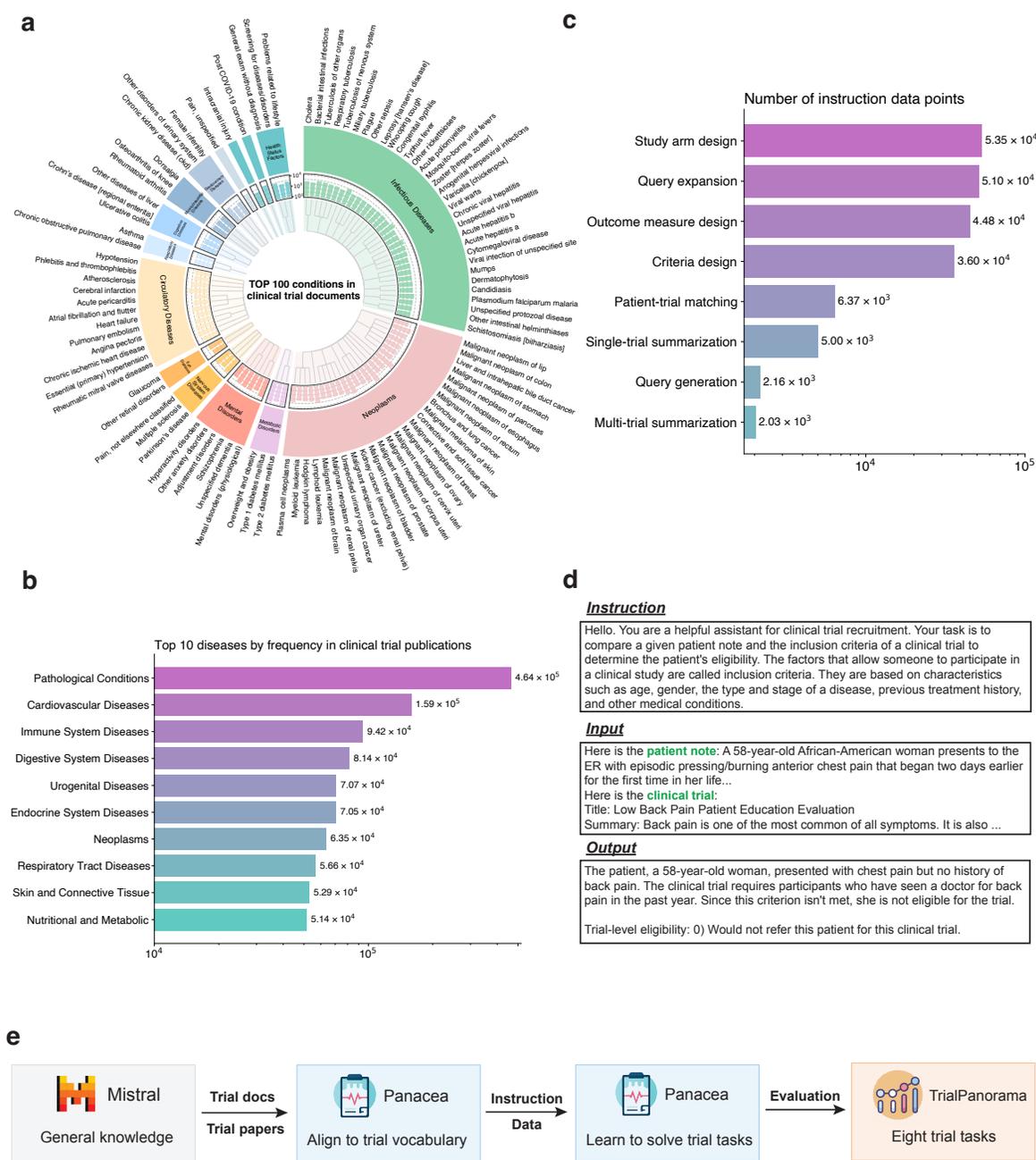


Figure 1: **Overview of Panacea.** **a**, Number of de-identified trial documents in each ICD-10 category. The top 100 conditions with the most number of trial documents are illustrated here. **b**, Bar plot showing the most frequent diseases in clinical trial publications according to the MeSH terms. **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.

102 Accurate trial search through query generation and expansion

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

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

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

127 A novel metric to evaluate trial summarization

128 Once similar trials are identified, the next task is to understand those trials via summarization.
129 We evaluated the performance of **Panacea** on trial summarization. We studied both single-
130 trial summarization, which aims to provide a concise summary of a specific trial study (**Fig.**
131 **3a**), and multi-trial summarization, which aims to summarize multiple trial studies that study
132 similar conditions and interventions (**Fig. 3b**).

133 Since it could be biased to evaluate summarization using lexical-based metrics, we propose
134 a novel metric based on large language models (see **Methods, Supplementary Figures 1**
135 **and 2**). In particular, we provided the ground truth summarization and the model-generated
136 summarization to Claude and asked if these summarizations studied the same problem and
137 made the same conclusion. We found that **Panacea** and comparison approaches can correctly
138 summarize the trial goal, while the summarization of the trial conclusion is less accurate (**Fig.**
139 **3c-d**). Moreover, summarizing multiple trials is more challenging than summarizing a single
140 trial based on the proposed metric. Nevertheless, our method still outperformed comparison
141 approaches in summarizing multiple trials, suggesting its potential to assist researchers in
142 extracting key information from many related trial studies.

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

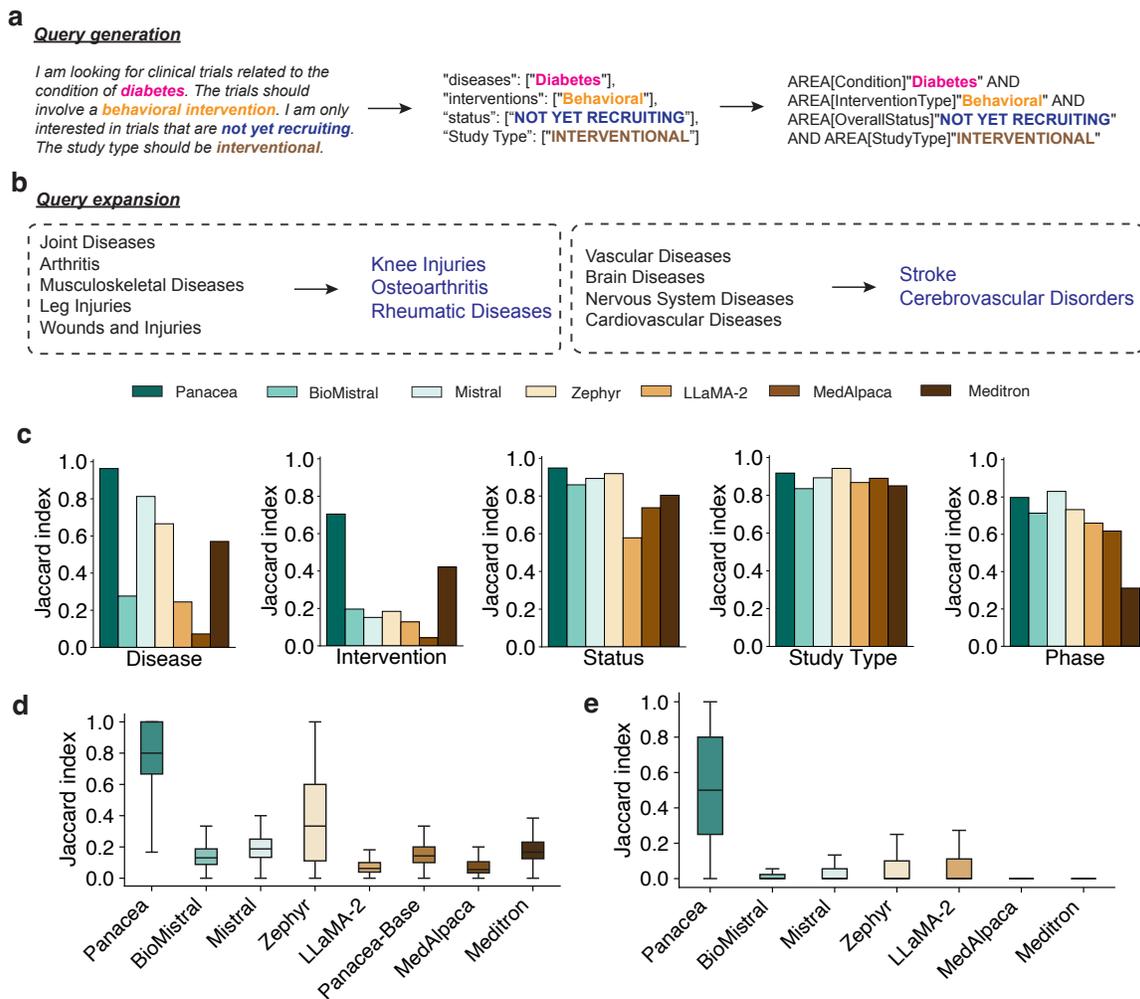


Figure 2: **Evaluation on trial search.** **a**, Query generation aims to convert free text user input into a structured query that contains five categories: disease, intervention, phase, status, and study type. **b**, Query expansion aims to expand a set of keywords. Candidate keywords are not provided.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

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

153 **Improved performance on clinical trial design**

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

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

176 Lexical similarity metrics are widely used to evaluate text generation problems, but might
177 not be clinically specific enough to evaluate the generations by **Panacea**. Recently, LLMs
178 have been used to evaluate the generated text by exploiting their strong ability in text un-
179 derstanding. Here, we exploit Claude²⁹ to evaluate these three tasks by asking the model
180 whether the generated task is clinically relevant (see **Methods, Supplementary Figures**
181 **3-5**). We found that **Panacea** outperforms all methods on criteria and study arms design,
182 demonstrating the high quality of generation by **Panacea** (**Fig. 4b**).

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

193 **Accurate patient-trial matching**

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

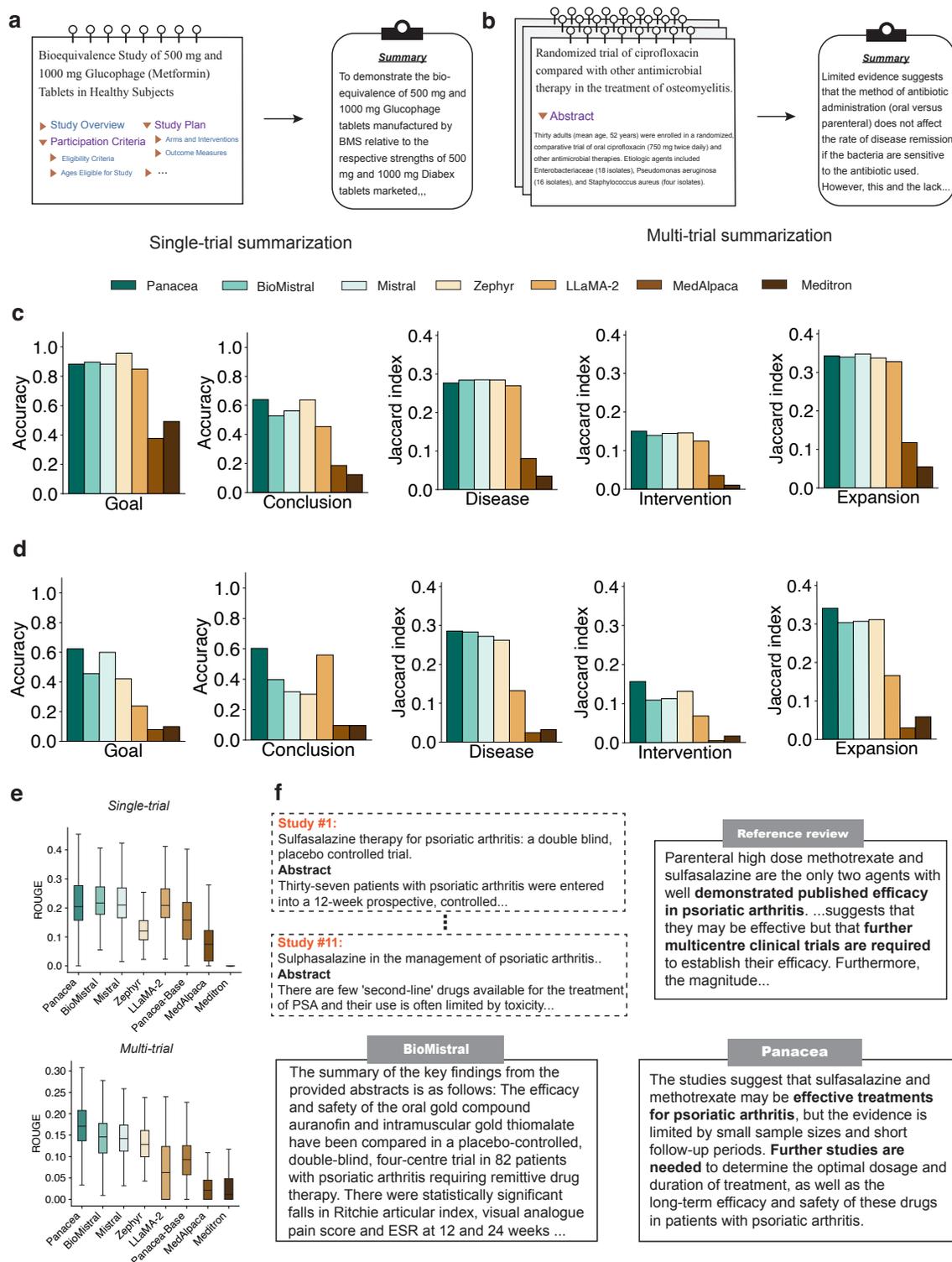


Figure 3: **Evaluating Panacea on trial summarization.** **a,b**, Trial summarization aims to provide a concise summary, including trial goal and conclusion, for a single trial (**a**) or multiple trials (**b**). **c,d**, Evaluation on single-trial summarization (**c**) and multiple-trial summarization (**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.

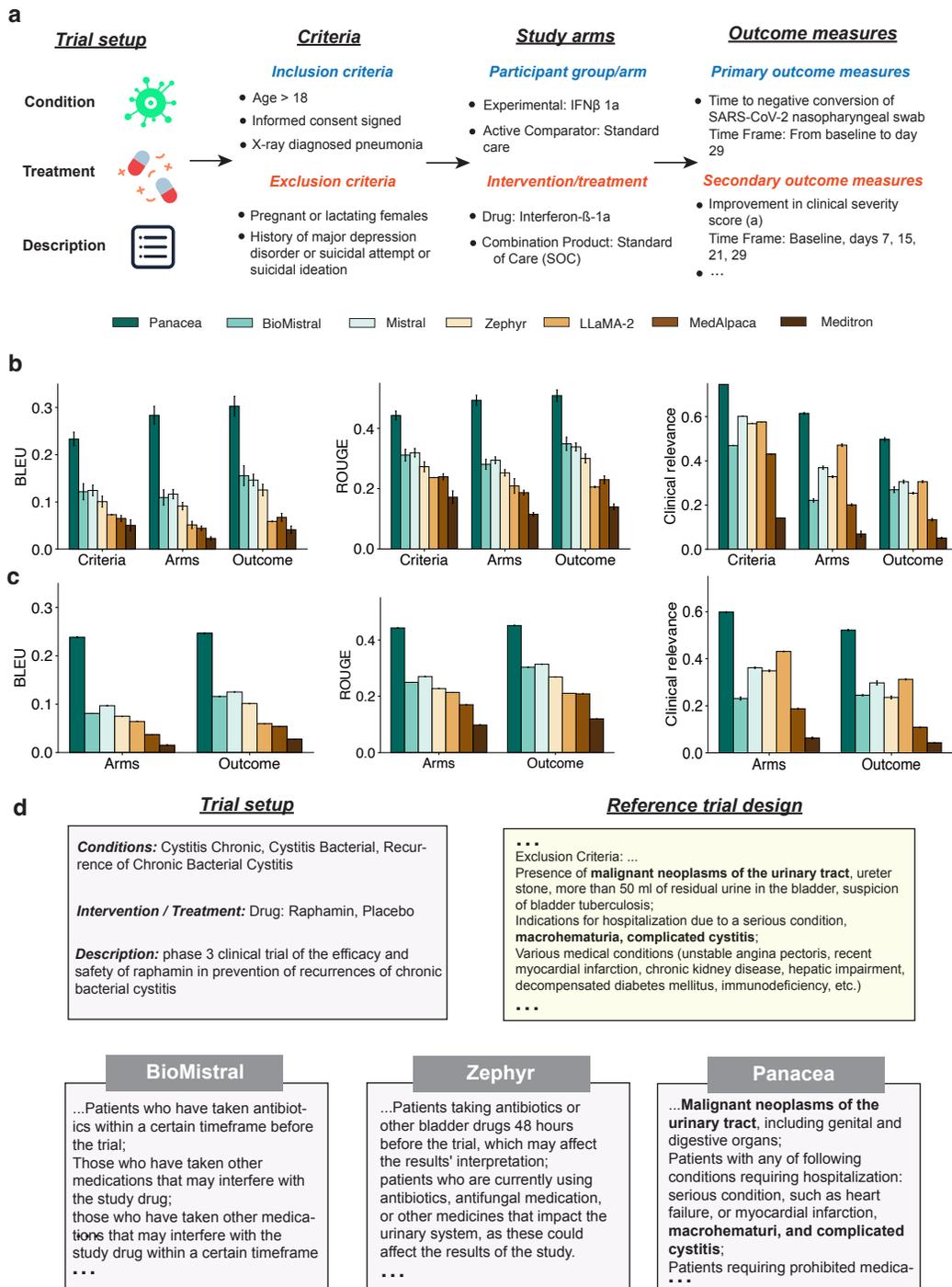


Figure 4: **Evaluation on clinical trial design.** **a**, Problem setting of clinical trial design, which aims to generate criteria, study arms, and outcome measures. Criteria are used as input to generate study arms. Criteria and study arms are used to generate the outcome measures. **b**, Evaluation 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.

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

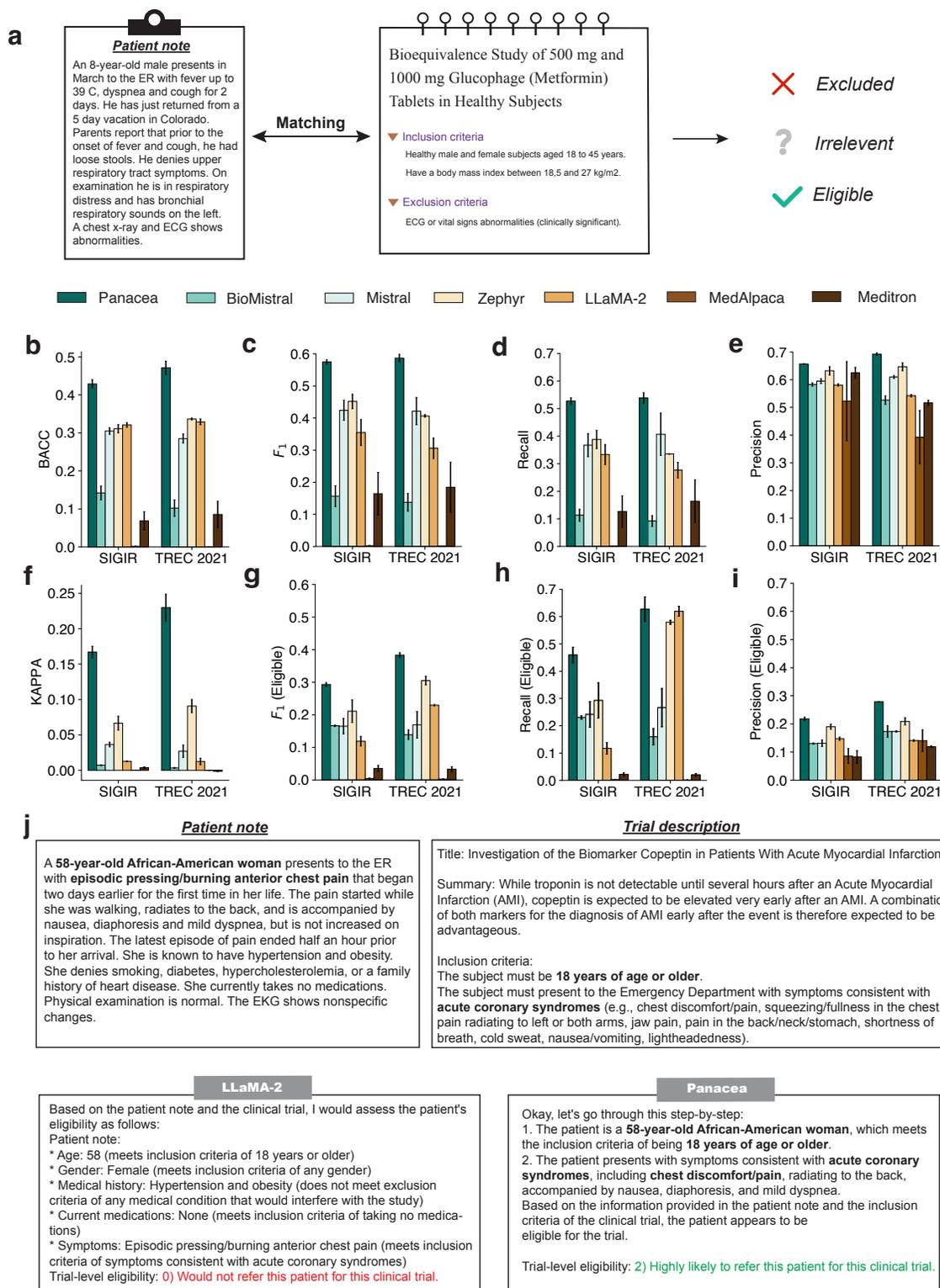
198 We first evaluated our method on the TREC 2021 dataset,³⁰ which consists of a training
199 set and a test set. We used the training set to construct instructions in `TrialAlign`, and then
200 assessed the performance of `Panacea` on the test set. We found that `Panacea` outperformed
201 all comparison approaches in terms of balanced accuracy (BACC), Cohen’s KAPPA score,
202 Recall, Precision, and F1, indicating the effectiveness of using `TrialInstruct` to fine-tune
203 the model (**Fig. 5b-f**). To investigate the generalizability of our method, we further tested
204 our method on the SIGIR dataset³¹ where the entire dataset is used as the test set. We found
205 that our method again attained the best performance on all three metrics, demonstrating the
206 strong generalizability of our method.

207 As the eligible class is crucial for patient-trial recruitment, we further examined a binary
208 classification setting. In this setting, we grouped “excluded” and “irrelevant” into one cat-
209 egory, and “eligible” into the other in order to determine whether a patient is eligible for a
210 trial. Our method outperformed all comparison approaches in terms of F1, precision, and
211 recall, indicating its applicability to real-world trial recruitment (**Fig. 5g-i**). Finally, we
212 used a case study to illustrate how our method successfully classified a patient as eligible by
213 examining each criterion and coming to a conclusion based on their criteria (**Fig. 5j**). In
214 contrast, LLaMA-2³² made an incorrect conclusion by hallucinating an exclusion criterion not
215 stated in the trial description.

216 Discussion

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

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



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

245 This study has several limitations that we would like to address in the future. First,
246 despite being fine-tuned on clinical trial instruction datasets, LLMs may still produce bi-
247 ased or low-quality outputs. Enhancing model alignment such as reinforcement learning from
248 human feedback³⁹ is crucial future work before **Panacea** can be deployed in production set-
249 tings. Second, for high-stakes applications such as clinical trials, it is essential to detect and
250 regulate LLM hallucinations, which can occur, particularly in areas not well-covered by the
251 LLM training data. It is worth exploring to enable LLMs to either reject an answer⁴⁰ or
252 utilize external knowledge bases to correct its outputs.⁴¹ Third, continually updating the
253 model's knowledge is vital for maintaining relevance and accuracy in a rapidly evolving med-
254 ical landscape. Therefore, it is worth exploring efficient knowledge updating techniques for
255 **Panacea**⁴² or enhancing it with retrieval-augmented generation.⁴³ Fourth, although **Panacea**
256 demonstrates significant improvements across various benchmark datasets, there is a need to
257 develop more evaluation metrics to comprehensively assess LLM performance in more clinical
258 trial tasks. Additionally, conducting user studies could further demonstrate the benefits of
259 **Panacea** in assisting experts with clinical development projects.

260 Method

261 Creating TrialAlign dataset

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

276 **Filtering** For trial documents, we further conduct intra- and inter-source de-duplication and
277 then remove the personally identifiable information (PII), finally obtaining 793k trial docu-
278 ment data. Further, to avoid information leakage, we selected documents with registration
279 dates before 2023-01-01 as the training corpus. The remaining is used for test data curation.
280 For trial papers, we de-duplicated all the papers and the final 1.11M trial paper corpus con-
281 sists of abstracts of all the papers and full text of 97k papers from PubMed Central (PMC).
282 Similarly, to avoid information leakage, we choose papers published before 2023-01-01, which
283 ensures the dates of related clinical trials of the selected papers are definitely before 2023-01-
284 01.

285 **Document/paper structure organization** For trial documents, we follow the format
286 shown in clinicaltrials.gov⁴⁵ to organize all the corpus for alignment. Each trial document
287 is arranged into a markdown format passage. For trial documents from clinicaltrials.gov,
288 each document contains section (1) “Public Title”; (2) “Study Overview” covering subsec-
289 tions “Brief Summary”, “Detailed Description”, “Official Title”, “Conditions” and “Inter-
290 vention/Treatment”; (3) Participation Criteria, including subsections “Eligibility Criteria”,
291 “Ages Eligibility for Study”, “Sexes Eligibility for Study” and “Accepts Healthy Volunteers”;
292 (4) “Study Plan”, including subsection “How is the study designed?” that contains “Design
293 Details” and “Arms and Interventions”, subsection “What is the study measuring?” con-
294 taining primary and secondary outcome measures; (5) Terms related to the study. For trial
295 documents from other sources, each document contains “Public Title”, “Scientific Title”,
296 “Study Type”, “Study Design”, “Intervention”, “Inclusion Criteria”, “Exclusion Criteria”,
297 “Primary Outcome Measures” and “Secondary Outcome Measures”. For trial paper data,
298 each paper contains “Title”, “Abstract” and full text (if any).

299 Creating TrialInstruct dataset

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

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

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

317 **Trial summarization** Trial summarization contains single-trial and multi-trial summariza-
318 tion. To curate single-trial summarization data, we leverage clinicaltrials.gov documents.
319 Specifically, the brief summary section serves as the output and the other parts serve as
320 the input. We finally have 5k training data (before 2023-01-01) and 1k test data (after
321 2023-01-01). For the multi-trial summarization data curation, we derived our dataset from
322 Cochrane dataset of systematic reviews,⁴⁶ i.e., we only selected data pairs containing clinical
323 trial papers. Specifically, each multi-trial summarization data contains a PMID set and a
324 review paper. The review is a high-level conclusion from papers in the PMID set. The data
325 curation process started with the matching between the PMID sets and all the trial paper
326 PMIDs in `TrialAlign`. We select those data pairs with at least three trial-related papers in
327 the PMID set. We finally constructed 2,029 samples for training and 252 for test, derived
328 from the Cochrane dataset’s training and validation sets due to the missing test labels in the
329 original Cochrane dataset.

330 **Trial design** We construct multi-turn conversation data for trial design due to the difficulty
331 of one-turn design, even for frontier models like GPT-4.⁶ Such conversation format data are
332 more realistic and benefit users to get more accurate designs as conversations progress. To
333 construct these conversation data, we focus on trial documents in clinicaltrials.gov and adopt
334 a two-stage strategy to construct the conversation data. For criteria design, we first input
335 criteria and trial setup, which contains title, conditions, drugs, and phase, to ask GPT-3.5 to
336 output the reasons for designing those criteria one by one. In the second stage, we input the
337 criteria, and reasons generated in the first stage, and trial setup, to ask GPT-3.5 to construct
338 multi-turn conversation data (see **Supplementary Figure 7**). This can ensure that GPT-
339 3.5 generated trial part data is actual. Likewise, for study arm design, we input study arms,
340 criteria, and trial setup. In the second stage, we collect the generated conversation data
341 given the study arms, reasons, criteria, and trial setup (see **Supplementary Figure 8**). For
342 outcome measures, the input in the first stage is outcome measures, study arms, criteria, and
343 trial setup, while the input in the second stage is outcome measures, reasons, study arms,
344 criteria, and trial setup (see **Supplementary Figure 9**). We use trial documents from
345 clinicaltrials.gov to construct these data, before 2023-01-01 for training and after 2023-01-01
346 for testing. We finally obtained 35,951 and 549 for the criteria design’s training and test set,
347 53,548 and 549 for the study arm design, and 44,809 and 549 for the outcome measure design.

348 **Patient-trial matching** We converted existing representative patient-trial matching datasets
349 into instruction format, i.e., SIGIR³¹ and TREC 2021³⁰ cohorts. Each instruction data of
350 patient-trial matching follows the structure: “Instruction”, “One-shot demonstration”, “Input
351 patient notes”, “Input Criteria” and “Output trial-level eligibility”, as illustrated in **Supple-**
352 **mentary Figure 10**. We split the TREC 2021 into the training (28,406 samples) and test
353 sets (7,424 samples), and all SIGIR data serves as the test set (3,869 samples). Specifically,
354 the patient-criteria pairs of 80% of patients in TREC 2021 formed into the training set, while
355 those pairs of the remaining 20% of patients in TREC 2021 are test data. For evaluation, we
356 trained our `Panacea` on the training set derived from TREC 2021 and evaluated on the test

357 set of TREC 2021 and all data in SIGIR.

358 **Creating TrialPanorama benchmark**

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

363 **Details of Panacea model**

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

366 **Alignment** We built on the Mistral-7B-Base model²⁷ in this study. After parameter initial-
367 ization, **Panacea** was trained on the 1.8M **TrialAlign** data. We trained the model using the
368 AdamW optimizer⁴⁷ with a batch size 512 for one epoch. We adopted a cosine learning rate
369 scheduler with a peak learning rate 2×10^{-6} and 10% warm-up steps. We set max sequence
370 length as 8192 tokens. To improve training speed and optimize the memory, we adopted
371 DeepSpeed ZeRO-3⁴⁸ and FlashAttention-2⁴⁹ strategies. After the alignment process, we ob-
372 tain the **Panacea-Base** model. During the alignment step, **Panacea** was trained on 4 Nvidia
373 A100 80G for four days.

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

380 **Details of experiments on trial search**

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

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

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

395 For the query expansion task, this process enhances the original query by adding seman-
396 tically related terms, thereby broadening the search scope to include relevant trials that may
397 not use the exact phrasing of the original query terms. **Panacea** was trained to suggest addi-
398 tional keywords based on the initial input terms. The model learned to recognize and predict
399 related terms that could be associated with the initial query, expanding the search breadth
400 effectively.

401 Details of experiments on trial summarization

402 The experiments on trial summarization were designed to test Panacea’s capabilities in con-
403 densing complex clinical trial information into succinct summaries. This component of our
404 research focused on two specific tasks: single-trial summarization and multi-trial summariza-
405 tion (see **Supplementary Figure 12**).

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

414 Details of experiments on clinical trial design

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

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

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

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

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

443 Details of experiments on patient-trial matching

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

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

468 Code and data availability

469 The TrialAlign data for the alignment step, the TrialInstruct data for the instruction
470 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
473 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 generate 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, sometimes 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, sometimes 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 outcome 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, sometimes 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 \geq 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 extraction 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 filled 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 (\geq 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 (Netherlands) 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	-

474 References

- 475 [1] Ling, A. L. *et al.* Clinical trial links oncolytic immunoactivation to survival in glioblas-
476 toma. *Nature* **623**, 157–166 (2023).
- 477 [2] Heitmann, J. S. *et al.* A covid-19 peptide vaccine for the induction of sars-cov-2 t cell
478 immunity. *Nature* **601**, 617–622 (2022).
- 479 [3] Hammond, T. C. *et al.* A phase 1/2 clinical trial of invariant natural killer t cell therapy
480 in moderate-severe acute respiratory distress syndrome. *Nature Communications* **15**, 974
481 (2024).
- 482 [4] Giamarellos-Bourboulis, E. J. *et al.* Activate: randomized clinical trial of bcg vaccination
483 against infection in the elderly. *Cell* **183**, 315–323 (2020).
- 484 [5] Gilbert, P. B. *et al.* Immune correlates analysis of the mrna-1273 covid-19 vaccine efficacy
485 clinical trial. *Science* **375**, 43–50 (2022).
- 486 [6] Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- 487 [7] Wang, Z., Xiao, C. & Sun, J. Autotrial: Prompting language models for clinical trial
488 design. In Bouamor, H., Pino, J. & Bali, K. (eds.) *Proceedings of the 2023 Conference on*
489 *Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December*
490 *6-10, 2023*, 12461–12472 (Association for Computational Linguistics, 2023).
- 491 [8] Gao, J., Xiao, C., Glass, L. M. & Sun, J. Compose: Cross-modal pseudo-siamese net-
492 work for patient trial matching. In *Proceedings of the 26th ACM SIGKDD international*
493 *conference on knowledge discovery & data mining*, 803–812 (2020).
- 494 [9] Wang, Z. & Sun, J. Trial2vec: Zero-shot clinical trial document similarity search using
495 self-supervision. In *Findings of the Association for Computational Linguistics: EMNLP*
496 *2022*, 6377–6390 (2022).
- 497 [10] Gligorijevic, J. *et al.* Optimizing clinical trials recruitment via deep learning. *Journal of*
498 *the American Medical Informatics Association* **26**, 1195–1202 (2019).
- 499 [11] Zhang, X., Xiao, C., Glass, L. M. & Sun, J. Deepenroll: patient-trial matching with
500 deep embedding and entailment prediction. In *Proceedings of the web conference 2020*,
501 1029–1037 (2020).
- 502 [12] Kim, J. H. *et al.* Towards clinical data-driven eligibility criteria optimization for interven-
503 tional covid-19 clinical trials. *Journal of the American Medical Informatics Association*
504 **28**, 14–22 (2021).
- 505 [13] Tu, T. *et al.* Towards generalist biomedical AI. *CoRR* **abs/2307.14334** (2023).
- 506 [14] Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature*
507 **616**, 259–265 (2023).
- 508 [15] Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nature*
509 *Medicine* **30**, 863–874 (2024).

- 510 [16] Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **30**, 850–862 (2024).
511
- 512 [17] Cui, H. *et al.* scgpt: toward building a foundation model for single-cell multi-omics using
513 generative ai. *Nature Methods* 1–11 (2024).
- 514 [18] Huang, Z., Bianchi, F., Yuksekogonul, M., Montine, T. J. & Zou, J. A visual–language
515 foundation model for pathology image analysis using medical twitter. *Nature medicine*
516 **29**, 2307–2316 (2023).
- 517 [19] Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data.
518 *Nature* 1–8 (2024).
- 519 [20] Jin, Q. *et al.* Matching patients to clinical trials with large language models. *ArXiv*
520 (2023).
- 521 [21] Yuan, J., Tang, R., Jiang, X. & Hu, X. Large language models for healthcare data
522 augmentation: An example on patient-trial matching. *arXiv preprint arXiv:2303.16756*
523 (2023).
- 524 [22] Wong, C. *et al.* Scaling clinical trial matching using large language models: A case study
525 in oncology. *CoRR* **abs/2308.02180** (2023).
- 526 [23] Li, C. *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine
527 in one day. *Advances in Neural Information Processing Systems* **36** (2024).
- 528 [24] Chaves, J. M. Z. *et al.* Training small multimodal models to bridge biomedical compe-
529 tency gap: A case study in radiology imaging. *arXiv preprint arXiv:2403.08002* (2024).
- 530 [25] DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B. & Wang, L. L. Ms2: Multi-document
531 summarization of medical studies. *arXiv preprint arXiv:2104.06486* (2021).
- 532 [26] Jiang, P. *et al.* Trisum: Learning summarization ability from large language models with
533 structured rationale. *arXiv preprint arXiv:2403.10351* (2024).
- 534 [27] Jiang, A. Q. *et al.* Mistral 7b. *arXiv preprint arXiv:2310.06825* (2023).
- 535 [28] Labrak, Y. *et al.* Biomistral: A collection of open-source pretrained large language models
536 for medical domains. *arXiv preprint arXiv:2402.10373* (2024).
- 537 [29] Anthropic, A. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*
538 (2024).
- 539 [30] Roberts, K., Demner-Fushman, D., Voorhees, E. M., Bedrick, S. & Hersh, W. R.
540 Overview of the trec 2021 clinical trials track. In *Proceedings of the thirtieth text re-*
541 *trieval conference (TREC 2021)* (2021).
- 542 [31] Koopman, B. & Zuccon, G. A test collection for matching patients to clinical trials. In
543 *Proceedings of the 39th International ACM SIGIR conference on Research and Develop-*
544 *ment in Information Retrieval*, 669–672 (2016).
- 545 [32] Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*
546 *arXiv:2307.09288* (2023).
- 547 [33] Luo, R. *et al.* Biogpt: generative pre-trained transformer for biomedical text generation
548 and mining. *Briefings Bioinform.* **23** (2022).

- 549 [34] Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180
550 (2023).
- 551 [35] Chen, Z. *et al.* Meditron-70b: Scaling medical pretraining for large language models.
552 *arXiv preprint arXiv:2311.16079* (2023).
- 553 [36] Van Veen, D. *et al.* Adapted large language models can outperform medical experts in
554 clinical text summarization. *Nature Medicine* 1–9 (2024).
- 555 [37] Tayebi Arasteh, S. *et al.* Large language models streamline automated machine learning
556 for clinical studies. *Nature Communications* **15**, 1603 (2024).
- 557 [38] Nori, H. *et al.* Can generalist foundation models outcompete special-purpose tuning?
558 case study in medicine. *CoRR* **abs/2311.16452** (2023).
- 559 [39] Ouyang, L. *et al.* Training language models to follow instructions with human feedback.
560 *Advances in Neural Information Processing Systems* **35**, 27730–27744 (2022).
- 561 [40] Lin, Z., Trivedi, S. & Sun, J. Generating with confidence: Uncertainty quantification for
562 black-box large language models. *arXiv preprint arXiv:2305.19187* (2023).
- 563 [41] Semnani, S., Yao, V., Zhang, H. & Lam, M. WikiChat: Stopping the hallucination of
564 large language model chatbots by few-shot grounding on wikipedia. In *Findings of the*
565 *Association for Computational Linguistics: EMNLP 2023*, 2387–2413 (2023).
- 566 [42] Hu, E. J. *et al.* LoRA: Low-rank adaptation of large language models. In *International*
567 *Conference on Learning Representations* (2021).
- 568 [43] Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Ad-*
569 *vances in Neural Information Processing Systems* **33**, 9459–9474 (2020).
- 570 [44] Collaboration, C. *et al.* Cochrane central register of controlled trials (central) (2014).
- 571 [45] Bergeris, A., Ide, N. C. & Tse, T. Clinicaltrials.gov (2005).
- 572 [46] Wallace, B. C., Saha, S., Soboczinski, F. & Marshall, I. J. Generating (factual?) narra-
573 tive summaries of rcts: Experiments with neural multi-document summarization. *AMIA*
574 *Summits on Translational Science Proceedings* **2021**, 605 (2021).
- 575 [47] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *7th International*
576 *Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9,*
577 *2019* (OpenReview.net, 2019). URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- 578 [48] Rajbhandari, S., Rasley, J., Ruwase, O. & He, Y. Zero: Memory optimizations toward
579 training trillion parameter models. In *SC20: International Conference for High Perfor-*
580 *mance Computing, Networking, Storage and Analysis*, 1–16 (IEEE, 2020).
- 581 [49] Dao, T. Flashattention-2: Faster attention with better parallelism and work partitioning.
582 *arXiv preprint arXiv:2307.08691* (2023).
- 583 [50] Irgs. Jsonformer: A bulletproof way to generate structured json from language models
584 (2023).
- 585 [51] Wu, T. *et al.* Chinese clinical trial registry: mission, responsibility and operation. *Journal*
586 *of evidence-based medicine* **4**, 165–167 (2011).

- 587 [52] Egger, G. F. *et al.* European union clinical trials register: on the way to more trans-
588 parency of clinical trial data. *Expert Review of Clinical Pharmacology* **6**, 457–459 (2013).
- 589 [53] Shiokawa, T. Background, introduction and activity of the japan primary registries
590 network. *Journal of Evidence-Based Medicine* **2**, 41–43 (2009).
- 591 [54] Askie, L. M. Australian new zealand clinical trials registry: history and growth. *Journal*
592 *of Evidence-Based Medicine* **4**, 185–187 (2011).
- 593 [55] Faure, H. & Hrynaszkiwicz, I. The isrctn register: achievements and challenges 8 years
594 on. *Journal of evidence-based medicine* **4**, 188–192 (2011).
- 595 [56] Laguardia, J. *et al.* Brazilian clinical trials registry and the challenges for clinical research
596 governance. *Journal of Evidence-Based Medicine* **4**, 156–160 (2011).
- 597 [57] Park, H.-Y. Primary registry of the who international clinical trial registry platform:
598 Clinical research information service (cris). *Journal of the Korean Medical Association*
599 **54**, 92–97 (2011).
- 600 [58] Hasselblatt, H., Dreier, G., Antes, G. & Schumacher, M. The german clinical trials
601 register: challenges and chances of implementing a bilingual registry. *Journal of Evidence-*
602 *Based Medicine* **2**, 36–40 (2009).
- 603 [59] Solaymani-Dodaran, M., Ostovar, A., Khalili, D. & Vasei, M. Iranian registry of clini-
604 cal trials: path and challenges from conception to a world health organization primary
605 register. *Journal of Evidence-Based Medicine* **2**, 32–35 (2009).
- 606 [60] Tulvatana, W., Kulvichit, K., Thinkhamrop, B. & Tatsanavivat, P. Thai clinical trials
607 registry. *Journal of Evidence-Based Medicine* **4**, 182–184 (2011).
- 608 [61] Driessen, M. *et al.* The dutch nationwide trauma registry: the value of capturing all
609 acute trauma admissions. *Injury* **51**, 2553–2559 (2020).
- 610 [62] Abrams, A. & Siegfried, N. The pan african clinical trials registry: year one data analysis
611 of the only african member of the world health organization network of primary registries.
612 *Journal of Evidence-Based Medicine* **3**, 195–200 (2010).
- 613 [63] Ranawaka, U. K. & Goonaratna, C. The sri lanka clinical trials registry—moving forward.
614 *Journal of Evidence-Based Medicine* **4**, 179–181 (2011).
- 615 [64] Elsevier Science. Embase [electronic database]. Electronic Database (1974). Produced
616 by Elsevier Science, Amsterdam, The Netherlands.
- 617 [65] Canese, K. & Weis, S. Pubmed: the bibliographic database. *The NCBI handbook* **2**
618 (2013).