CT Open: An Open-Access, Uncontaminated, Live Platform for the Open Challenge of Clinical Trial Outcome Prediction

Jianyou Wang[1,*], **Youze Zheng**[1,*], **Longtian Bao**[1,*], **Hanyuan Zhang**[1,*], **Qirui Zheng**[1], **Yuhan Chen**[1], **Yang Zhang**[1], **Matthew Feng**[1], **Maxim Khan**[3], **Aditya K. Sehgal**[3], **Christopher D. Rosin**[3], **Ramamohan Paturi**[1], **Umber Dube**[2], **Leon Bergen**[1]
[1]Laboratory for Emerging Intelligence, University of California San Diego
[2]Department of Dermatology, University of California San Diego
[3]Elsevier
{jiw101, yoz018, haz146}@ucsd.edu

Abstract

Scientists have long sought to accurately predict outcomes of real-world events before they happen. Can AI systems do so more reliably? We study this question through clinical trial outcome prediction, a high-stakes open challenge even for domain experts. We introduce **CT Open**, an open-access, live platform that would run four challenge every year. Anyone can submit predictions for each challenge. CT Open evaluates those submissions on trials whose outcomes were not yet public at the time of submission but became public afterwards. Determining if a trial's outcome is public on the internet before a certain date is surprisingly difficult. Outcomes posted on official registries may lag behind years, while the first mention may appear in obscure articles. To address this, we propose a novel, fully automated decontamination pipeline that uses iterative LLM-powered web search to identify the earliest mention of trial outcomes. We validate the pipeline's quality and accuracy by human expert's annotations. Since CT Open's pipeline ensures that every evaluated trial had no publicly reported outcome when the prediction was made, it allows participants to use any methodology and any data source. In this paper, we release a training set and two time-stamped test benchmarks, Winter 2025 and Summer 2025. We believe CT Open can serve as a central hub for advancing AI research on forecasting real-world outcomes before they occur, while also informing biomedical research and improving clinical trial design. Code to reproduce our experiments is available at https://github.com/ClinicalTrial-OpenChallenge/CT_Open.

1 Introduction

Large language model's performance is increasingly competitive with humans on a wide range of tasks. This progress raises a natural question: beyond assisting humans on tasks they can already perform, can AI systems make reliable predictions on high-stakes problems whose answers are not yet known? In this paper, we study one such problem: predicting the outcome of a clinical trial before the trial has published results or before it has even begun.

Clinical trial outcome prediction is both difficult and consequential. If the problem were easy, clinical development would not remain so uncertain, with many launched trials failing their objectives. Accurate prediction matters to many, such as clinicians, patients and pharmaceutical companies. For AI research, this challenge is also compelling because it tests whether LLM-based systems can make useful predictions in a domain with immediate and substantial real-world consequences.

*equal contributions

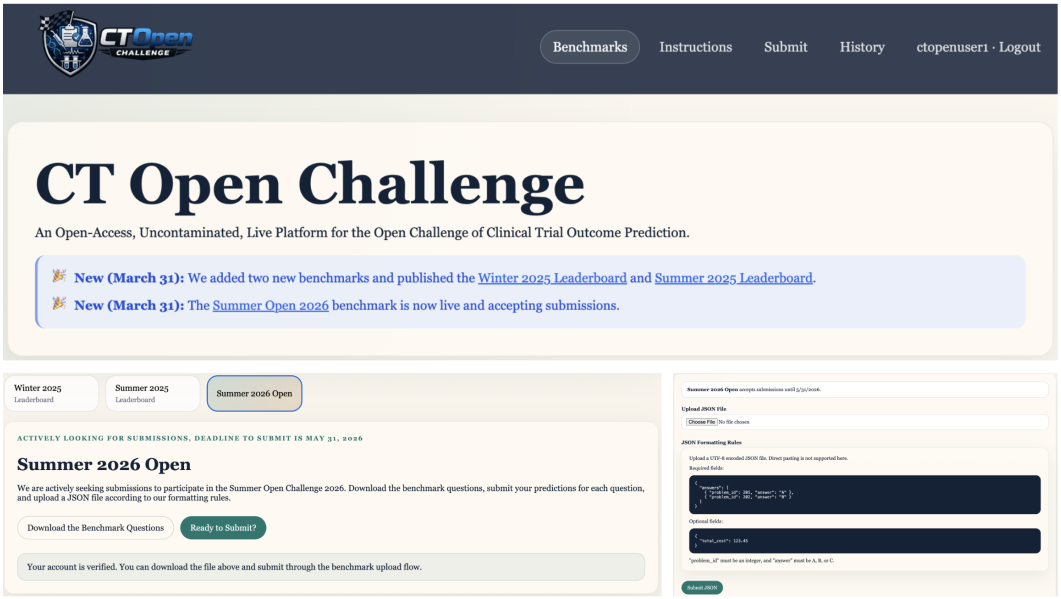

Figure 1: CT Open Platform Screenshots. It has released leaderboards for Winter 2025 and Summer 2025 benchmarks. It is now accepting submissions for Summer 2026 Open.

To facilitate finding a solution to this challenge, we introduce **CT Open**, a dynamic platform for clinical trial outcome prediction. See Figure 1 for website design. CT Open hosts four challenges each year: Winter Open, Spring Open, Summer Open, and Fall Open. Each challenge corresponds to a three-month window and a fixed set of clinical trial prediction questions. Participants entering a given challenge must submit their predictions before that challenge's window begins. After the window ends, CT Open evaluates all submissions and releases a leaderboard the following week. Specifically, from the pool of clinical trials associated with that challenge's question set, CT Open identifies the subset of trials that had no public results before the start of the challenge window but acquired usable public results during it. CT Open ranks submissions using only questions from this subset and their derived answers from public results, which are then released as a time-stamped benchmark.

CT Open is distinguished by three properties. First, it is **dynamic**. Unlike static benchmarks, CT Open will release new time-stamped benchmarks every three-month via its automated, expert-validated pipelines. This helps to evaluate newly released models, while reducing concerns that performance is driven by memorization or benchmark leakage. Second, it is **open-access** and **open-method**. Anyone can submit predictions, and participants are not restricted to a predefined evaluation protocol. All methods are permitted. We view this openness as essential: clinical trial prediction is a genuinely difficult open problem, and it is unlikely that a single fixed methodology will be sufficient. Third, CT Open is designed to be **contamination-resistant**. Our novel and robust decontamination pipeline identifies trials with no publicly available results prior to each challenge's window, including preliminary, interim, subgroup, conference, journal article, news, blog, financial-disclosure evidence, etc. Since submissions are made before each challenge's window, strong performance on CT Open for each challenge is much more plausibly attributable to real predictive ability rather than recall of previously published (partial) results.

Furthermore, CT Open includes a training set of questions and two time-stamped test benchmarks, Winter 2025 and Summer 2025, with cutoff dates of 2025-01-31 and 2025-08-31, respectively. By construction, no clinical trial included in Winter 2025 or Summer 2025 had any publicly available results before its corresponding cutoff date. Therefore, these benchmarks are suitable for evaluating LLMs whose knowledge cutoff, or preferably whose release date, predates the corresponding cutoff timestamp.

In this paper, we present initial evaluations that illustrate the range of methods CT Open can support. We compare prompt-based LLMs, retrieval-augmented LLMs, agentic LLMs, traditional machine learning methods and neural networks, and find that simple ML and NN baselines remain competitive with, and sometimes outperform, LLM-based approaches. This highlights an important gap in current LLM capabilities: despite their strong performance on many existing benchmarks, they still struggle on difficult predictive tasks such as clinical trial outcome forecasting. In addition, we show that RAG can provide useful evidence from similar completed trials, although its gains are limited and inconsistent, while agentic search can uncover valuable external information but remains computationally expensive and yields mixed results. Finally, by analyzing certain LLM's inconsistent behavior across Winter 2025 and Summer 2025, we show why dynamic, time-stamped, contamination-resistant benchmarks are necessary for fair evaluation.

2 Related Work

Clinical trial outcome prediction has been studied under a range of formulations, from machine learning and neural network methods over structured drug-development and registry features to more recent multimodal methods that incorporate richer trial context (Lo et al., 2019; Siah et al., 2021; Fu et al., 2022; Wang et al., 2023; Zheng et al., 2025; Hu et al., 2025). Clinical trial benchmarks (Chen et al., 2025; Gao et al., 2024) mostly target trial-level result, whereas CT Open consists with fine-grained outcome measure level questions with respect to specific study arms.

Retrieval-augmented generation (RAG) augments an LLM's parametric knowledge with retrieved information at inference time, improving factual grounding while enabling access to updated external information (Lewis et al., 2021). Subsequent work has expanded this paradigm through iterative, self-reflective, and domain-specialized retrieval strategies (Asai et al., 2023; Jeong et al., 2024; Xiong et al., 2024). This is directly relevant to our setting, where access to trial-relevant evidence such as registry entries, prior publications, and drug- or disease-related context may materially affect outcome prediction.

Agentic Systems with Tools and Web Access. A related line of work studies agentic LLM systems that interleave reasoning with external actions such as web browsing, tool use, API calls, and code execution (Nakano et al., 2022; Yao et al., 2023; Schick et al., 2023; Tang et al., 2024; Yue et al., 2024). Interestingly, most benchmarks, static or dynamic, do not allow agentic system with unrestricted web-search because they can find these published benchmarks and download their answers, while CT Open's unique design makes it suitable to evaluate such systems.

LLM Auto-Research and AI Scientist Systems. Recent work has pushed LLM-based systems beyond single-query question answering toward more autonomous, multi-step research workflows, including literature synthesis, experimentation, and report writing (Schmidgall et al., 2025; Wang et al., 2024; Lu et al., 2024; Yamada et al., 2025; Gottweis et al., 2025). Our benchmark connects naturally to this line of work by providing an evaluation setting for systems that can run experiments to produce more informed scientific judgments.

Dynamic Benchmark. Most LLM benchmarks are static: they are released once and then reused across successive model generations. Even difficult benchmarks such as GPQA and OlympiadBench become increasingly vulnerable to contamination as newer models are trained on larger web-scale corpora (Rein et al., 2023; He et al., 2024; Balloccu et al., 2024; Deng et al., 2024). This has motivated dynamic evaluation, from human-and-model-in-the-loop data collection in Dynabench (Kiela et al., 2021) to temporally refreshed benchmarks such as LiveBench (White et al., 2025), LiveCodeBench (Jain et al., 2024), ForecastBench (Karger et al., 2025), MathArena Balunović et al. (2026), and LiveMedBench (Yan et al., 2026). Our benchmark follows this temporally refreshed perspective for clinical trial outcome prediction by updating quarterly with newly available outcome evidence from the internet, enabling continual evaluation on previously unseen trial results.

3 Definition of Key Terms

CT Open focuses on randomized controlled clinical trials that evaluate drug interventions and are registered on clinicaltrials.gov. We highlight a few important definitions below.

Definition (NCT ID)**.** The NCT ID is a trial's identifier on clinicaltrials.gov.

Definition (Intervention/Treatment)**.** For CT Open's clinical trials, the intervention (i.e. treatment) is a drug, or a combination of drugs and medical procedures.

Definition (Study, Treatment and Comparator Arms)**.** A study arm is a group of participants assigned to a intervention or control. A treatment arm is a study arm receiving the intervention, while a comparator arm is control, such as placebo.

Definition (Outcome Measure and Endpoint)**.** An outcome measure, or endpoint, is a pre-specified quantity used to evaluate an intervention in a clinical trial. It defines what will be measured, and is part of the trial design rather than the trial result. For example, 5-year overall survival rate is an outcome measure.

Definition (Result and Outcome)**.** A result, or outcome, is the observed value of an outcome measure after the trial is conducted. It describes what actually happened in the trial. For example, 85% of patients survived for at least 5 years after treatment is a result, whereas 5-year overall survival rate is the corresponding outcome measure.

Definition (Challenge, Window, Time-Stamp, Cutoff Date)**.** A challenge's window is the period of time where we wait for results of clinical trials get published online. A time-stamped benchmark is released the first week after a challenge's window ends. The time-stamp (i.e. the cutoff date) is one day before the challenge's window begins. CT Open ensures no trial in this time-stamped benchmark has any form of results before the cutoff date.

4 The Design of CT Open

CT Open's long-term goal is to facilitate building AI systems that can truly predict clinical trial outcomes, especially at the earliest stage, when a trial is still being planned or has only recently begun. To this end, CT Open uses a novel, robust, carefully-designed **decontamination pipeline** to ensure that any system evaluated by CT Open cannot rely on partial, interim, or already published trial results, which often appear online in incomplete forms and in scattered places, long before final results are officially posted.

4.1 CT Open's Questions

CT Open's question is defined at the level of a specific outcome measure of the trial and is represented as a tuple of a clinical trial, a particular outcome measure, and the associated study arms (when applicable). CT Open includes three classes of questions. See the Appendix F for example questions and additional details for each class.

- The **Superiority** class consists of binary questions on whether the treatment arm shows a statistically significant improvement over the comparator arm.
- The **Comparative Effect** class consists of three-way questions on whether the treatment arm is statistically significantly better than, worse than, or not different from the comparator arm.
- The **Endpoint** class consists of two closely-related binary subtypes of questions: one asks whether the trial meets the endpoint, and the other asks whether at least one arm in the trial meets the endpoint.

Throughout this paper, meeting an endpoint refers to achieving a statistically significant positive result on an outcome measure. Unless otherwise noted, statistical significance is defined at the 95% confidence level.

Because clinical trials specify their treatment arms, comparator arms, and outcome measures (i.e. endpoints) even before they start, CT Open can automatically generate questions from

	Winter 2025	Summer 2025	Train
Unique Trials	314	240	7,292
Total Questions	605	857	15,444
Superiority / Comparative / Endpoint (%)	78% / 9% / 13%	77% / 6% / 17%	51% / 31% / 18%
Trials started in or after 2020 (%)	89.1	88.4	8.8
Median Enrollment	234	212	275
Unique Countries	82	81	131

Table 1: Overall Statistics of the three datasets of CT Open. All three trial sets are disjoint.

the trial information without access to the trial results. When the results are later reported, they can be used to answer those questions via CT Open's **answer generation pipeline**.

4.2 The Components of CT Open

CT Open has two main components: a static component and a dynamic component.

As shown in Table 1, the static component consists of a training set of 7292 unique clinical trials with officially reported results, together with approximately 15,444 questions derived from those trials. The static component also includes two time-stamped test benchmarks, Winter 2025 (605 questions) and Summer 2025 (857 questions), with respective time-stamps at 2025-01-31 and 2025-08-31. They encompass a diverse range of biological and medical topics, including oncology, cardiovascular diseases, autoimmune disorders and neurology. For details, refer to Table 4 and Table 5 in Appendix A.

The dynamic component consists of a series of time-stamped benchmarks released every three months. The first benchmark is scheduled to be constructed via CT Open's **decontamination pipeline** and **answer generation pipeline**, and released on 2026-09-07 and will use a cutoff date of 2026-05-31. Future benchmarks will be constructed and released on the same quarterly schedule. Note the Winter 2025 benchmark has a different cutoff date.

4.3 The Schedule of CT Open

Challenge	Window	Submission Deadline	Leaderboard Release
Winter Open	December–March	November 30	First week of March
Spring Open	March–June	Last day of February	First week of June
Summer Open	June–September	May 31	First week of September
Fall Open	September–December	August 31	First week of December

Table 2: CT Open hosts four challenges every year. Submission begins when the question pool for each challenge is released, usually 2-3 months before deadline.

CT Open hosts four challenges each year: Winter Open, Spring Open, Summer Open, and Fall Open (Table 2). The CT Open website is scheduled to go live in the first week of April 2026, so the first active challenge will be Summer Open 2026. Participants may submit predictions until May 31, 2026, and the leaderboard will be released in the first week of September 2026. Submissions received after that deadline will be evaluated in the Fall Open.

For each challenge, CT Open releases a downloadable question set derived from trials that are likely to report results within that challenge's window. At the end of the window, CT Open runs its **answer generation pipeline** and **decontamination pipeline** to identify the subset of trials with usable results posted during that window and no public results posted before the window begins. Trials that fail this condition are considered unusable or contaminated, and are removed. The remaining trials, and their questions and answers form the time-stamped benchmark for that challenge, which is then used to evaluate submissions and produce the leaderboard.

Importantly, CT Open only evaluates submissions made before a challenge window begins, and it excludes any clinical trial for which results are publicly available before that window.

As a result, systems evaluated on CT Open cannot benefit from prior access to the outcomes of the trials used for evaluation. Therefore, any substantial improvement in performance is more likely to reflect genuine predictive ability rather than retrieval or recall of existing results.

4.4 The Decontamination Pipeline

To construct a time-stamped benchmark for each challenge, CT Open applies a novel and robust decontamination pipeline to identify suitable clinical trials from the pool of trials associated with that challenge's question set. This pipeline combines multiple search mechanisms, including LLM-based web search and a customized web and database search system, to detect publicly available trial results and exclude contaminated trials.

We define the following notation:

$$\textsc{Search}(query, [start, end], mode) \longrightarrow \{doc_1, doc_2 ...\}$$

as a search procedure that uses *query* to retrieve all documents about a clinical trial, subject to the publication-date constraint $[start, end]$. Each returned document must be verifiably about the queried trial and must report trial results. A document can take any form: journal article, press release, financial statement, blog posts, etc. The *query* may be a keyword (e.g. trial's NCT ID), a sentence-length query or a multi-paragraph description. If *start* date is $-\infty$, no lower bound is imposed on publication date; if *end* date is $+\infty$, results may be as recent as the day of search.

The parameter *mode* specifies the retrieval system. If *mode* is an LLM (e.g., GPT-5.2 or Gemini 3), search is performed through the corresponding model's web-search capability. If *mode* is Brave, we use a customized pipeline in which Brave Search API returns candidate URLs, webpages are fetched with Python Requests or ZenRows, and downloadable files are converted to text with DeepSeek OCR (Wei et al., 2025). Scraped webpages are then processed with a customized text-extraction system derived from trafilatura (Barbaresi, 2021), designed for long pages and for isolating the main article content while removing HTML, sidebars, and other articles' text. If *mode* = Database, we use the trial's NCT ID to search PubMed, PMC, BioRxiv, and MedRxiv for matching articles. If *mode* is Brave or Database, all retrieved texts, whether from webpages, downloadable files, or databases, first undergo our customized pipeline for published date extraction to verify its within [*start*, *end*], and then undergo two rounds of GPT-5 verification: first, to confirm that the document refers to the same clinical trial; and second, to confirm that it reports some form of results for that trial. See Figure 8 and Figure 9 for prompts.

Input: Initial trial pool $\mathcal{P}$, cutoff date c
Output: Filtered trial set $\mathcal{F}$
$\mathcal{F} \leftarrow \emptyset$;
foreach trial $\in \mathcal{P}$ **do**
  info $\leftarrow$ {NCT ID, aliases, title, description};
  $U_{\text{before}} \leftarrow \text{SEARCH}(\text{info}, [-\infty, c), \text{GPT-5.2}) \cup \text{SEARCH}(\text{info}, [-\infty, c), \text{Gemini 3})$;
  if $U_{\text{before}} \neq \emptyset$ **then continue**;
  $U_{\text{after}} \leftarrow \text{SEARCH}(\text{info}, [c, +\infty), \text{GPT-5.2}) \cup \text{SEARCH}(\text{info}, [c, +\infty), \text{Gemini 3})$;
  if $U_{\text{after}} = \emptyset$ **then continue**;
  Generate three rewritten queries Q_1, Q_2, Q_3 from info using GPT5;
  $U_{\text{brave}} \leftarrow \bigcup_{i \in \{1,2,3\}} \text{SEARCH}(Q_i, [-\infty, c), \text{Brave})$;
  $U_{\text{db}} \leftarrow \text{SEARCH}(\text{NCT ID}, [-\infty, c), \text{Database})$;
  if $U_{\text{brave}} \neq \emptyset$ **or** $U_{\text{db}} \neq \emptyset$ **then continue**;
  add t to $\mathcal{F}$;
end
return $\mathcal{F}$

Algorithm 1: Filtering trials with pre-cutoff results or no post-cutoff results

Algorithm 1 summarizes our decontamination pipeline. To assess its robustness, we performed an additional validation using Embase, a proprietary Elsevier database that includes

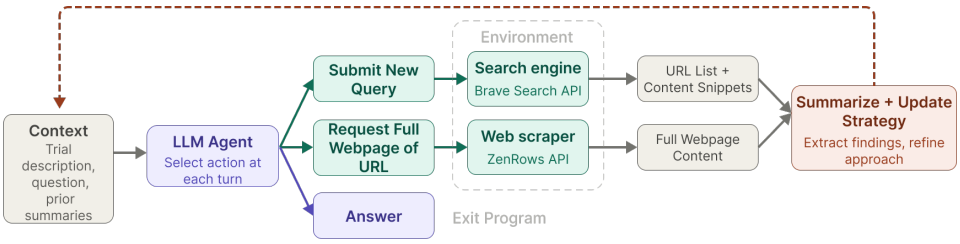

Figure 2: Agentic LLM Workflow

biomedical literature and conference abstracts not necessarily accessible through public search. Because our pipeline relies entirely on publicly available information, Embase provides a useful external check for residual contamination. We applied this validation to all clinical trials that passed our pipeline for both the Winter 2025 and Summer 2025 benchmarks and found no trial with reported results in Embase prior to the corresponding cutoff date. This shows our pipeline is robust even when it can only access public information.

We also randomly sampled 50 NCT IDs for manual review. For each sampled trial, one researcher conducted extensive search using Google Search with manually designed queries, coupled with GPT-5 web search, and GPT-5 Deep Research as supplements, examining more than 200 webpages per trial. For each trial, this process could take 2 to 4 hours. Our annotators confirmed that 49 of the 50 sampled NCT IDs had no reported results before their respective cutoff dates. Only one trial was classified as ambiguous. Our annotator identified a 2024 LinkedIn post by the trial's principal investigator that included a blurry photograph of a conference poster appearing to show partial trial results. This single image is the only evidence available prior to the cutoff date, and, to the best of our knowledge, current LLMs are pretrained on textual data. To check the impact of this image, we showed that o3-mini, GPT-5, Opus 4.5 Qwen-2.5-14B and DeepSeek-Distilled-Qwen-2.5-14B have no recall of this image's information. GPT-5 and Opus 4.5 could not use web-search to find it. We asked each model to predict the trial's outcome and inspected its reasoning traces; none referenced or alluded to the poster.

Nevertheless, conservatively, the decontamination pipeline achieves an accuracy of at least 98%. As a precaution, we exclude this trial and its associated questions from the test benchmarks.

4.5 Answer Generation Pipeline

Each remaining trial is associated with a set of result documents published within the challenge window. We design a multi-stage GPT-5.4-based pipeline to determine whether the reported results are statistically rigorous; can be clearly linked to specific outcome measures and study arms; and are sufficient to support answering the questions associated with that trial. If the pipeline determines that the available results cannot support a question with sufficient certainty, that question is removed from the benchmark. See Figure 10 and Figure 11 for prompts.

To validate this pipeline, we randomly sampled 100 questions, answers and reported results. We then asked a physician-scientist experienced in clinical trial design and analysis to annotate these 100 cases and assess whether the proposed answers were supported by the reported results. The expert's annotation shows an 99% accuracy.

5 Evaluation

We evaluate o3-mini (OpenAI, 2025), GPT-5 (OpenAI, 2025), Claude Opus 4.5 (Anthropic, 2025) using the following approaches on Winter 2025 and Summer 2025. All prompts, agent

Model	Endpoint		Superiority		Comparative Effect	
	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc
Winter 2025						
Traditional Baselines						
Random Forest	66.60	71.54	55.39	63.76	**80.70**	**79.12**
Feed-Forward NN	67.74	72.29	54.23	57.91	58.87	<u>74.60</u>
Large Language Models						
o3-mini	68.40	70.23	<u>68.09</u>	68.33	51.91	51.53
+ RAG	**76.88** ↑↑	73.48 ↑	**69.21** ↑	69.18 ↑	56.81 ↑	57.14 ↑
+ Agent	64.83 ↓	66.67 ↓	67.46 ↓	68.06 ↓	54.74 ↑	55.72 ↑
GPT-5	65.29	77.63	66.17	<u>71.20</u>	54.25	54.70
+ RAG	67.98 ↑	78.03 ↑	67.54 ↑	**71.86** ↑	56.96 ↑	57.49 ↑
Claude Opus 4.5[†]	70.17	<u>80.23</u>	62.31	68.91	55.52	54.83
+ RAG	<u>70.36</u> ↑	**81.85** ↑	63.43 ↑	69.62 ↑	56.00 ↑	55.19 ↑
Open-Source Large Language Models						
Qwen2.5-14B	53.29	53.68	60.04	60.21	58.00	62.94
+ RAG	68.77 ↑	66.27 ↑	62.67 ↑	62.34 ↑	<u>60.32</u> ↑	60.64 ↓
+ Agent	61.41 ↑	55.63 ↑	59.74 ↓	61.53 ↑	34.20 ↓	33.78 ↓
DeepSeek-R1-14B	53.54	54.18	56.68	58.02	45.69	55.05
+ RAG	67.56 ↑	64.21 ↑	61.26 ↑	61.09 ↑	32.30 ↓	41.31 ↓
+ Agent	66.23 ↑	64.07 ↑	61.21 ↑	61.77 ↑	51.21 ↑	56.78 ↑
Summer 2025						
Traditional Baselines						
Random Forest	**65.57**	<u>66.48</u>	55.88	61.63	46.15	45.65
Feed-Forward NN	<u>65.25</u>	**67.78**	64.95	67.58	57.56	<u>85.87</u>
Large Language Models						
o3-mini	59.83	59.46	72.48	71.93	54.94	59.78
+ RAG	59.07 ↓	58.98 ↓	<u>73.15</u> ↑	72.94 ↑	56.75 ↑	60.87 ↑
+ Agent	61.75 ↑	62.84 ↑	**73.33** ↑	<u>73.96</u> ↑	**69.81** ↑↑	**92.75** ↑↑
GPT-5	51.54	58.24	70.23	73.71	64.81	64.49
+ RAG	50.13 ↓	55.55 ↓	70.18 ↓	73.78 ↑	<u>68.95</u> ↑	70.05 ↑
Claude Opus 4.5	53.71	57.86	68.76	73.68	48.24	49.64
+ RAG	58.69 ↑↑	62.72 ↑↑	69.59 ↑	**74.10** ↑	48.42 ↑	50.00 ↑
Open-Source Large Language Models						
Qwen2.5-14B	53.42	53.61	58.84	59.43	44.74	44.69
+ RAG	58.91 ↑	58.55 ↑	61.00 ↑	61.02 ↑	45.35 ↑	44.20 ↓
+ Agent	60.60 ↑	59.82 ↑	55.13 ↓	57.85 ↓	32.14 ↓	23.68 ↓
DeepSeek-R1-14B	52.04	52.25	57.95	59.65	41.39	51.21
+ RAG	58.32 ↑	57.83 ↑	61.48 ↑	61.79 ↑	32.92 ↓	47.71 ↓
+ Agent	57.31 ↑	56.90 ↑	57.59 ↓	58.33 ↓	29.63 ↓	21.05 ↓

[†] Opus 4.5 knowledge cutoff date postdates Winter 2025's timestamp, possibly contaminated.

Table 3: Acc is balanced accuracy. Compared to base model, ↑and ↓indicate nominal increases and decreases, while ↑↑denotes statistically significant improvements (95% CI via hierarchical bootstrap sampling). Best and second-best results within each dataset and column are marked in **bold** and <u>underline</u>, respectively. All LLM results are averaged over three runs.

protocols, and hyperparameters for the traditional machine learning and neural network methods are developed on the validation set, which is 10% of the training set.

Prompt-based LLM approach: We provide the model with the trial title, trial description, drug description when available, patient eligibility criteria, the question and its outcome measure, and the relevant study arms when available. The model is then asked to predict

the most likely answer choice. We found that adding numerical and categorical variables reduced performance, so they were excluded. See Appendix E Figure 12 for prompt.

Machine Learning and Neural Network methods: We train on 90% of the training set and reserve the remaining 10% for hyperparameter tuning. We include all available categorical and numerical trial features, such as phase and enrollment. See Appendix B.2 for details.

Agentic LLM: We design an iterative search-and-browse workflow. At each step, the agent chooses either to issue a new query to a Search engine or to inspect the full content of a previously retrieved URL. Query results are filtered to include only documents whose publication dates are verified to precede the benchmark cutoff, and the agent receives the corresponding URLs together with their content snippets. After each step, the agent summarizes the newly observed information and may revise its search strategy; only these summaries, rather than the full retrieved content, are retained in context. Agent can interact with the environment for up to 25 rounds before producing answer. See Figure 2 for agentic workflow and see Appendix E for agent protocol in Figure 16.

LLM with RAG: In preliminary experiments, retrieving preclinical evidence, such as animal and observational studies, consistently degraded baseline LLM performance on the validation set. We found that retrieval is more useful when it focuses on previously completed clinical trials with publicly available results that closely resemble the target trial in drug mechanism, design, outcome measure, and patient population. Because these prior trials are not exact matches, their outcomes cannot be used by simple aggregation; rather, the model must reason about how small differences in trial design may affect the final result. Motivated by this observation, we developed a customized retrieval pipeline to get a list of similar completed trials, averaging about 15 per target trial, each paired with its most relevant endpoint. We then provide the LLM with these similar trials and their reported results from `clinicaltrials.gov`, and ask it to reason over the differences between the retrieved trials and the target trial before predicting the final answer. See Figure 13, 14, 15 for prompts.

6 Analyses and Future Directions

Traditional ML and NN methods remain competitive with LLM-based approaches: This suggests that these models can effectively learn correlations between categorical, numerical features and the target labels. By contrast, off-the-shelf LLMs appear less effective at using such variables, consistent with our initial experiments. These findings point to a promising direction for future work: improving LLMs' ability to learn and exploit correlational structure in tabular and other structured data, or enabling LLM-driven systems to train more effective classical predictors, such as random forests and lightweight neural networks, in an auto-research or AI-scientist-style setting. More broadly, the results highlight a persistent challenge for LLM research: although modern LLMs perform strongly on many tasks that humans can already do, they still struggle on open predictive problems that remain difficult for humans themselves. For brevity, Table 3 reports only the strongest traditional baseline; full baseline results are provided in Table 6 in the Appendix B. A grouped permutation-importance analysis of which feature groups these baselines rely on is provided in Appendix B.3.

RAG yields modest improvements over the baseline in certain cases: We manually examined cases in which the LLM with RAG produced an incorrect answer while the baseline LLM answered correctly. In most such cases, the retrieved similar trials were not themselves irrelevant; rather, the model appeared unable to properly account for subtle but consequential differences between the retrieved trials and the target trial. Nevertheless, for question classes where RAG did improve performance, for example, comparing o3-mini with o3-mini + RAG on Winter 2025, and Opus 4.5 with Opus 4.5 + RAG on Summer 2025, the reported outcomes of similar prior trials appear to provide a useful starting point for reasoning about the target trial and forming a prediction.

The Agent setting produces mixed results: We found the agent setting to be both computationally expensive and double-edged. Running the current agent configuration with

o3-mini costs more than $1000 in tokens, so we were only able to evaluate it with o3-mini. On Winter 2025, o3-mini with the agent setting nominally underperformed the o3-mini baseline. On Summer 2025, however, o3-mini with the agent setting outperformed the baseline, particularly on Comparative Effect questions. Manual inspection suggests that the agent can sometimes discover unexpected but highly informative sources that substantially alter its search strategy. For example, in one case it encountered a `withpower.com` result, which prompted the agent to include `withpower` in a subsequent query, allowing Brave Search to target the site directly. Because WithPower contains relevant information and analyses on clinical trials, the agent was able to extract additional useful evidence from related trials hosted there. In another case, the agent discovered `https://copingwithmsa.com/`, a patient-focused website containing firsthand accounts related to clinical trials. By refining its queries to search that site more deeply, the agent located a relevant patient blog describing experiences with a similar trial, which provided useful evidence for prediction.

Open-weight models benefit from retrieval but not from the agent setting: Qwen2.5-14B and DeepSeek-R1-14B score 52.04-53.54 macro-F1 on Endpoint in both benchmarks, close to random assignment on a binary task. RAG improves both models on the two binary tasks, with the largest gains on Winter 2025 Endpoint (+15.48 and +14.02 macro-F1), but its effect on three-way Comparative Effect is mixed. The agent setting improves both models on Endpoint only and lowers Qwen2.5-14B's macro-F1 on the other two tasks in both benchmarks. Error analysis with our biomedical expert identified two failure modes. First, the models retrieve relevant prior trials without adjusting for design differences: on NCT06360458 (MIRACLE), Qwen2.5-14B retrieved the comparable MARVEL trial but did not account for MIRACLE enrolling the subgroup MARVEL explicitly excluded. Second, despite explicit instruction otherwise, both models spent much of their query budget searching for unpublished outcomes that our decontamination pipeline ensures were unavailable before the cutoff date, providing further evidence that the benchmark is not solvable by retrieval.

Pretrained knowledge may affect performance on contaminated benchmarks: We intentionally evaluated Opus 4.5 on Winter 2025, whose cutoff date is January 31, 2025, even though Opus 4.5 has a knowledge cutoff of August 31, 2025. On Winter 2025, Opus 4.5 outperformed o3-mini and GPT-5 in five settings: Endpoint macro-F1 and accuracy, Comparative Effect macro-F1 and accuracy, and Superiority accuracy. By contrast, on Summer 2025, whose cutoff date is August 31, 2025 and is therefore technically uncontaminated with respect to Opus 4.5's knowledge cutoff, Opus 4.5 underperformed o3-mini and GPT-5 in four settings: Endpoint accuracy, Superiority macro-F1, and Comparative Effect macro-F1 and accuracy. These results illustrate the importance of dynamic, uncontaminated benchmarks, which enable recently released models such as Opus 4.5 to be evaluated more fairly and reward genuine predictive ability rather than recall from pretraining data.

7 Conclusion

In this paper, we present CT Open and its automated, expert-validated benchmark construction algorithms. The design of CT Open will ensure that it is contamination-resistant and that it would properly differentiate systems that display genuine predictive abilities over systems that are over-fitted or contaminated. CT Open will serve as a central hub that brings together the community interested in clinical trial outcome prediction or advancing the capabilities of LLMs. Our analyses have revealed critical gaps of current RAG and agentic systems for performing predictive tasks, which highlights many opportunities to further improve these systems, and even potentially conceive new methodology.

Ethics Statement

CT Open does not and will never publish patient's personal information or display them on the platform. All questions, answers, and evaluations use publicly available information that is aggregate statistics, and not individual's information. All clinical trials at CT Open are approved and standard trials registered at `clinicaltrials.gov`.

References

Anthropic. Introducing claude opus 4.5, November 2025. URL https://www.anthropic.com/news/claude-opus-4-5. Anthropic release page.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023. URL https://arxiv.org/abs/2310.11511.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms, 2024. URL https://arxiv.org/abs/2402.03927.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, 2026. URL https://arxiv.org/abs/2505.23281.

Adrien Barbaresi. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In Heng Ji, Jong C. Park, and Rui Xia (eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pp. 122–131, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.15. URL https://aclanthology.org/2021.acl-demo.15/.

Jintai Chen, Yaojun Hu, Mingchen Cai, Yingzhou Lu, Yue Wang, Xu Cao, Miao Lin, Hongxia Xu, Jian Wu, Cao Xiao, Jimeng Sun, Yuqiang Li, Lucas Glass, Kexin Huang, Marinka Zitnik, and Tianfan Fu. Trialbench: Multi-modal artificial intelligence-ready clinical trial datasets, 2025. URL https://arxiv.org/abs/2407.00631.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models, 2024. URL https://arxiv.org/abs/2311.09783.

Zhao Fan, Fanyu Xu, Cai Li, and Lili Yao. Application of kpca and adaboost algorithm in classification of functional magnetic resonance imaging of alzheimer's disease. Neural Computing and Applications, 32(10):5329–5338, 2020.

Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Hint: Hierarchical interaction network for clinical-trial-outcome predictions. Patterns, 3(4), 2022.

Chufan Gao, Jathurshan Pradeepkumar, Trisha Das, Shivashankar Thati, and Jimeng Sun. Automatically labeling clinical trial outcomes: A large-scale benchmark for drug development. arXiv preprint arXiv:2406.10292, 2024.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an ai co-scientist, 2025. URL https://arxiv.org/abs/2502.18864.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL https://arxiv.org/abs/2402.14008.

Qixuan Hu, Xumou Zhang, Jinman Kim, Florence Bourgeois, and Adam G. Dunn. A novel language model for predicting serious adverse event results in clinical trials from their prospective registrations, 2025. URL https://arxiv.org/abs/2507.22919.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024. URL https://arxiv.org/abs/2403.07974.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models, 2024. URL https://arxiv.org/abs/2401.15269.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities, 2025. URL https://arxiv.org/abs/2409.19839.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in nlp, 2021. URL https://arxiv.org/abs/2104.14337.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. arXiv preprint arXiv:2405.17428, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Andrew W Lo, Kien Wei Siah, and Chi Heem Wong. Machine learning with statistical imputation for predicting drug approvals, volume 60. SSRN, 2019.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL https://arxiv.org/abs/2408.06292.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL https://arxiv.org/abs/2112.09332.

OpenAI. Openai o3-mini. https://openai.com/index/openai-o3-mini/, January 2025. OpenAI release page.

OpenAI. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023. URL https://arxiv.org/abs/2302.04761.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants, 2025. URL https://arxiv.org/abs/2501.04227.

Kien Wei Siah, Nicholas W Kelley, Steffen Ballerstedt, Björn Holzhauer, Tianmeng Lyu, David Mettler, Sophie Sun, Simon Wandel, Yang Zhong, Bin Zhou, et al. Predicting drug approvals: The novartis data science and artificial intelligence challenge. Patterns, 2(8), 2021.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Findings of the Association for Computational Linguistics: ACL 2024, pp. 599–621, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.33. URL https://aclanthology.org/2024.findings-acl.33/.

Léon-Charles Tranchevent, Francisco Azuaje, and Jagath C Rajapakse. A deep neural network approach to predicting clinical outcomes of neuroblastoma patients. BMC medical genomics, 12(Suppl 8):178, 2019.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. Autosurvey: Large language models can automatically write surveys, 2024. URL https://arxiv.org/abs/2406.10252.

Zifeng Wang, Cao Xiao, and Jimeng Sun. Spot: Sequential predictive modeling of clinical trial outcome with meta-learning, 2023. URL https://arxiv.org/abs/2304.05352.

Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression, 2025. URL https://arxiv.org/abs/2510.18234.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited llm benchmark, 2025. URL https://arxiv.org/abs/2406.19314.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Improving retrieval-augmented generation in medicine with iterative follow-up questions, 2024. URL https://arxiv.org/abs/2408.00727.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search, 2025. URL https://arxiv.org/abs/2504.08066.

Zhiling Yan, Dingjie Song, Zhe Fang, Yisheng Ji, Xiang Li, Quanzheng Li, and Lichao Sun. Livemedbench: A contamination-free medical benchmark for llms with automated rubric evaluation, 2026. URL https://arxiv.org/abs/2602.10367.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. In Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400713026. doi: 10.1145/3698587.3701359. URL https://doi.org/10.1145/3698587.3701359.

Wenhao Zheng, Liaoyaqi Wang, Dongshen Peng, Hongxia Xu, Yun Li, Hongtu Zhu, Tianfan Fu, and Huaxiu Yao. LIFTED: Multimodal clinical trial outcome prediction via large language models and mixture-of-experts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 7503–7517, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.396. URL https://aclanthology.org/2025.findings-emnlp.396/.

Appendix: Table of Contents

A. Dataset Statistics and Full Evaluation Results 15
B. Traditional Baseline Implementation Details 16
  1. Full data feature set for tabular machine learning models 16
  2. Baseline Implementation Details and Hyperparameters 16
  3. Feature Group Importance Analysis 16
C. Pipelines 16
  1. Decontamination Pipeline Implementation Detail 16
  2. Answer Generation Pipeline Implementation Detail 19
D. Dataset Generation Prompts 23
E. Evaluation Prompts 28
F. Question Classes, Label Distribution, and Example Questions 29

A Dataset Statistics and Full Evaluation Results

Tables 4 and 5 summarize the scale, temporal distribution, and disease-area composition of each data split. Table 6 presents the full evaluation results across all baselines and models. Table 10 in Appendix F reports the label distribution within each question class.

	Winter 2025	Summer 2025	Train
Scale			
Unique trials	314	240	7,292
(Trial, endpoint) pairs	561	781	10,842
Total questions	605	857	15,444
Superiority / Comparative / Endpoint (%)	78 / 9 / 13	77 / 6 / 17	51 / 31 / 18
Primary / Secondary / Other (%)	54 / 44 / 2	34 / 65 / 1	32 / 66 / 2
Temporal distribution *(study start year)*			
Year range	2011–2025	2009–2025	1989–2023
<2020	35 (11.1%)	25 (10.4%)	6,649 (91.2%)
2020–2023	229 (72.9%)	131 (54.6%)	625 (8.6%)
2024–2025	51 (16.2%)	81 (33.8%)	0 (0.0%)
Enrollment			
Median / Mean	234 / 922	212 / 382	275 / 660
Range	50–149,201	50–6,300	50–144,000
Trial phase			
Phase 3	81 (57.6%)	130 (54.2%)	4,250 (58.3%)
Phase 2/3	213 (7.3%)	12 (5.0%)	259 (3.6%)
Phase 2	91 (29.0%)	78 (32.5%)	2,534 (34.7%)
Phase 1/2	3 (1.0%)	3 (1.3%)	159 (2.2%)
Phase 1 / Early	16 (5.1%)	17 (7.1%)	90 (1.2%)
Geography			
Unique countries	82	81	131
Top-3 countries	US, ES, CN	US, CN, ES	US, CA, DE
Trials w/ US sites	181 (57.6%)	115 (47.9%)	5,282 (72.4%)
Intervention type			
Drug	690 (84.3%)	519 (80.2%)	17,814 (84.6%)
Biological	50 (6.1%)	67 (10.4%)	1,536 (7.3%)
Other	78 (9.5%)	61 (9.4%)	1,738 (8.3%)

Table 4: Dataset overview and key characteristics. All trials are interventional with no shared NCT IDs across any pair of splits. A trial may map to multiple disease categories.

Disease Area	Winter 2025	Summer 2025	Train
Oncology	24.5	21.3	21.1
Cardiovascular	13.7	8.3	5.9
Metabolic / Endocrine	8.3	14.2	10.7
Autoimmune / Inflamm.	8.0	9.6	6.7
Neurology / Psychiatry	7.3	9.2	9.6
Infectious Disease	6.4	6.7	8.7
Rare / Genetic	5.7	2.5	4.3
Respiratory	4.5	4.6	7.2
Dermatology	4.5	3.3	2.9
Pain / Anesthesia	4.5	2.5	4.5
Nephrology	3.8	3.8	2.5
Musculoskeletal	3.5	3.3	4.7
Other[†]	15.6	19.6	20.8

[†] Includes hematology, gastroenterology, ophthalmology, and unmatched conditions.

Table 5: Disease-area distribution (% of unique trials). Categories below 2% in all splits are grouped under "Other."

B Traditional Baseline Implementation Details

B.1 Full data feature set for tabular machine learning models

Table 7 presents a detailed overview of the feature schema for tabular machine learning baseline models.

B.2 Baseline Implementation Details and Hyperparameters

All baselines share a common feature preprocessing pipeline unless noted otherwise. **Numerical features** are used directly; missing values are imputed with the column median (KNN+RF instead uses k-nearest-neighbor imputation following Lo et al. (2019)). **Categorical features** are one-hot encoded; missing values are imputed with the most-frequent category. **Textual features** are embedded with NV-Embed-v2 (Lee et al., 2024) and reduced to 32 dimensions via Principal Component Analysis; missing values are replaced by a zero vector. The only exception is Logistic Regression, where it uses the full NV-Embed-v2 embeddings without PCA reduction. For FFNN, the epoch with the best validation performance is selected for evaluation on both test benchmark datasets.

B.3 Feature Group Importance Analysis

We ran a grouped permutation-importance analysis for the Random Forest and L2 Logistic Regression baselines. For each of 15 runs per classifier (3 task types $\times$ 5 random seeds), we permuted one feature group at a time, held all remaining groups and the labels fixed, and recorded the resulting drop in macro-F1. Textual features are encoded as vectors. All embedding dimensions of a given feature were permuted jointly, so that the contribution of a feature is not scaled by its embedding dimensionality. Each reported score is the mean drop across runs, weighted by per-task sample size, aggregated over both benchmarks. A larger value corresponds to a larger drop in macro F1 under permutation of that group.

Table 8 reports the union of the two models' top-10 groups, sorted by the larger of the two scores; "—" indicates that the group falls outside that model's top 10. For both models, the largest drops come from textual features: endpoint details, comparator arm, trial conditions, and eligibility criteria. Among non-textual features, trial locations and lead sponsor class produce the largest drops. Scores are not directly comparable across the two models, since Logistic Regression uses the full NV-Embed-v2 embeddings while Random Forest uses their 32-dimensional PCA reduction (Appendix B.2).

C Pipelines

C.1 Decontamination Pipeline Implementation Detail

We provide implementation details for the decontamination pipeline summarized in Algorithm 1. In Algorithm 1 we use a simplified notation $\textsc{Search}(query, [start, end], mode)$. In fact, we used this notation $\textsc{Search}(query, [start, end], mode, k)$, which retrieves up to k documents about a clinical trial that report results, subject to the publication date constraint $[start, end]$.

Initial Screening Pass. As shown in Algorithm 1, the first step of the pipeline is a quick screening pass over the full initial trial pool $\mathcal{P}$, which typically contains over 10,000 trials. For each trial t, we run $\textsc{Search}(t, [-\infty, +\infty], \text{GPT-5.2}, 1)$, which terminates as soon as a single document reporting results is found. Trials for which no result is found anywhere on the web are immediately discarded, since they cannot serve as valid evaluation targets. This step reduces the candidate set from approximately 10,000 trials to approximately 2,000, significantly lowering the cost of the more intensive searches that follow. The prompt used for this initial screening is shown in Figure 3.

LLM Mode Search. When *mode* is an LLM (GPT-5.2 or Gemini 3), search is performed through the corresponding model's web search capability, as described in the main text. We implement this in two rounds. For each trial, we flatten the protocol section (excluding eligibility criteria and any date containing fields) into a text representation. In Round 1, we ask the model to perform a grounded web search and return a list of URLs that report results for that exact trial (not related trials or earlier phases). The prompts used for the before-cutoff and after-cutoff searches are shown in Figure 4 and Figure 5, respectively. We normalize the raw LLM output into clean URL lists, handling varied formats such as Python literals, JSON, and raw text with URL extraction. In Round 2, we send the discovered URLs back to the model with a second prompt asking it to visit each URL and extract the publication date and concise result summaries (e.g., pCR rates, hazard ratios, efficacy/safety findings). The prompt used for this extraction step is shown in Figure 6. We parse this second round output using regex to segment multi-URL responses into structured dictionaries mapping each URL to its date and extracted results.

As specified in Algorithm 1, we run LLM mode searches in four configurations: $\textsc{Search}(\text{info}, [-\infty, c), \text{GPT-5.2}, \infty)$, $\textsc{Search}(\text{info}, [-\infty, c), \text{Gemini 3}, \infty)$, $\textsc{Search}(\text{info}, [c, +\infty), \text{GPT-5.2}, \infty)$, and $\textsc{Search}(\text{info}, [c, +\infty), \text{Gemini 3}, \infty)$. If either model finds results before the cutoff date c, the trial is removed. If neither model finds any results at all, the trial is also removed, since the initial screening pass may have been a false positive.

Brave Mode Search. When *mode* is Brave, as described in the main text, the Brave Search API returns candidate URLs, webpages are fetched with Python Requests or ZenRows, and downloadable files are converted to text with DeepSeek OCR. We use three query strategies derived from the trial's information info $= \{$NCT ID, aliases, title, description$\}$: (1) the trial's brief title, (2) its NCT ID, and (3) an LLM-rewritten query generated by GPT-5 that incorporates identifiers, tags, eligibility criteria, and other metadata from the trial protocol. The prompt used for query rewriting is shown in Figure 7. Each query type is run both with and without a freshness date filter, where the freshness parameter restricts results based on the trial's study start date and the earliest available completion or results submission date. For each query configuration, we fetch up to 100 results, and all searches are executed in parallel using multithreading. The raw search results from all query and freshness combinations are then collected. For each returned URL, we determine its type to select the appropriate content retrieval method. If the URL points to a known paper repository (e.g., PubMed, PMC, or a preprint server), we use the corresponding database API to retrieve the paper's full text and publication date directly. If the URL points to a downloadable PDF, we download the file and convert it to text using DeepSeek-OCR2, deployed on our local H100 server. For all other URLs, we use the Python Requests library along with the ZenRows API to scrape the full webpage content. The scraped and converted texts are then stored in a database for downstream processing and verification.

Database Mode Search. When *mode* is Database, we use the trial's NCT ID to search PubMed, PMC, BioRxiv, and MedRxiv for matching articles, as described in the main text.

Text Extraction from Scraped Webpages. For Brave mode and Database mode searches, all retrieved texts, including webpages, downloadable files, and database articles, undergo our customized processing pipeline. As noted in the main text, scraped webpages are processed with a customized text extraction system derived from trafilatura, designed for long pages and for isolating the main article content while removing HTML, sidebars, and other articles' text. We provide further details on this system below.

We parse each page's raw HTML in parallel using multiprocessing. We first remove all scripting, styling, and non-rendered elements, then perform a depth-first traversal of the DOM tree, converting the HTML into a flat list of text and path tuples while preserving structural semantics such as headings, paragraphs, and lists in Markdown format. We then apply a series of domain-specific and general-purpose filters to remove irrelevant content. For ResearchGate pages, we truncate everything after the Citations or References section to remove recommended papers. For ichgcp.net, we remove "similar clinical trials" sections,

and for other domains we detect and strip "similar trials" or "similar articles" blocks. We also remove PubMed tail sections such as conflict of interest statements and comment blocks, site-specific footers (e.g., Mayo Clinic), and newsletter signup sections that appear after references. A broad marketing slogan filter detects over 50 common promotional patterns such as "You Might Also Like," "Related Posts," "Recommended Reading," "Subscribe," and similar calls to action, removing them only when they appear in the latter half of the page to avoid false positives on actual content. For WithPower pages, we strip "Other People Viewed" sections, and for Academia pages, we remove "Related papers" blocks. Finally, we detect reference, bibliography, and conflict of interest sections using extensive pattern matching that handles numerous edge cases where section headers may be split across multiple HTML elements (e.g., "R" + "eferences", "Re" + "ferences", or even character by character), as well as non-English variants. When exactly one such section boundary is found, we truncate the content there; when multiple candidates exist, we flag them for further disambiguation.

Publication Date Extraction. As described in the main text, for Brave mode and Database mode searches, all retrieved texts undergo our customized pipeline for publication date extraction to verify that each document falls within the required date range $[start, end]$. We run a multi-strategy prioritized pipeline in parallel across all scraped pages to determine each webpage's publication date as accurately as possible. We first skip known problematic domains and pages that failed scraping. For preprint servers such as bioRxiv and medRxiv, we use a specialized parser that extracts the posting date from their unique HTML layout. For YouTube pages, we parse the embedded JavaScript data object to retrieve the video's publication date. For all other pages, we apply a cascading series of extraction strategies in priority order: (1) structured data embedded in the page's JSON-LD blocks, (2) high-priority Open Graph and article metadata tags, (3) semantic publish date attributes on `time` and other elements, (4) domain-specific handlers such as definition list pairs for ScienceDaily and dedicated publish date containers, (5) citation metadata commonly used by academic publishers, (6) dates embedded within inline JavaScript variables, (7) semantic date spans with consistency checks for uniqueness, (8) the earliest date found across all time-related elements on the page, and (9) as a last resort, visible text scanning within the main content area for date patterns near keywords like "Published," "Posted," or "Updated." When multiple strategies yield dates, we take the earliest one to be conservative. All raw date strings are normalized through a robust converter that handles over a dozen formats including ISO 8601 datetimes, year-first and day-first numeric dates, two-digit year pivoting, compact date strings, month name variants in multiple orderings, year-plus-month only, and year only, with validation for month and day ranges including leap years. Dates with years before 1990 or after the current year are rejected as invalid.

Two-Round GPT-5 Verification. As described in the main text, for Brave mode and Database mode searches, after publication date filtering, all documents undergo two rounds of GPT-5 verification: first, to confirm that the document refers to the same clinical trial; and second, to confirm that it reports some form of results for that trial. We provide further details on each round below.

In Round 1 (trial matching), for each trial and document pair, we strip all date-related fields from the trial's protocol metadata to prevent the model from using temporal cues when making its judgment. We then ask GPT-5 to determine whether the document discusses the same exact clinical trial, or at least a subset of it. The prompt used for this step is shown in Figure 8. A key design choice is that we do not rely solely on NCT ID matching, since many documents, particularly news articles, press releases, and research papers, discuss clinical trials without ever mentioning a registry identifier. Instead, we instruct the model to systematically compare both sides across multiple dimensions: registry IDs and other study identifiers, intervention treatments including specific drugs and procedures, study design and masking techniques, sponsor and investigator organizations, trial phase, geographic locations, enrollment numbers, and participant eligibility criteria. By triangulating across all of these fields, the model can confidently identify a trial even when no NCT ID is present in the document, using the combination of matching drug names, sponsor organizations, study design, and enrollment characteristics as converging evidence. We enforce asymmetric

matching rules: certain fields such as identifiers, treatments, study design, and sponsors must match exactly or be a strict subset, while other fields such as enrollment and phase are allowed to differ if the document could plausibly be reporting on a subset or arm of the given trial. For example, a document reporting fewer treatments or a smaller enrollment is acceptable as a potential subset, but a document with additional or different drugs, a conflicting NCT ID, or a completely distinct set of sponsors is grounds for rejection.

In Round 2 (results verification), only pairs that the model identifies as matching in Round 1 proceed. We ask GPT-5 again with the same trial and document pair, this time to determine whether the document reports some form of results for that trial. The prompt used for this step is shown in Figure 9. We explicitly define "results" as any mention of outcomes such as success, failure, or signs of efficacy or safety, and instruct the model to reject cases where the document only reports results from related trials, studies, or experiments rather than from this exact trial, or where it merely discusses the launch or initiation of the trial without reporting outcomes. This two-round approach ensures that we first confirm identity between the document and the trial before checking for result contamination, reducing false positives from documents that discuss similar but distinct trials.

C.2 Answer Generation Pipeline Implementation Detail

Answer Verification for Endpoint/Superiority Questions. After automatic question generation, we perform an additional LLM-based verification step to ensure the correctness of each answer label against the reported results identified earlier. Specifically, we prompt `GPT-5.4-2026-03-05` with the trial title, the endpoint under evaluation, the relevant study arms, the two answer options, the selected answer, and an indexed list of reported results collected from online sources. The model is asked to judge whether the reported results support the provided answer, returning a binary `yes` or `no` decision. Cases flagged as unsupported are removed from the benchmark. The full answer verification prompt is provided in Figure 10.

Answer Verification for Comparative Effect Questions. For Comparative Effect questions, which involve a three-way classification between two study arms, we used a separate verification prompt. We provide `GPT-5.4-2026-03-05` with the clinical trial description, the specific endpoint under evaluation, the two study arms being compared, and the list of reported results. The model is asked to determine which of three options is correct: (a) the first arm is statistically significantly worse, (b) the second arm is statistically significantly worse, or (c) no statistical significant difference exists between the two arms. Because all three options hinge on statistical significance, the prompt includes explicit guidance on how to interpret significance from reported results. For example, a p-value exceeding 0.05, an odds ratio whose 95% confidence interval crosses 1, or overlapping confidence intervals between arms all indicate non-significance. Conversely, if a result explicitly states that the outcome measure was met or provides a direct statement about significance, this is taken as sufficient evidence. Results that contains only numerical values without any indication of significance are treated as inconclusive. If the model cannot select an correct option and outputs `None`, the corresponding question is discarded. The full verification prompt is provided in Figure 11.

Model	Endpoint		Superiority		Comparative Effect	
	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc
Winter 2025						
Traditional Baselines						
Random Forest	66.60	71.54	55.39	63.76	**80.70**	**79.12**
Feed-Forward NN	67.74	72.29	54.23	57.91	58.87	74.60
KNN + Random Forest	65.49	70.78	54.87	63.55	75.45	72.87
L2 Logistic Regression	62.38	70.56	54.46	58.58	65.01	67.55
HINT	57.28	64.72	51.87	62.33	46.08	50.00
AdaBoost	67.55	70.24	54.85	58.41	50.45	64.10
XGBoost	65.52	74.89	58.29	63.47	69.47	66.62
Large Language Models						
o3-mini	68.40	70.23	68.09	68.33	51.91	51.53
+ RAG	**76.88** ↑↑	73.48 ↑	**69.21** ↑	69.18 ↑	56.81 ↑	57.14 ↑
+ Agent	64.83 ↓	66.67 ↓	67.46 ↓	68.06 ↓	54.74 ↑	55.72 ↑
GPT-5	65.29	77.63	66.17	71.20	54.25	54.70
+ RAG	67.98 ↑	78.03 ↑	67.54 ↑	**71.86** ↑	56.96 ↑	57.49 ↑
Claude Opus 4.5†	70.17	80.23	62.31	68.91	55.52	54.83
+ RAG	70.36 ↑	**81.85** ↑	63.43 ↑	69.62 ↑	56.00 ↑	55.19 ↑
Open-Source Large Language Models						
Qwen2.5-14B	53.29	53.68	60.04	60.21	58.00	62.94
+ RAG	68.77 ↑	66.27 ↑	62.67 ↑	62.34 ↑	60.32 ↑	60.64 ↓
+ Agent	61.41 ↑	55.63 ↑	59.74 ↓	61.53 ↑	34.20 ↓	33.78 ↓
DeepSeek-R1-14B	53.54	54.18	56.68	58.02	45.69	55.05
+ RAG	67.56 ↑	64.21 ↑	61.26 ↑	61.09 ↑	32.30 ↓	41.31 ↓
+ Agent	66.23 ↑	64.07 ↑	61.21 ↑	61.77 ↑	51.21 ↑	56.78 ↑
Summer 2025						
Traditional Baselines						
Random Forest	**65.57**	66.48	55.88	61.63	46.15	45.65
Feed-Forward NN	65.25	**67.78**	64.95	67.58	57.56	85.87
L2 Logistic Regression	60.62	67.20	70.92	73.10	43.46	49.28
KNN + Random Forest	62.13	63.29	57.56	62.88	46.74	46.74
HINT	64.08	66.86	56.28	63.89	46.15	45.65
AdaBoost	60.40	61.05	58.15	61.03	39.51	34.78
XGBoost	61.53	63.27	65.26	69.79	45.56	44.57
Large Language Models						
o3-mini	59.83	59.46	72.48	71.93	54.94	59.78
+ RAG	59.07 ↓	58.98 ↓	73.15 ↑	72.94 ↑	56.75 ↑	60.87 ↑
+ Agent	61.75 ↑	62.84 ↑	**73.33** ↑	73.96 ↑	**69.81** ↑↑	**92.75** ↑↑
GPT-5	51.54	58.24	70.23	73.71	64.81	64.49
+ RAG	50.13 ↓	55.55 ↓	70.18 ↓	73.78 ↑	68.95 ↑	70.05 ↑
Claude Opus 4.5	53.71	57.86	68.76	73.68	48.24	49.64
+ RAG	58.69 ↑↑	62.72 ↑↑	69.59 ↑	**74.10** ↑	48.42 ↑	50.00 ↑
Open-Source Large Language Models						
Qwen2.5-14B	53.42	53.61	58.84	59.43	44.74	44.69
+ RAG	58.91 ↑	58.55 ↑	61.00 ↑	61.02 ↑	45.35 ↑	44.20 ↓
+ Agent	60.60 ↑	59.82 ↑	55.13 ↓	57.85 ↓	32.14 ↓	23.68 ↓
DeepSeek-R1-14B	52.04	52.25	57.95	59.65	41.39	51.21
+ RAG	58.32 ↑	57.83 ↑	61.48 ↑	61.79 ↑	32.92 ↓	47.71 ↓
+ Agent	57.31 ↑	56.90 ↑	57.59 ↓	58.33 ↓	29.63 ↓	21.05 ↓

† Opus 4.5 knowledge cutoff date postdates Winter 2025's timestamp, possibly contaminated.

Table 6: Acc is balanced accuracy. Compared to base model, ↑and ↓indicate nominal increases and decreases, while ↑↑denotes statistically significant improvements (95% CI via hierarchical bootstrap sampling). Best and second-best results within each dataset and column are marked in **bold** and underline, respectively. All LLM results are averaged over three runs.

Feature Name	Description & Examples	Type
Trial Overview		
Brief title	Short official title of the clinical trial.	Textual
Study type	Trial study type. *(We only examine interventional studies in this paper)*	Categorical
Trial allocation	How trial participants are assigned to different study arms. *(We only examine randomized trials in this paper)*	Categorical
Intervention model	The structural design of the trial. *(EXAMPLES: parallel; crossover; factorial; sequential; single group)*	Categorical
Primary purpose	The main intent of the trial. *(EXAMPLES: treatment; prevention; device feasibility, diagnostic)*	Categorical
Trial masking	The blinding scheme used in the trial, indicating who does or does not know the treatment assignment. *(EXAMPLES: single; double; none)*	Categorical
Enrollment count	The total number of participants enrolled or planned for the study.	Numerical (discrete)
Protocol arms count	The number of study arms defined in the protocol.	Numerical (discrete)
Protocol interventions count	The number of distinct interventions being studied across the trial.	Numerical (discrete)
Clinical Features		
Trial conditions	The diseases, disorders, or other medical conditions being targeted in the study.	Textual
Intervention "MeSH" terms	Standardized medical subject headings describing the interventions, including drug therapies, surgeries, and physical treatments.	Textual
Patient Information		
Eligibility criteria	The inclusion and exclusion criteria that determine who can or cannot participate in the study.	Textual
Sex	Which sex groups are eligible for enrollment. *(EXAMPLES: male; female; all)*	Categorical
Minimum age	The youngest allowed participant age (converted to years).	Numerical (discrete)
Maximum age	The oldest allowed participant age (converted to years).	Numerical (discrete)
Healthy volunteers	Whether people without the target disease are allowed to participate.	Categorical (binary)
Sponsors & Sites		
Lead sponsor name	The main organization responsible for leading or funding the trial. *(EXAMPLES: Eli Lilly and Company; AbbVie; Merck Sharp & Dohme LLC; Novartis Pharmaceuticals)*	Categorical
Lead sponsor class:	The high-level category of the lead sponsor. *(EXAMPLES: industry; network; FED; NIH)*	Categorical
Collaborator classes	The categories of any collaborating organizations involved in the study. *(EXAMPLES: FED / industry; NIH / industry)*	Categorical (multi)
Locations	The country or countries where the trial is being conducted. *(EXAMPLES: United States; China; multiple_countries)*	Categorical
Endpoint & Study Arms		
Trial protocol arms	A structured description of the study arms, including which groups receive which treatments.	Textual
Endpoint details	A detailed description of the endpoint, explaining exactly what outcome is being evaluated.	Textual
Endpoint time frame	When the endpoint is assessed (converted to number of days).	Numerical (discrete)
Focal arm	Description of the main study arm whose outcome is being evaluated in the comparison.	Textual
Comparator arm	Description of the reference study arm against which the focal arm is compared.	Textual

Table 7: Feature schema with description, examples and feature types.

Feature Group	Type	Logistic Regression	Random Forest
Endpoint details	Textual	0.0423	0.0247
Comparator arm	Textual	0.0381	0.0104
Trial conditions	Textual	0.0055	0.0223
Locations	Categorical	0.0165	—
Eligibility criteria	Textual	0.0044	0.0128
Lead sponsor class	Categorical	0.0119	0.0025
Brief title	Textual	0.0023	0.0115
Trial protocol arms	Textual	—	0.0086
Intervention MeSH terms	Textual	0.0017	0.0067
Trial lead sponsor name	Categorical	0.0040	—
Trial maximum age (years)	Numerical	0.0040	—
Focal arm	Textual	—	0.0034
Enrollment count	Numerical	—	0.0013

Table 8: Grouped permutation importance, reported as mean drop in macro-F1 when a feature group is permuted. Rows are the union of the two models' top-10 groups, sorted by the larger of the two scores. "—" indicates the group falls outside that model's top 10.

Model	Hyperparameter	End. / Sup. / Comp.
Logistic Regression (L2) (Lo et al., 2019)	penalty	L2
	C (inverse regularization strength)	1.0
	solver	lbfgs
	max_iter	3000
Random Forests (RF) (Lo et al., 2019)	n_estimators	300 / 200 / 200
	max_depth	None
	min_samples_leaf	6 / 4 / 4
KNN+RF (Lo et al., 2019)	knn_neighbors	5
	knn_weights	uniform
	n_estimators	200 / 100 / 100
	max_depth	None
	min_samples_leaf	4
FFNN (Tranchevent et al., 2019)	hidden_layer_sizes	(512, 128)
	activation	ReLU
	α	1×10^{-4}
	learning_rate_init	1×10^{-3}
	max_iter	200
AdaBoost (Fan et al., 2020)	n_estimators	300 / 300 / 200
	learning_rate	1.0
	base_max_depth	2
	base_min_samples_leaf	2
XGBoost (Siah et al., 2021)	n_estimators	300
	learning_rate	0.1 / 1.0 / 1.0
	max_depth	12

Table 9: Hyperparameters for each baseline. Where a single value is listed, the same setting is used across all three task types (Endpoint / Superiority / Comparative Effect). Otherwise, task-specific values are shown separated by "/".

D Dataset Generation Prompts

This section collects all prompt templates used throughout the CT Open pipeline, including those for decontamination search, trial matching, results verification, and answer generation.

```
You are given the NCTID of a clinical trial. You are also given a clinical trial's title, description, and outcome measures. Search the internet to find any mention of the clinical trial's results (it can be interim results or any indication of results on any of its outcome measures) from any webpage such as news, company statements, papers (e.g. pubmed, pmc, bioRxiv, medRxiv, clinicaltrials.gov), blogs and online forums. When you find one webpage that mentions this exact clinical trial's NCTID and mentions at least some interim results of this exact clinical trial, stop the search all together. If you find such a webpage, output "yes" after the keyword "OUTPUT:"

Note, if you still cannot find any webpage with any results (good or bad) for this specific clinical trial, output "no" after the keyword "OUTPUT:"

NCTID:
{nctid}

Clinical trial description:
{ct_description}

Clinical trial outcome measures:
{ct_outcome_measures}
```

Figure 3: Prompt for the initial screening pass, used to determine whether any results exist on the web for a given trial.

```
Find me results of this exact same Clinical Trial prior to this date: {cutoff_date}.

Any mention of interim outcomes for the specific trial—such as success, failure, or signs of efficacy/safety—counts as reporting results for that trial.

However, if a webpage only reports results from related trials, studies, papers, or experiments (and not from this exact trial), then it is not reporting results for the specific trial.

Give me all the urls you found. Note you should only output urls that are reporting results for the specific trial, not urls that are related to the trial or a earlier phase of the trial.

Output format (must follow exactly)
- Start a new line with LIST: then a single valid Python list of strings.
- No extra commentary before or after the list.
- Example:
  LIST: [url1, url2, url3, url4, url5, ...]
```

Figure 4: Prompt for LLM web search before the cutoff date.

```
Find me results of this exact same Clinical Trial on or after this date: {cutoff_date}.

Any mention of interim outcomes for the specific trial—such as success, failure, or signs of efficacy/safety—counts as reporting results for that trial.

However, if a webpage only reports results from related trials, studies, papers, or experiments (and not from this exact trial), then it is not reporting results for the specific trial.

Give me all the urls you found. Note you should only output urls that are reporting results for the specific trial, not urls that are related to the trial or a earlier phase of the trial.

Output format (must follow exactly)
- Start a new line with LIST: then a single valid Python list of strings.
- No extra commentary before or after the list.
- Example:
  LIST: [url1, url2, url3, url4, url5, ...]
```

Figure 5: Prompt for LLM web search after the cutoff date.

```
Task: Given some publicly available webpages that report outcomes for the exact clinical trial described below, extract the date and the result of the
webpages for the exact clinical trial.

What counts as "results": Any explicit reporting of outcomes (e.g., pCR, MFS, investigator/BICR assessments, efficacy signals, safety findings, success/failure
of interim analyses) for this exact trial.

Dates: Extract the page's publication date. If multiple dates appear, use the one that reflects when the results were first posted. If you cannot determine a
date, write "unknown".

Result content: Under result:, add concise bullet points (ideally 1-3) summarizing the evidence of outcomes reported on that page. Prefer specific numbers
(e.g., pCR %, MFS hazard ratios), time frames, and clear efficacy/safety statements. Do not include commentary beyond what the page states.

Output format (must follow exactly; no extra text before or after):

url1: <url1>
date: <YYYY-MM-DD or unknown>
result:

<data point 1 demonstrating results>

<data point 2 demonstrating results>


url2: <url2>
date: <YYYY-MM-DD or unknown>
result:

<data point 1 demonstrating results>

<data point 2 demonstrating results>

(Repeat for all qualifying URLs.)

Trial to match (must be this exact study):
{trial}

URLs to process:
{urls}
```

Figure 6: Prompt for extracting publication dates and result summaries from URLs discovered by LLM web search.

```
Given some information about a clinical trial, output a one-sentence single search query without logical operators that searches for relevant outcomes on
the web for the given clinical trial after the keyword "OUTPUT:"

Clinical trial information:
{clinical_trial_info}
```

Figure 7: Prompt for rewriting trial information into search queries for the Brave Search API.

Given the description of a clinical trial and a scraped webpage, (if the webpage is a research paper, ignore any scraped part that is in the reference section or similar articles or external links, because they are not the main contents of the webpage, and should not be considered by you), decide if the webpage contains information about the same exact clinical trial, or at least a part of the same exact clinical trial.

If the webpage does not reference any clinical trial, then you should output "no". Note, if a webpage contains case reports of patients, these cases could be part of a clinical trial. Even if the webpage does not explicitly mention a clinical trial but does mention case reports, this is not evidence for you to output "no".

The webpage may reference multiple clinical trials, with different level of details provided. However, no matter how detailed each clinical trial is described, you must check all of them carefully against the given clinical trial, instead of purely focusing on the main clinical trial mentioned on the webpage.

Note, if the webpage discusses a similar clinical trial and for this trial it explicitly gives a different NCT ID from the given clinical trial, this is not the same trial as the given trial and this cannot be a part of the given trial, therefore you should output "no". If the webpage references the clinical trial using a different study identifier that is not NCT ID, and if the given clinical trial also has another identifier that has the same format as the webpage's trial's study identifier, and if they exactly match, output "yes". Otherwise, output "no". For example, if the webpage's trial has an identifier that's SPX-017 and the given trial has an identifier that's SPX-011, then they are different trials and you should output "no". If the webpage's trial has an identifier that's ABC-017, and the given trial has an identifier that's SPX-011, because they do not have the same format, clearly they don't use the same identification convention. Then you cannot compare them. This is neither evidence for you to output "yes" nor "no".

For both the webpage's trial, and the given clinical trial, identify and output the following information for both, NCT IDs (if available), other study identifiers (for example SPX-017, if available), treatment/intervention (e.g., drugs and procedures), study design, sponsor/investigator/collaborator/affiliated parties/organization, phase, location, enrollment (or the number of patients reported). Below is a detailed guideline of how you should use this information for both trials to determine if the webpage's trial is the same as (or a subset of) the given trial, or if they are different trials.

If the webpage discusses a similar trial, but the intervention treatment is not the same (e.g. different combination of drugs, different dosage, different delivery method, different control etc.) as any arm or group of the given clinical trial, then the trial cannot be the same as the given trial nor can it be a subset of the given trial, therefore you should output "no".

Note it is acceptable if the webpage's trial has fewer intervention treatments than the given trial, because the webpage's trial could be a subset or a few arms of the given trial. Thus, if the webpage's trial has fewer treatments, it is not evidence for you to output "no". Importantly, however, if the webpage's trial has more treatment, e.g., more drugs or more procedures, or has slightly different treatment, e.g., some different drugs or procedures, then the webpage's trial has no chance of being a subset of a few arms of the given trial. Thus, if the webpage's trial has more or different treatments than the given trial, this is evidence for you to output "no".

If the webpage's trial uses a different study design, such as a different masking/blinding technique, than the given clinical trial, you should output "no".

Importantly, you must immediately identify investigator, sponsor, collaborator, affiliated parties, organization information from both the webpage's trial and the given trial, when available. If the webpage's trial has a completely distinct set of sponsor/investigator/collaborator/affiliated parties/organization from the set of information for the given trial, and you make sure there is no way this set of information from the webpage's trial overlaps with the set of information from the given trial, then you should output "no".

You should output "no" if any of the two following situations happens: The webpage discusses a similar trial but the location or phase information is drastically different from the given clinical trial, which means there is no way the webpage's discussed trial is the same as or a part of the given clinical trial. For example, if the given trial is phase 2 and phase 3 (i.e. the given trial has two components, one for phase 2 and the other for phase 3), but the webpage's trial is phase 1. For another example, if the given trial only has a location in San Francisco, but the webpage's trial has multiple locations, then there is no way the webpage's trial is a part of the given trial. Location information is usually city specific, therefore, two different cities in the same country does not mean the same location. You should output "no" if any of the above two situations happens.

However, different sources can have slightly different definitions of phase information. For instance, if the given trial is phase 2/3, and the webpage's trial is explicitly mentioned to be phase 3 or phase 2a or phase 2b, it is possible they have slightly different definitions of phase, thus this is not evidence that they are not the same trial. If the given trial's location is multi-location, and the webpage's trial only reports a single location, it is possible the webpage's trial is a subset of the given trial. Therefore, you should not use these as evidence to output "no".

If the webpage's trial has a much larger enrollment number than the given trial, then there is no way the webpage is a subset or the same as the given trial, therefore you should output "no".

However, if the webpage's trial has a much lower enrollment number than the given trial, and you find the given trial has multiple cohorts or groups and some cohorts have approximately the same number of participants as the webpage trial's enrollment number, then it is possible the webpage is discussing a group or subset of the given trial. You should not use this as evidence to output "no".

If the webpage's trial uses very different participants eligibility such as inclusion/exclusion criteria, or if you decide that at least one patient in the webpage's trial does not explicitly meet the inclusion criteria of the given clinical trial, or if you find that at least one patient in the webpage's trial meet the exclusion criteria of the given trial, this means there is no way this trial is the same as or a part of the given trial, then you should output "no".

Overall, the guiding principle is: there is certain information such as enrollment number, phase information that can be different but it is still possible the webpage's trial is a subset or the same as the given trial. However, there is other information, such as study identifier, treatment, drug, study design, sponsor/investigator/collaborator/affiliated parties/organization, and eligibility criteria that cannot be different.

To repeat the previous instruction, for both the webpage's trial, and the given clinical trial, identify and output the following information for both, NCT IDs (if available), other study identifiers (for example SPX-017, if available), treatment/intervention (e.g., drugs and procedures), study design, sponsor/investigator/collaborator/affiliated parties/organization, phase, location, enrollment (or the number of patients reported). Above is the aforementioned detailed guideline of how you should use this information for both trials to determine if the webpage's trial is the same as (or a subset of) the given trial, or if they are different trials.

After you have carefully compared the webpage's trial (if there is any) and the given trial, you should have decided if they are the same trial or at least the webpage is reporting part of the given trial, if you decide yes, you should output "yes". Otherwise, directly output "no".

Clinical trial description:
{clinical_trial_description}

Scraped Webpage:
{scraped_webpage}

Output either "yes" or "no" after the keyword "OUTPUT:"

Figure 8: Prompt for Round 1 of GPT-5 verification, used to determine whether a document discusses the same clinical trial.

You are given a webpage that contains some mention of a specific clinical trial. You are also given details of this specific clinical trial. Your task is to determine if the webpage's trial contains any indication of results of this specific given trial. Any mention of interim results for this specific given trial such as success, failure, or signs of success, is indication of results. Note, if the webpage only reports results for related trials/studies/papers/experiments, but not this specific given trial, then the webpage is only reporting results relevant to this specific given trial, which means the webpage is not reporting results from this specific given trial, therefore you should output "no". In other words, if all results in the webpage are not generated from this specific given trial, then you should output "no".

If the webpage does not contain any indication of the result for this specific given trial, for example, it only talks about the launching and release of this specific clinical trial, then you should still output "no".

If in the end, you have decided that the webpage's trial contains some indication of results about this specific given trial, then, and only then, you should output "yes".

Output either "yes" or "no" after the keyword "OUTPUT:"
Then output all indication of results for this specific clinical trial after the keyword "RESULTS:"

Clinical trial description:
{clinical_trial_description}

Scraped Webpage:
{scraped_webpage}

Figure 9: Prompt for Round 2 of GPT-5 verification, used to determine whether a document reports results for the matched trial.

Given a clinical trial's descriptions, endpoint, a few listed study arms, a question that contain two options, and an answer that indicate which option is the correct answer, and an indexed list of reported results from various url sourced, determine if these reported results support the provided answer being correct. Output either "yes" or "no" after the keyword "SUPPORT:"
If you answer yes, then examine all indexed list of reported results and find a comprehensive list of indices whose reported results support the provided answer being correct. You do not have to include all indices, because it is possible some indices' reported results are not relevant to support the provided answer being correct. Output this list of indices as a python list, e.g., [0, 3, ...], after the keyword "LIST:"
If you answer no the prior question, then output None after the keyword "LIST:"

Trial Title:
{title}

Endpoint:
{endpoint}

Examine the following Study Arms:
{List_Study_Arms}

Options:
{option}

Answer:
{answer}

Reported Result:
{report_result}

Figure 10: Prompt for answer verification for Endpoint/Superiority Questions.

You are given descriptions of a clinical trial, and a specific proposed outcome measure (endpoint). You will also see a question with three options based on two study arms in the clinical trial. You will also see a list of reported results. Based on this list of reported results, determine if you can select option a or option b or option c as the correct answer to the question, output either "a" or "b" or "c" after the keyword "ANSWER:". If you cannot select either option a or option b or option c, then output "none" after the keyword "ANSWER:"

Note, because the three options all mention statistical significance, when you decide which option to choose or output "none", you must decide whether the reported results conform to our definition of statistical significance.

For your reference, Statistical significance (or insignificance) can be expressed in several ways, including but not limited to: a p-value greater than 0.05 indicating no significance; an odds ratio whose 95% confidence interval straddles 1, indicating no significance; or overlapping confidence intervals between the comparator and experimental arms, indicating no significance, etc. If a result contains only a numerical value without any indication of significance, it cannot be classified as either statistically significant or statistically insignificant. If a result explicitly states that the outcome measure was met, including but not limited to: through a direct statement about this specific outcome measure, or a direct statement about statistical significance or no significance, or a statement that all secondary outcome measures were met (and this specific outcome measure is a secondary outcome measure), etc., then statistical significance (or insignificance) is implied.

Afterwards, output your reasoning after the keyword "REASON:"

Clinical Trial Descriptions:
{trial_description}

Specific Proposed Endpoint:
{endpoint}

Two Study Arms to be Examined:
{study_arms}

Question with three Options:
a) Study Arm {arm_b} is statistically significantly worse than Study Arm {arm_a}.
b) Study Arm {arm_a} is statistically significantly worse than Study Arm {arm_b}.
c) No statistically significant difference between Study Arm {arm_a} and Study Arm {arm_b}.

List of Reported Results:
{reported_results}

Figure 11: Prompt for answer verification for Comparative Effect Questions.

E Evaluation Prompts

This section describes the prompts used to evaluate different models. Baseline evaluation prompt for prompt-only LLMs is provided in Figure 12. RAG evaluation prompts for datapoints with historical similar trials, historical me-too trials, and without historical similar trials are provided in Figures 13, 14 and 15. Protocol used in agent pipeline are in Figure 16.

```
You are an expert in assessing biological and medical clinical trials. You are given a clinical trial's
description and meta information, drugs' biological mechanism, patients eligibility criteria, study arms
information, and a specific proposed outcome measure. Based on these information and your expertise
in assessing clinical trials, predict the results of this clinical trial, and answer the following question. An
endpoint can be met for a specific study arm, or an endpoint is met in general for this trial. We consider
an endpoint is met if statistical significant (95% confidence level) improvements over control is observed.
The question has two or three options, assign probabilities for these two or three options, make sure
these two or three probabilities sum up to 1, and select the option with the highest probability.

If there are two options, output the two probabilities respectively, in a tuple, (p, 1-p), after the keyword
"DECISION:"
If there are three options, output the three probabilities respectively, in a triple, (p, q, 1-p-q), after the
keyword "DECISION:"

Afterwards, output your reasoning after the keyword "REASON:"

Trial Title and Description:
{trial_title_description}

Trial Eligibility Criteria:
{eligibility}

Trial Specific Proposed Outcome Measure (Endpoint):
{endpoint}

{arms_information}

Question for you to consider and assign probabilities for both options:
{question}
```

Figure 12: Prompt for Prompt-only LLM Evaluation.

F Question Classes, Label Distribution, and Example Questions

Answer Options. The meaning of each answer option depends on the question class. Throughout, "statistically significant" refers to the 95% confidence level.

For the two binary classes, Endpoint and Superiority, Option A is the positive outcome and Option B is the negative outcome. For Superiority, Option A indicates that the treatment arm achieved a statistically significant improvement over the comparator arm, and Option B indicates that it did not. For Endpoint, Option A indicates that the endpoint was met, or, in the "at least one arm" subtype, that at least one arm met it; Option B indicates that it was not met.

Comparative Effect is a three-way class comparing the treatment (experimental) arm with the comparator arm, which may be a placebo or an active comparator. Option A indicates that the comparator arm is statistically significantly worse than the treatment arm. Option B indicates that the treatment arm is statistically significantly worse than the comparator arm. Option C indicates that there is no statistically significant difference between the two arms. We include this class because some trials are designed to establish non-inferiority of the experimental treatment relative to the comparator, which is common when the outcome measure concerns safety, tolerability, or the rate of side effects. In these cases a three-way distinction carries more information than a binary one.

Table 10 reports the label distribution within each class and Table 11 provides one example question per class.

Question Class	**Label**	**Winter 2025** ($n = 605$)	**Summer 2025** ($n = 857$)
Endpoint ($n = 80$ / 146)	Option A	66 (82.5%)	109 (74.7%)
	Option B	14 (17.5%)	37 (25.3%)
Superiority ($n = 470$ / 662)	Option A	299 (63.6%)	400 (60.4%)
	Option B	171 (36.4%)	262 (39.6%)
Comparative Effect ($n = 55$ / 49)	Option A	6 (10.9%)	2 (4.1%)
	Option B	2 (3.6%)	1 (2.0%)
	Option C	47 (85.5%)	46 (93.9%)

Table 10: Label distribution by question class. Percentages are within class and split. Class sizes are listed as Winter 2025 / Summer 2025 and sum to 605 and 857 respectively. Option B contains 2 and 1 instances in the Comparative Effect class, so per-class scores on that class are sensitive to individual predictions.

Table 11: Example questions from the benchmark dataset.

	Superiority Example	**Comparative Effect Example**	**Endpoint Example**
NCT ID	NCT03335839	NCT06177912	NCT05321082
Outcome Measure	Post-procedural MBG 0/1 or Distal Embolization	Percentage of participants with solicited systemic AEs	Time to first occurrence of composite endpoint: acute ILD exacerbation, hospitalization for respiratory cause, or death
Outcome Type	Primary	Primary	Secondary
Time Frame	30 days	Up to 5 days	Up to 31 months
Study Arms	*Arm 1:* Intracoronary tPA 10 mg (Experimental) *Arm 2:* Placebo (Placebo Comparator)	*Arm 1:* V116 – Pneumococcal 21-valent conjugate vaccine (Experimental) *Arm 2:* PPSV23 – Pneumococcal 23-valent conjugate vaccine (Active Comparator)	*Arm 1:* BI 1015550 low dose / Nerandomilast (Experimental) *Arm 2:* BI 1015550 high dose / Nerandomilast (Experimental)
Question	(a) Intracoronary tPA 10 mg achieved statistically significant improvements over Placebo. (b) Intracoronary tPA 10 mg has *not* achieved statistically significant improvements over Placebo.	(a) PPSV23 is statistically significantly worse than V116. (b) V116 is statistically significantly worse than PPSV23. (c) No statistically significant difference between V116 and PPSV23.	(a) At least one arm met this endpoint. (b) No arm met this endpoint.
Answer	**(b)**	**(c)**	**(b)**

You are an expert in assessing biological and medical clinical trials. You are given a clinical trial's description and meta information, drugs' biological mechanism, patients eligibility criteria, study arms information, and a specific proposed outcome measure. Based on these information and your expertise in assessing clinical trials, predict the results of this clinical trial, and answer the following question. An endpoint can be met for a specific study arm, or an endpoint is met in general for this trial. We consider an endpoint is met if statistical significant (95% confidence level) improvements over control is observed.

For your reference, you are also given a list of historical trials whose investigated drugs share similar biological mechanism as this given trial's drugs. This list of historical trials all have results on all of their endpoints. They are for references only, you do not have to consider them at all when making prediction for the current given trial.

The question has two or three options, you must assign probabilities for these two or three options, make sure these two or three probabilities sum up to 1.
If there are two options, output the two probabilities respectively, in a tuple, (p, 1-p), after the keyword "DECISION:"
If there are three options, output the three probabilities respectively, in a triple, (p, q, 1-p-q), after the keyword “DECISION:”
Afterwards, output a brief summary of your reasoning after the keyword "REASON:"

The current given trial's info:
{a}

The current given trial's question for you to predict:
{b}

The list of historical trials and their results for your reference:
{c}

Recall the instruction, You are an expert in assessing biological and medical clinical trials. You are given a clinical trial's description and meta information, drugs' biological mechanism, patients eligibility criteria, study arms information, and a specific proposed outcome measure. Based on these information and your expertise in assessing clinical trials, predict the results of this clinical trial, and answer the following question. An endpoint can be met for a specific study arm, or an endpoint is met in general for this trial. We consider an endpoint is met if statistical significant (95% confidence level) improvements over control is observed.

For your reference, you are also given a list of historical trials whose investigated drugs share similar biological mechanism as this given trial's drugs. This list of historical trials all have results on all of their endpoints. They are for references only, you do not have to consider them at all when making prediction for the current given trial.

The question has two or three options, you must assign probabilities for these two or three options, make sure these two or three probabilities sum up to 1.
If there are two options, output the two probabilities respectively, in a tuple, (p, 1-p), after the keyword "DECISION:"
If there are three options, output the three probabilities respectively, in a triple, (p, q, 1-p-q), after the keyword “DECISION:”
Afterwards, output a brief summary of your reasoning after the keyword "REASON:"

The current given trial's info:
{a}

The current given trial's question for you to predict:
{b}

Figure 13: Prompt for RAG Evaluation for datapoints with Matched Historical Similar Trials.

You are an expert in assessing biological and medical clinical trials. You are given a clinical trial's description and meta information, drugs' biological mechanism, patients eligibility criteria, study arms information, and a specific proposed outcome measure. Based on these information and your expertise in assessing clinical trials, predict the results of this clinical trial, and answer the following question. An endpoint can be met for a specific study arm, or an endpoint is met in general for this trial. We consider an endpoint is met if statistical significant (95% confidence level) improvements over control is observed.

For your reference, you are also given a list of historical trials that are considered to be me-too trials for the current given trial. These historical trials have their actual results posted. When using these historical trials' results, you must also take into account their difference (w.r.t drugs, eligibility criteria, outcome measure, etc) to the current given trial. You must decide to what extent to use these historical trails results, and to what extent to ignore these historical trials results, depending on their differences to the current trial.

The question has two or three options, you must assign probabilities for these two or three options, make sure these two or three probabilities sum up to 1.
If there are two options, output the two probabilities respectively, in a tuple, (p, 1-p), after the keyword "DECISION:"
If there are three options, output the three probabilities respectively, in a triple, (p, q, 1-p-q), after the keyword “DECISION:”

Afterwards, output a brief summary of your reasoning after the keyword "REASON:"

The current given trial's info:
{a}

The current given trial's question for you to predict:
{b}

The list of historical trials and their results for your reference:
{c}

Recall the instruction, You are an expert in assessing biological and medical clinical trials. You are given a clinical trial's description and meta information, drugs' biological mechanism, patients eligibility criteria, study arms information, and a specific proposed outcome measure. Based on these information and your expertise in assessing clinical trials, predict the results of this clinical trial, and answer the following question. An endpoint can be met for a specific study arm, or an endpoint is met in general for this trial. We consider an endpoint is met if statistical significant (95% confidence level) improvements over control is observed.

For your reference, you are also given a list of historical trials that are considered to be me-too trials for the current given trial. These historical trials have their actual results posted. When using these historical trials' results, you must also take into account their difference (w.r.t drugs, eligibility criteria, outcome measure, etc) to the current given trial. You must decide to what extent to use these historical trails results, and to what extent to ignore these historical trials results, depending on their differences to the current trial.

The question has two or three options, you must assign probabilities for these two or three options, make sure these two or three probabilities sum up to 1.
If there are two options, output the two probabilities respectively, in a tuple, (p, 1-p), after the keyword "DECISION:"
If there are three options, output the three probabilities respectively, in a triple, (p, q, 1-p-q), after the keyword “DECISION:”

Afterwards, output a brief summary of your reasoning after the keyword "REASON:"

The current given trial's info:
{a}

The current given trial's question for you to predict:
{b}

Figure 14: Prompt for RAG Evaluation for datapoints with Matched Historical Me-Too Trials.

You are an expert in assessing biological and medical clinical trials. You are given a clinical trial's description and meta information, drugs' biological mechanism, patients eligibility criteria, study arms information, and a specific proposed outcome measure. Based on these information and your expertise in assessing clinical trials, predict the results of this clinical trial, and answer the following question. An endpoint can be met for a specific study arm, or an endpoint is met in general for this trial. We consider an endpoint is met if statistical significant (95% confidence level) improvements over control is observed.

Note, this given trial does not have a historical me-too trial or a highly similar trial with similar endpoint. Therefore, to predict their results, it is important to consider the merit of the drug, maybe from prior research papers from your internal knowledge, and it is also important to consider the difficulty of the outcome measure, is it a tough outcome measure that cannot be easily met, is it a rare outcome measure that has not been used before, or is it a commonplace outcome measure? You must use your internal knowledge.

The question has two or three options, you must assign probabilities for these two or three options, make sure these two or three probabilities sum up to 1.
If there are two options, output the two probabilities respectively, in a tuple, (p, 1-p), after the keyword "DECISION:"
If there are three options, output the three probabilities respectively, in a triple, (p, q, 1-p-q), after the keyword “DECISION:”
Afterwards, output a brief summary of your reasoning after the keyword "REASON:"

The current given trial's info:
{a}

The current given trial's question for you to predict:
{b}

Figure 15: Prompt for RAG Evaluation for datapoints without Matched Historical Trials.

You are an AI agent (with expertise in assessing biological and medical clinical trials) and you are tasked with gathering the information necessary to predict the result of a clinical trial for a specified outcome measure, and then answering a question based on that prediction. You will be given descriptions of the clinical trial (title, summary, eligibility criteria), the specific outcome measure of interest, the study arms of interest, and the multiple choice question you must answer by choosing only one of the option.

Based on these information and your expertise in assessing clinical trials, predict the results of this clinical trial, and answer the following question. An endpoint can be met for a specific study arm, or an endpoint is met in general for this trial. We consider an endpoint is met if statistical significant (95% confidence level) improvements over control are observed.
The question has two or three options, assign probabilities for these two or three options, make sure these two or three probabilities sum up to 1, and select the option with the highest probability.

The setup is a multi-turn conversation with the environment. One turn is defined to be one environment input + your output after seeing the environment input plus the history of prior turns. In the first turn, the environment input is the protocol plus the clinical trial and its question that you must answer. In each subsequent turn, the environment input is a new input plus the protocol, the trial question and a summarization of prior turns of conversations (especially what you have learned in previous turns). At each turn, after seeing the input, you are allowed to take certain actions, and with each action, the environment will return you an input for the next turn of the conversation.

Control Flow and Logic:
Rule 1: You cannot answer the question before processing over {lower_bound} tokens. At each turn of the conversation, you will be reminded of the total amount of token you have seen in prior conversation plus the amount of the token you are seeing for this environment input. Again, you must not answer the question if the total prior token + current input token is less than {lower_bound} tokens.

Rule 2: You must answer the question after seeing {upper_bound} tokens. You must give the answer to the question by selecting one of its option if the total prior token + current input token is greater than {upper_bound} tokens.

Rule 3: You have access to search engine. At each of your turn, you are allowed to submit a one-sentence query that the environment will run it on a search engine. You can expect the environment input to be a list of urls (at most 20 urls), each with a short content snippet. You can submit any query you want, but the search engine will only return you urls published before the cutoff date, {cutoff_date}.

Rule 4: After you have seen the list of urls and their content snippets, you must decide to see at least one url's full webpage. The environment input will be the requested url's full webpage, along with the entire list of urls. Afterwards, you can decide to see another url's full webpage from the list. Then, the system will input this other url's full webpage. After you have run out of urls that you want to see, or you decide that you do not want to see any other urls from the list, you can submit a new query. IMPORTANTLY, you must request to see at least one url from the list recently retrieved for you. You CANNOT issue a new query without first requesting to see at least one url's full webpage. In other words, you cannot issue two queries in two consecutive turns. It must be query then url (and possibly more urls) then new query...
Finally, you must not request to see a url that you have previously seen before (possibly with a different query).

Rule 5: If at any point, you have exceeded the {upper_bound} token limit, it does not matter what is your current stage, you must answer the question in this current turn, following Rule 2. If you have exceeded the {lower_bound} token limit but not the {upper_bound} token limit, you can answer the question or see the full webpage of a url (in accordance to Rule 4) or submit a new query (in accordance to Rule 3). If you have not exceeded the {lower_bound} token limit, you are forbidden from answering the question (in accordance to Rule 1) and you must either see a url's full webpage or submit a new query.

Rule 5.1: At every turn, even the first turn, it does not matter what action you choose to take, you must give a multi-paragraphs long elaborate and detailed account of your current strategy for how to predict the results of this clinical trial and how to answer the question and which option do you think is most likely. Your strategy should be influenced by the nature of the trial, endpoint, study arms, question, and any inputs from the environment, and also what your previous strategies are and what you previously have learned from prior turns of conversations. You can refer to the **General Guidance** section for how to formulate a rudimentary strategy. Output your current strategy after the keyword "STRATEGY:"
Note, IMPORTANTLY, you must always output a strategy at each turn of the conversation.

Rule 5.2 At every turn, even the first turn, you must summarize what you have learned from the current environment input with regard to the question. Environment input can be a list of urls and content snippets, or it can be the full webpage of a url. From the current environment input, you must extract the potentially useful and relevant information that your query has requested (you can check your previous intention, explanation, strategy and query), or information that may help you predict the results of the clinical trial and answer the question. When you summarize, you must faithfully output any useful data, statistics, or information that you extracted. You must not omit these information you extracted. You summary should not include a summary of the given clinical trial in question, or the question or the protocol. Your summary must solely be about the current environment input, which is either a list of urls with content snippets or a full webpage.
After your extraction and summarization, output it after the keyword "SUMMARY:"
Note, IMPORTANTLY, you must always output a summary at each turn of the conversation.

Output Format:
To ensure accurate parsing of your output, you must strictly adhere to the following output formatting rules.

Rule 6: When you decide to answer the question provided you have exceeded seeing {lower_bound} token limit (in accordance to Rule 1 and Rule 2), note that the question has two or three options, assign probabilities for these two or three options, make sure these two or three probabilities sum up to 1, and select the option with the highest probability. If there are two options, output the two probabilities respectively for options "a" and "b", in a tuple, (p, 1-p), after the keyword "ANSWER:"
If there are three options, output the three probabilities respectively (i.e. p is the probability for option a, q is the probability for option b, and 1-p-q is the probability for option c), for option "a", "b" and "c", in a triple, (p, q, 1-p-q), after the keyword "ANSWER:"
Afterwards, you must output your reasoning to support this option, which must be a multiple-paragraphs long detailed and careful consideration and deliberation that takes into account all prior rounds of conversations, each round has a summary of your own findings, after the keyword "REASON:"

Rule 7: When you decide to submit a new query if you have not exceeded seeing {upper_bound} token limit (in accordance to Rule 2 and Rule 3), you must formulate the query after the keyword "QUERY:"
The query must be within one-sentence (for instance, the drug's codename, or a combination of search terms).

Rule 8: When you decide to see a the full webpage of a url if you have not exceeded seeing {upper_bound} token limit (in accordance to Rule 2 and Rule 4), you must make sure the url is from one of the list of urls that have previously returned to you triggered by your submitted queries. The url must be accurate, otherwise it cannot be parsed and full webpage content cannot be retrieved. You will only output the url after the keyword "URL:"

Rule 9: If you do not strictly follow the formatting Rule 6, Rule 7 and Rule 8, you will see an format error message which compels you to fix the formatting and correctly give the output.

Rule 10: If you do not strictly follow Rule 1, 2, 3, 4, 5 from the **Control Flow and Logic** Section, you will see a non-compliance error message that compels you to strictly follow the rules and correctly give the output.

Rule 11: Whatever action you decide to take, whether it is a new query, a url, or answer the question, you must provide an explanation for your action and your current thinking process. Specifically, when you issue a new query instead of checking on urls retrieved from your previous query, you must explain why you think it is better to issue a new query instead of looking into details and the full webpages of the list of urls retrieved from your previous query. Output after the keyword "EXPLAIN:"

General Guidance

Advice 1: Do not search for the target trial's unpublished result because the it is unavailable before the cutoff date. Instead, Search for relevant evidence that can predict the result.

Advice 2: If you do not know the drugs used in the trial, either because they are new drugs or because they use internal codenames designated by pharmaceutical companies, you should try to find out the background information about this drug including its biological mechanism, its intended effect, any potential concerns, etc.

Advice 3: Importantly, writing query is an important part of your job. After a few rounds of exploration, you should be able to quickly find out key terms and search phrases that are ineffective. When you write a new query, it should accurately reflect what you wish to search for and your current strategy. Your new query should not contain generic or ineffective search phrases, nor should it be a permutation, a rephrasing, or a rearrangement from previous queries. New queries should strive to explore previous unknown information.

Advice 4: Equally importantly, you must Search for disconfirming evidence and failure modes, not only supportive evidence. You must obtain a balanced view, listening to evidence from both proponents and opponents, in order to make an unbiased decision when you answer the question.

Advice 5: Prioritize evidence by directness to the target question: same intervention (drugs), same comparator, same population, same endpoint, same timing, and similar design.

Advice 6: Do not treat related endpoints as interchangeable. Benefit on one endpoint does not guarantee benefit on another.

Advice 7: Prefer quantitative evidence over generic summaries whenever possible.

Advice 8: Reason explicitly about which arm should do better or worse under plausible success and failure scenarios.

Advice 9: Down-weight evidence that differs materially in dose, schedule, population, endpoint definition, assay, background therapy, or follow-up.

Advice 10: Use eligibility criteria to find studies in comparable patients.

Advice 11: Distinguish biological plausibility from likely statistical success in the trial.

Advice 12: Reflect uncertainty honestly when the evidence is mixed.

Figure 16: Protocol used during the agent evaluation pipeline.