CTBench: A Comprehensive Benchmark for Evaluating Language Model Capabilities in Clinical Trial Design

Nafis Neehal
*Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
neehan@rpi.edu

Bowen Wang
Center of Biotechnology and Interdisciplinary Studies
Rensselaer Polytechnic Institute
Troy, NY 12180
wangb19@rpi.edu

Shayom Debopadhaya
Albany Medical College
Albany, NY 12208
debopas@amc.edu

Soham Dan
IBM Research
1101 Kitchawan Rd, NY 10598
soham.dan@ibm.com

Keerthiram Murugesan
IBM Research
1101 Kitchawan Rd, NY 10598
keerthiram.murugesan@ibm.com

Vibha Anand
Healthcare and Life Sciences, IBM Research
314 Main Street, Cambridge, MA
anand@us.ibm.com

Kristin P. Bennett
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY 12180
bennek@rpi.edu

Abstract

We introduce CTBench, a benchmark to assess language models (LMs) in aiding clinical study design. Given metadata specific to a study, CTBench examines how well AI models can determine the baseline features of the clinical trial (CT) which include demographic and relevant features collected at the start of the trial from all participants. The baseline features, typically presented in CT publications (often as Table 1), are crucial for characterizing study cohorts and validating results. Baseline features, including confounders and covariates, are also required for accurate treatment effect estimation in studies involving observational data. CTBench consists of two datasets: "CT-Repo", containing baseline features from $1,690$ clinical trials sourced from `clinicaltrials.gov`, and "CT-Pub", a subset of 100 trials with more comprehensive baseline features gathered from relevant publications. We develop two LM-based evaluation methods for evaluating the actual baseline feature lists against LM-generated responses. "ListMatch-LM" and "ListMatch-BERT" use GPT-4o and BERT scores (at various thresholds), respectively, to perform the evaluation. To establish baseline results, we apply

*Github Link - https://github.com/nafis-neehal/CTBench_LLM.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

advanced prompt engineering techniques using LLaMa3-70B-Instruct and GPT-4o in zero-shot and three-shot learning settings to generate potential baseline features. We validate the performance of GPT-4o as an evaluator through human-in-the-loop evaluations on the CT-Pub dataset, where clinical experts confirm matches between actual and LM-generated features. Our results highlight a promising direction with significant potential for improvement, positioning CTBench as a useful tool for advancing research on AI in CT design and potentially enhancing the efficacy and robustness of CTs.

1 Introduction

Medical research can be broadly categorized into clinical trials (CTs) and observational studies, among other types. CTs aim to test one or more interventions for the improvement of health outcomes, where human subjects are recruited and assigned prospectively to the interventions or respective placebo controls. In contrast, observational studies are where the causal effects of health outcomes are observed by the investigators without controlling the independent variables. Randomized CT remains the "gold standard" in evaluating the safety and efficacy of the intervention. At the same time, observational studies allow for much less expensive and larger-scale investigations using existing or prospective data [1–3]. In either case, it is crucial to ensure the balance between the study groups at the baseline, and that no systemic difference between study groups interferes with the causal relationship between the variables of interest and study outcomes [4]. Baseline characteristics, typically found in "Table 1" in CT publications, describe the demographic and relevant features collected at the beginning of the study for all participants between study groups. Depending on the study outcomes, the baseline characteristics may include sociodemographics, anthropometrics, confounding medical conditions, etc. For observational studies, the baseline features can help design the study by matching the cohort by the confounders and covariates. The showcase of baseline characteristics shows the reader how representative the study population is and how applicable the results would be. It validates the study design, increases the statistical efficiency, and helps the investigators draw logical conclusions [5–7].

Currently, general guidelines and considerations for the selection of baseline features exist [8]. However, most of the relevant features are study-specific and require the investigators' judgment. This may lead to an overlook of relevant confounders or covariates. Alternatively, for observational studies in particular, the improper selection of confounders/covariates from baseline features may lead to over-adjustment bias [9]. In addition, the reporting of baseline feature variables is not standardized and consistent across studies even for similar interventions or health outcomes. To tackle this issue in clinical research, we introduce CTBench, a benchmark to assess the role of language models (LMs) in aiding clinical study design. CTBench requires these models to predict the baseline characteristic variables of a clinical study based on the CT metadata. This study is the first to use LMs to solve the challenging task of designing the baseline features for both CTs and observational studies.

To achieve this, we create the benchmark from the centralized CT repository along with human annotation. We create two expansive datasets: 1) "CT-Pub" which includes the metadata and baseline features from 1,690 CTs collected from the `clinicaltrials.gov` API, and, 2) "CT-Repo" which contains a subset of 100 trials where the baseline features are retrieved from the related clinical publications via human curation.

The main contributions of this work include: 1) we propose a benchmark (CTBench) to use LMs to develop AI support tools for CT, assist researchers in selecting baseline features and design more efficient and robust clinical studies; 2) we create two CT metadata datasets with associated baseline features derived from a definitive repository and published papers; 3) we develop two automated evaluation methods for comparing predicted and actual trial baseline features, "ListMatch-LLM" and "ListMatch-BERT", and validate them with "human-in-the-loop" evaluations; and 4) we demonstrate CTBench by using robust prompt engineering techniques on several LLMs to generate the baseline

feature variables and evaluate their performance results. 5) Our data, code, and demo examples are available at https://github.com/nafis-neehal/CTBench_LLM.

2 Related Work

Recent applications of LLMs show that they can serve as powerful tools alongside human evaluators [10, 11]. They have been efficiently deployed for extracting clinical information with models such as the CT-BERT and MT-clinical BERT [12, 13]. CliniDigest showed a similar value, reducing 10,000-word CT descriptions into 200-word summaries using GPT 3.5 [14]. LLMs have been shown to have further uses in comparing similarity among trials to improve result comparison and aid in the precision design of subsequent studies [15]. Advances in prompting have additionally increased the use cases, both in specific medical specialties and generalized contexts [16–18]

Research exists on using LLMs to aid in creating eligibility criteria for CTs [19–22]. Critical2Query was validated on 10 CTs of different medical contexts to produce inclusion and exclusion criteria for the resolution of previous conditions, disease severity, and disease duration [19]. TrialGPT proposed an LLM that could potentially reduce 42.6% of the screen time needed to match CTs by domain experts without compromising in near-expert level grouping [20]. AutoCriteria similarly shows promising extraction of eligibility criteria through a set of 180 manually annotated trials [21].

However, automation of proposing baseline features of CTs is lacking. Since baseline features of CTs have become significantly more complex from 2011-2022 [23], better approaches for suggesting a generalizable and standardized set of cohort demographics and features are needed. Adequately training and validating LLMs for these clinical tasks requires relevant and feature-rich datasets. Several works have leveraged the `clinicaltrials.gov` database that has information for over 300,000 research studies conducted in more than 200 countries [12, 15, 16, 19, 21]. However, the prioritization of creating CT eligibility tools has left patient descriptor data relatively understudied.

CTBench addresses gaps between study criteria and features that are reported in databases such as `clinicaltrials.gov` in comparison to what appears in the final publication. For example, where age, sex, race, ethnicity, region of enrollment, and hemoglobin A1C may be reported on databases [24], investigators ensured that additional baseline characteristics of fasting serum glucose, duration of diabetes, BMI, weight, waist circumference, estimated GFR, albumin-to-creatinine ratio, medication use, and cardiovascular parameters were included in the final report [25]. As only 4 baseline features are consistently reported by greater than 10% of studies on these well-used databases, the development of publicly available and accurate baseline feature databases is necessary [26]. Current datasets that attempt to address this are limited by low CT cohort size or have sufficient patient data but are sourced from general clinical notes in place of CTs [27, 28]. Other projects do create datasets from high-quality, manually annotated CTs, but do not provide public access [21]. Here, our constructed datasets are relevant to baseline demographics (CT-Repo, CT-Pub), with human annotation to include all the features of a reported clinical study (CT-Pub), and larger than previously available CT data sets with a complete set of patient demographic data [27, 28].

3 Methodology

3.1 Data Construction

We collect CT data from `clinicaltrials.gov` using their publicly available API. Our selection criteria include studies that are: 1) interventional trials, 2) completed with results reported, 3) related to one of five common chronic diseases: hypertension, chronic kidney disease, obesity, cancer, diabetes, and 4) reported at least six baseline features. The requirement for a minimum of six baseline features ensures the inclusion of studies with more comprehensive data beyond commonly reported features such as age group, race/ethnicity, and sex. This criterion is implemented to ensure the robustness of our dataset, as some features from the publication about CT may not be reported on the `clinicaltrials.gov`.

Table 1: A sample example from CTBench with CT metadata and corresponding baseline features.

Field	Data
Trial ID	NCT00000620
Trial Title	Action to Control Cardiovascular Risk in Diabetes (ACCORD)
Brief Summary	The purpose of this study is to prevent major cardiovascular events (heart attack, stroke, or cardiovascular death) in adults with type 2 diabetes mellitus using intensive glycemic control, intensive blood pressure control, and multiple lipid management.
Eligibility Criteria	*Inclusion Criteria:* * Diagnosed with type 2 diabetes mellitus, as determined by the new American Diabetes Association guidelines, which include a fasting plasma glucose level greater than 126 mg/dl (7.0 mmol/l), or a 2-hour postload value in the oral glucose tolerance test of greater than 200 mg/dl, with confirmation by a retest .... *Exclusion Criteria:* ...
Conditions	Atherosclerosis, Cardiovascular Diseases, Hypercholesterolemia, ...
Primary Outcomes	First Occurrence of a Major Cardiovascular Event (MCE), ...
Interventions	Anti-hyperglycemic Agents, Anti-hypertensive Agents, ...
Baseline Features	Age, Gender, Ethnicity (NIH/OMB), Race, Region of Enrollment, Previous cardiovascular disease (CVD) event, Glycated hemoglobin, Blood pressure, Cholesterol, Triglycerides, Diabetes duration

For each CT, we collect several types of information (see Table 1). We initially started with 1798 studies returned from the API query. After thorough pre-processing steps, including removing duplicate trials and trials with missing values, we are left with 1693 CTs for our final study.

From our 1693 CTs, we construct two datasets: "CT-Repo" and "CT-Pub" summarized in Table 2 The CT-Repo dataset consists of 1690 trials, with the remaining three trials used as example trials for three-shot learning in LMs. We randomly pick 100 CTs from the CT-Repo dataset to build the CT-Pub dataset. For each trial in CT-Pub, human annotators manually collect the list of baseline features reported in the publications associated with the CT and ensure that: 1) each CT has at least one relevant publication reporting the trial results, 2) the publication contains a table where the baseline features featured for the trial are fully reported, and 3) the publication is evidenced to be connected to the trial by mentioning the trial ID in the publication and/or in the publisher's website.

Table 2: Dataset description for CTBench.

	Total n	Cancer n (%)	Chronic Kidney Disease n (%)	Diabetes n (%)	Hypertension n (%)	Obsesity n (%)
CT-Repo	1690	484 (28.64%)	169 (10.00%)	479 (28.34%)	266 (15.74%)	292 (17.27%)
CT-Pub	100	16 (16.00%)	18 (18.00%)	34 (34.00%)	14 (14.00%)	18 (18.00%)

Challenges: The data extracted from `clinicaltrials.gov` include title, summary, conditions, eligibility criteria, interventions, primary outcomes, and baseline features in free-text format (Table 1). The trial titles and brief summaries provide an overview of the study in plain language, often without consistent terminology. Conditions refer to health issues/diseases being studied written in free text, which can lead to inconsistencies in interpretation due to polysemy (multiple meanings) and synonymy (different terms for the same concept). Eligibility criteria, encompassing both inclusion and exclusion criteria, are detailed as paragraphs, bulleted lists, or enumeration lists, without adherence to common standards or controlled vocabularies. Interventions describe the treatments or procedures being tested, in unstructured text. Primary outcomes and baseline features outline the main objectives and initial data points of the study, respectively, and are similarly unstructured, lacking standardization in terms of medical dictionaries or ontologies. This variability and lack of standardized language across all these fields pose significant challenges for both data extraction and results analysis.

3.2 Generation Task

The CTBench task is to predict the baseline features of a study given the metadata. We demonstrate our benchmarking process and evaluate performance results on two state-of-the-art LMs, open-source LLaMa3-70B-Instruct [29] and commercial GPT-4o [30]. For GPT-4o, we used the API provided by

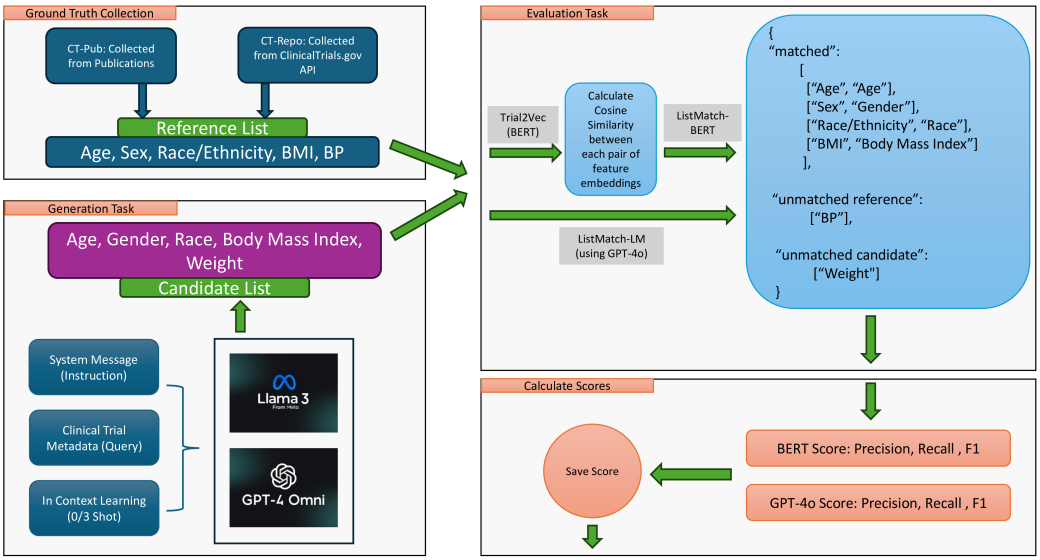

Figure 1: Workflow of CTBench.

OpenAI [31]. For LLaMa-3-70B-Instruct, we used APIs from GROQ [32] and HuggingFace's serverless inference service [33]. We investigate two in-context learning settings for feature generation: zero-shot and three-shot [34]. Each query has the system message and the user query (Figure 1). For the zero-shot setting, we provide CT metadata (excluding the baseline features) as input context to these models (Figure 2), and query the models to generate a list of probable baseline features relevant to the clinical trial. In the three-shot setting (see Appendix C for full prompt template), we extend the zero-shot system prompt by appending trial metadata and corresponding answers (i.e., list of baseline features) for three example trials. All our generation prompts are in Appendix C. For CT-Repo, the generation task involves predicting the list of baseline features reported in the `clinicaltrials.gov` portal using the CT metadata presented in Table 1. For the CT-Pub dataset, the generation task is to predict the baseline features collected from the publications relevant to each trial.

3.3 Evaluation Task

The evaluation task compares the "candidate features" suggested by each LLM with the "reference baseline features" from the CT publications for CT-Pub or `clinicaltrials.gov` API for CT-Repo. The objective is to evaluate each pair of features, one from the reference list and one from the candidate list, to determine if they are contextually and semantically similar, i.e., if they match. We remove noisy keywords from the feature lists (e.g., "Customized," "Continuous") during pre-processing. After identifying all matched pairs, the final results are categorized into three lists: matched pairs, unmatched reference features, and unmatched candidate features. We employ two approaches for identifying matched pairs: "ListMatch-BERT" and "ListMatch-LM." For the evaluation task, we use Trial2Vec and GPT-4o for ListMatch-BERT and ListMatch-LM, respectively. The Trial2Vec implementation requires local installation and a GPU for inference, as it is not readily available through HuggingFace or other inference service providers. We utilized NVIDIA Ampere A100 and NVIDIA T4 GPUs via Google Colab for our work. For GPT-4o as an evaluator, we again used the OpenAI APIs available through their public site. All hyperparameters related to our generation and evaluation tasks are presented in Appendix B. We use a fixed seed and a temperature value of 0.0 across all experiments to ensure the outputs are deterministic and reproducible [35].

ListMatch-BERT: Here we consider a variation of the BERTScore [36]. We utilize Trial2Vec architecture proposed for CTs, built on top of TrialBERT [15] (MIT license) to generate embeddings for each feature and then calculate a cosine similarity matrix for each set of pairs. We explore

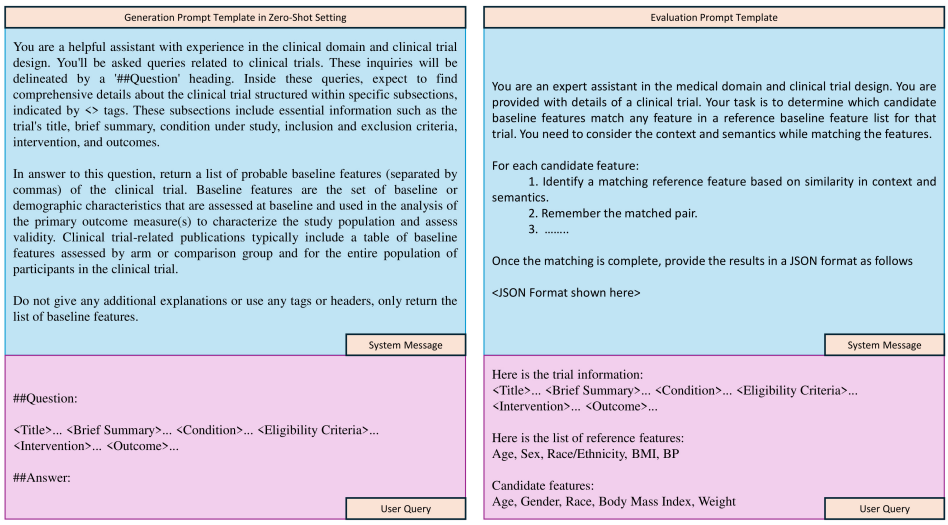

Figure 2: Prompt template for generation (in zero-shot setting) and evaluation

different matching threshold values $T_h \in \{0.6, 0.7, 0.8, 0.9\}$, and recommend using the value of 0.7 (see Appendix D for detailed comparison and reasoning). Matches are considered starting from the pair with the highest cosine similarity above T_h, and these pairs are added to the matched list, and removed from their respective lists and the similarity matrix. Matching continues until: 1) no more matches are found with similarity greater than T_h, or 2) no more features remain to match in either the reference or candidate list. A detailed description of the ListMatch-BERT process is provided in Appendix A.

We report mean Precision, mean Recall, and mean F1 scores across all studies for each dataset. Once the lists of matched pairs, unmatched references, and unmatched candidates are established, and given: TP (True Positives): $n_{matched\_pairs}$, FP (False Positives): $n_{remaining\_candidate\_features}$, FN (False Negatives): $n_{remaining\_reference\_features}$, we calculate precision and recall:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{n_{matched\_pairs}}{n_{matched\_pairs} + n_{remaining\_candidate\_features}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{n_{matched\_pairs}}{n_{matched\_pairs} + n_{remaining\_reference\_features}} \tag{2}$$

ListMatch-LM: Here GPT-4o is prompted to identify matched pairs and the remaining unmatched sets (see Figures 1 and 2). For each study, GPT-4o receives the reference features and candidate features as input. Trial metadata (excluding the actual baseline features) is provided as context. GPT-4o is tasked with identifying matched pairs and generating unmatched lists, which are returned as a JSON object. Mirroring the procedure used in ListMatch-BERT, the model is instructed to remove matched pairs from further consideration immediately upon identification, ensuring that no reference feature is matched to multiple candidate features, and vice versa. Once the matches are generated and the unmatched items are identified, we calculate precision, recall, and F1 scores similarly as described above and report their means over all the studies. Appendix C provides the full evaluation prompt.

Human Evaluation: To evaluate the accuracy of GPT-4o as an evaluator, we employ clinical domain experts to serve as human annotators. Their task is to identify matched pairs for each of the 100 CT studies in the CT-Pub dataset. To streamline the evaluation, we focus exclusively on the candidate responses generated by GPT-4o in the three-shot setting. The annotators receive the same information

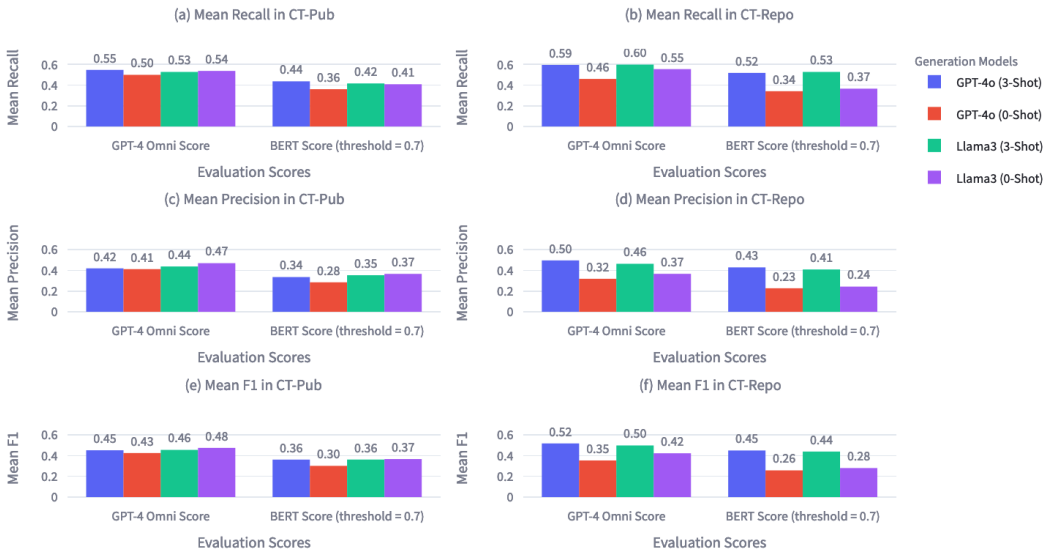

Figure 3: Performance Comparison for CT-Pub and CT-Repo datasets

provided to GPT-4o during its evaluation and are instructed to match features using the same criteria. We developed a web tool to collect and store the responses from all annotators for each of the 100 studies in a database. We also solicit evaluations from human annotators regarding the remaining unmatched candidate features that may merit further examination. Our findings indicate a high level of agreement between the human annotator and GPT-4 Omni's evaluations, underscoring the reliability of GPT-4o in capturing nuanced similarities between features. Detailed results of these experiments are provided in Appendix D.

4 Results and Discussion

In CTBench, precision measures the proportion of predicted baseline features that are accurate, while recall measures the proportion of actual baseline features that the model successfully identifies. We find recall to be of more interest as it ensures comprehensive identification of all relevant baseline features, which is crucial for accurately characterizing study cohorts and maintaining the validity and robustness of clinical trial results. High recall minimizes the risk of missing critical features that could undermine the study's conclusions. Figure 3 shows the performance comparison of GPT-4o and LLaMa3 for CT-Pub and CT-Repo datasets. We find that GPT-4o (3-Shot) leads in recall in the CT-Pub dataset, while LLaMa3 (0-Shot) excels in the CT-Pub dataset for precision and F1 scores. In the CT-Repo dataset, GPT-4o (3-shot) outperforms LLaMa3 across all ICL settings and metrics.

4.1 Performance Analysis in Generation Tasks

4.1.1 Analysis on CT-Pub Dataset

Observation about Metric Values and Model Performance: The values of recall, precision, and F1 scores are not particularly high, indicating a moderate performance of LLaMa3 and GPT-4o in predicting baseline features. This suggests there is room for improvement in the models' ability to generate accurate and comprehensive baseline features.

Comparison of Precision, Recall, and F1 Scores Across Models: The models exhibit varied strengths across different metrics. LLaMa3 (0-Shot) demonstrates the highest precision and F1 score, with an F1 score of 0.48, indicating its strong capability to accurately identify relevant baseline features without requiring prior examples. GPT-4o (3-Shot) leads in the recall, highlighting its

superior ability to retrieve a comprehensive list of relevant baseline features when examples are provided. This suggests that GPT-4o benefits significantly from example-based learning, whereas LLaMa3 performs robustly even in a zero-shot setting, making it a versatile choice for scenarios with limited training data.

ICL Setting Analysis:

- **Zero-shot vs. Three-shot:** In the CT-Pub dataset, LLaMa3 performs better in the zero-shot setting, particularly in precision and F1 score. GPT-4o, however, benefits more from the examples, performing better in the three-shot setting in the recall.
- **Model Benefit from Examples:** GPT-4o shows a significant improvement in recall when examples are provided (3-shot), whereas LLaMa3 shows a higher overall performance in the zero-shot setting.

4.1.2 Analysis on CT-Repo Dataset:

Observation about Metric Values and Model Performance: Similar to the CT-Pub dataset, the values are not exceptionally high, reflecting moderate performance in predicting baseline features. This emphasizes the need for enhanced models to improve prediction accuracy and comprehensiveness.

Comparison of Precision, Recall, and F1 Scores Across Models: The CT-Repo dataset reveals that GPT-4o (3-Shot) outperforms LLaMa3 in precision and F1 score, achieving a notable F1 score of 0.52, while providing comparable performance in recall. This highlights GPT-4o's robustness and effectiveness when prior examples are available, making it highly suitable for matching or adjusting treatment and control subjects in clinical trials and observational studies. LLaMa3 (3-Shot) also demonstrates strong performance, particularly in the recall, indicating its capability to retrieve a broad range of relevant features when examples are provided. The overall moderate performance of both models reflects the complexity and challenging nature of accurately predicting baseline features from clinical trial metadata.

ICL Setting Analysis:

- **Zero-shot vs. Three-shot:** In the CT-Repo dataset, both models perform better in the three-shot setting. GPT-4o significantly benefits from examples, especially in precision and recall.
- **Model Benefit from Examples:** GPT-4o shows substantial improvement with examples (3-shot), indicating its dependency on context for better performance. LLaMa3 also shows improved performance with examples but retains good performance in the zero-shot setting. Since the ground-truth baseline features for CT-Repo were collected from the `clinicaltrials.gov` API, there are specific nuances, such as reporting 'Region of Enrollment' as a baseline feature, which is not typically seen in CT-Pub publications. We believe this context explains why both GPT-4o and LLaMa3 benefit from example-based learning in this scenario.

4.1.3 Why is GPT-4o under-performing significantly and consistently in zero-shot setting?

GPT-4o (zero-shot) underperforms across all cases and scores in both datasets due to the lack of contextual learning from prior examples, which is crucial for accurately interpreting and predicting complex, domain-specific clinical trial features. This setting relies solely on pre-trained knowledge, which is insufficient for the nuanced and detailed task of baseline feature prediction in clinical trials.

4.2 Performance on Evaluation Tasks

GPT-4 Omni Scores: GPT-4 evaluation scores generally surpass BERT scores at a 0.7 threshold due to GPT-4o's broader understanding and contextual evaluation, which captures more nuanced similarities between reference and candidate baseline features. This results in a more generous and context-aware assessment compared to the stricter, more literal BERT scoring.

BERT Scores (threshold = 0.7): After examining several thresholds, we recommend 0.7 to be used as the threshold value for producing BERT scores using ListMatch-BERT. The 0.7 threshold for BERT scores signifies a balance between generous and strict evaluation criteria, requiring high similarity for matches to be considered valid. This, however, reduces precision and recall by demanding closer alignment between generated and actual features compared to lower threshold values. Lowering the threshold would allow for more matches but could increase false positives and false negatives, affecting the precision and recall negatively. We present a thorough evaluation of BERT scores at different threshold values in Appendix D.

Comparing both metrics, we believe that GPT-4 Omni scores suggest a comprehensive and context-sensitive evaluation, crucial for accurately assessing the quality of LM-generated baseline features in clinical trial design.

5 Limitations

CT Data Expansion: Our results, derived from CT data, demonstrate the potential of LLMs to significantly aid in the design and implementation of clinical studies. But the CTBench consists of only RCTs for 5 chronic diseases gathered from `clinicaltrials.gov` with only a subset annotated with additional "gold-standard" from CT-related papers. Using our tools and framework, CTBench could be expanded with other CT repositories, more published CT results, and more diseases. Future work should also explicitly incorporate and evaluate observational studies.

Evaluation Methods: We have presented two LLM-based matching methods and associated evaluation metrics, but how to best evaluate predicted descriptors is an interesting research question in itself. Currently, each reference or candidate item is permitted to be matched only once to provide a standardized fair evaluation across models. But other strategies allowing multiple matches are possible. We hope that the human-in-the-loop evaluation tools provided to compare the LM and human evaluations assist in the further evolution of effective evaluation strategies.

Additional Methods for Generation: Our baseline CTBench study focuses on benchmarking the two state-of-the-art LLaMa3-70B-Instruct and GPT-4o models only with zero-shot and three-shot prompts due to resource constraints. By contrasting an open-source model (LLaMa3-70B-Instruct) with a closed-source model (GPT-4o), we aim to provide a preliminary evaluation of current leading technologies. In our experiments, both for the text generation and evaluation API calls, we have maintained a consistent approach by using a fixed seed and a temperature value set to 0.0. This methodological choice is based on OpenAI's documentation [35], which claims that a fixed seed and a temperature parameter of 0.0 are likely to produce reproducible and deterministic results. But many other possibilities exist. Running each API call multiple times with the same question and considering aggregated answers could improve results. We hope that CT-bench will spur new prompt and model research to expand the scope and depth of AI methods for CT design support.

Impact of Societal Bias: Societal biases present in language models (LMs) can potentially be transferred to clinical trials through the models' baseline feature predictions. This bias could skew the characterization of study cohorts, leading to biased clinical results and affecting the generalizability and applicability of the findings. Such biases in baseline features can undermine the validity of clinical trials, resulting in health outcomes that do not accurately reflect the broader population.

6 Conclusion

CTBench serves as a pioneering benchmark for evaluating LLMs in predicting baseline features from CT metadata - a critical component in CT design. By leveraging datasets from `clinicaltrials.gov` and curated from trial publications, and utilizing advanced evaluation methods such as ListMatch-LM and ListMatch-BERT, CTBench provides a robust framework for assessing AI-generated baseline features. Our results establish a promising baseline, validated through expert human evaluations, and underscore CTBench's potential to significantly enhance the efficacy and robustness of clinical trials through advanced AI research.

Acknowledgments and Disclosure of Funding

This work was supported by IBM Research and the Rensselaer Institute for Data Exploration and Applications.

References

[1] Stuart L Silverman. From randomized controlled trials to observational studies. *The American journal of medicine*, 122(2):114–120, 2009.

[2] David Faraoni and Simon Thomas Schaefer. Randomized controlled trials vs. observational studies: why not just live together? *BMC anesthesiology*, 16:1–4, 2016.

[3] Ravi Thadhani. Formal trials versus observational studies. *Fabry disease: perspectives from 5 years of FOS*, 2006.

[4] Chris Roberts and David J Torgerson. Baseline imbalance in randomised controlled trials. *Bmj*, 319(7203):185, 1999.

[5] Mathias J Holmberg and Lars W Andersen. Adjustment for baseline characteristics in randomized clinical trials. *JAMA*, 328(21):2155–2156, 2022.

[6] Emir Festic, Bhupendra Rawal, and Ognjen Gajic. How to improve assessment of balance in baseline characteristics of clinical trial participants—example from proseva trial data? *Annals of translational medicine*, 4(4), 2016.

[7] Zhongheng Zhang, Alberto Alexander Gayle, Juan Wang, Haoyang Zhang, and Pablo Cardinal-Fernandez. Comparing baseline characteristics between groups: an introduction to the cbcgrps package. *Annals of Translational Medicine*, 5(24), 2017.

[8] ClinicalTrials.gov. Data prep checklist. `https://prsinfo.clinicaltrials.gov/data-prep-checklist-bl.pdf`, 2017. [Accessed 04-06-2024].

[9] Anita van Zwieten, Jiahui Dai, Fiona M Blyth, Germaine Wong, and Saman Khalatbari-Soltani. Overadjustment bias in systematic reviews and meta-analyses of socio-economic inequalities in health: a meta-research scoping review. *International Journal of Epidemiology*, 53(1):dyad177, 2024.

[10] Matthew Hutson. How ai is being used to accelerate clinical trials. *Nature*, 627(8003):S2–S5, 2024.

[11] Jong-Lyul Ghim and Sangzin Ahn. Transforming clinical trials: the emerging roles of large language models. *Translational and Clinical Pharmacology*, 31(3):131, 2023.

[12] Xiong Liu, Greg L Hersch, Iya Khalil, and Murthy Devarakonda. Clinical trial information extraction with bert. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 505–506. IEEE, 2021.

[13] Andriy Mulyar, Ozlem Uzuner, and Bridget McInnes. Mt-clinical bert: scaling clinical information extraction with multitask learning. *Journal of the American Medical Informatics Association*, 28(10):2108–2115, 2021.

[14] Renee White, Tristan Peng, Pann Sripitak, Alexander Rosenberg Johansen, and Michael Snyder. Clinidigest: a case study in large language model based large-scale summarization of clinical trial descriptions. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, pages 396–402, 2023.

[15] Zifeng Wang and Jimeng Sun. Trial2vec: Zero-shot clinical trial document similarity search using self-supervision. *arXiv preprint arXiv:2206.14719*, 2022.

[16] Zifeng Wang, Cao Xiao, and Jimeng Sun. Autotrial: prompting language models for clinical trial design. *arXiv preprint arXiv:2305.11366*, 2023.

[17] Kyeryoung Lee, Hunki Paek, Liang-Chin Huang, C Beau Hilton, Surabhi Datta, Josh Higashi, Nneka Ofoegbu, Jingqi Wang, Samuel M Rubinstein, Andrew J Cowan, et al. Seetrials: Leveraging large language models for safety and efficacy extraction in oncology clinical trials. *medRxiv*, 2024.

[18] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[19] Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, et al. Criteria2query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, 26 (4):294–305, 2019.

[20] Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. Matching patients to clinical trials with large language models. *ArXiv*, 2023.

[21] Surabhi Datta, Kyeryoung Lee, Hunki Paek, Frank J Manion, Nneka Ofoegbu, Jingcheng Du, Ying Li, Liang-Chin Huang, Jingqi Wang, Bin Lin, et al. Autocriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *Journal of the American Medical Informatics Association*, 31(2):375–385, 2024.

[22] Danny M den Hamer, Perry Schoor, Tobias B Polak, and Daniel Kapitan. Improving patient pre-screening for clinical trials: Assisting physicians with large language models. *arXiv preprint arXiv:2304.07396*, 2023.

[23] Nigel Markey, Ben Howitt, Ilyass El-Mansouri, Carel Schwartzenberg, Olga Kotova, and Christoph Meier. Clinical trials are becoming more complex: a machine learning analysis of data from over 16,000 trials. *Scientific Reports*, 14(1):3514, 2024.

[24] ClinicalTrials.gov. Nct03987919 study results. `https://clinicaltrials.gov/study/NCT03987919?tab=results`, 2022. [Accessed 05-06-2024].

[25] Juan P Frías, Melanie J Davies, Julio Rosenstock, Federico C Pérez Manghi, Laura Fernández Landó, Brandon K Bergman, Bing Liu, Xuewei Cui, and Katelyn Brown. Tirzepatide versus semaglutide once weekly in patients with type 2 diabetes. *New England Journal of Medicine*, 385(6):503–515, 2021.

[26] Amos Cahan and Vibha Anand. Second thoughts on the final rule: An analysis of baseline participant characteristics reports on clinicaltrials. gov. *PloS one*, 12(11):e0185886, 2017.

[27] Bevan Koopman and Guido Zuccon. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 669–672, 2016.

[28] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, Steven Bedrick, and Willian R Hersh. Overview of the trec 2021 clinical trials track. In *Proceedings of the thirtieth text retrieval conference (TREC 2021)*, 2021.

[29] AI@Meta. Llama 3 model card. 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`. [Accessed 03-01-2024].

[30] OpenAI. GPT-4 Technical Report. `https://cdn.openai.com/papers/gpt-4.pdf`, 2023. [Accessed 05-06-2024].

[31] OpenAI. OpenAI API. https://openai.com/index/openai-api/, 2021. [Accessed 05-06-2024].

[32] Groq. Groq builds the world's fastest AI inference technology — groq.com. https://groq.com/, 2023. [Accessed 05-06-2024].

[33] HuggingFace. Inference for PROs — huggingface.co. https://huggingface.co/blog/inference-pro, 2023. [Accessed 05-06-2024].

[34] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[35] OpenAI. Reproducable Outputs. https://platform.openai.com/docs/guides/text-generation/reproducible-outputs, 2022. [Accessed 05-06-2024].

[36] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] see section 3 - 5
   (b) Did you describe the limitations of your work? [Yes] see section 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] see section 5
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see section 3 and Appendix
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] see section 5
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] see Appendix
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] see section 3.3
   (b) Did you mention the license of the assets? [Yes] see 3.3
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] see Github link + Appendix
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]