Are Large Language Models Effective in Clinical Trial Design? A Study on Baseline Feature Generation

Anonymous ACL submission

Abstract

In clinical trial design, baseline feature selection is one of the crucial tasks for characterizing study cohorts and ensuring accurate study outcomes. Large Language Models (LLMs) show promise in automating this process by analyzing trial data and identifying key features. To assess the capabilities of LLMs in generating appropriate baseline features for clinical trials, we create two datasets: *CT-Repo*, which contains baseline features from 1,690 clinical trials sourced from `clinicaltrials.gov`, and *CT-Pub*, a curated subset of 100 clinical trials with more detailed baseline features extracted from published studies. In this paper, we consider GPT-4o and LLaMa3-70B-Instruct models in three configurations: zero-shot, three-shot with a fixed set of examples, and three-shot using an adaptive set of examples based on Retrieval-Augmented Generation (RAG) approach. We evaluate the model performance of baseline feature generation using the *LLM-as-a-Judge* framework. We further validate the LLM-as-a-judge evaluation on the CT-Pub dataset using assessments from human experts in a clinical trial. The results indicated that the RAG-based three-shot learning approach significantly improved performance by providing relevant, context-specific examples. This study marks an important initial advancement in using LLM for the robust design of clinical trials and observational studies.

1 Introduction

Clinical trials (CTs) are crucial for medical research, with randomized CTs considered the gold standard for assessing the effectiveness of drug interventions. Baseline features, often presented as "Table 1" in clinical publications, include essential demographic and relevant characteristics collected from participants before the commencement of the clinical study. These features are essential for demonstrating population representativeness, validating study design, and drawing logical conclusions (Holmberg and Andersen, 2022; Festic et al., 2016; Zhang et al., 2017). The preselection of baseline features ensures unbiased randomization, allows for pre-specified covariate adjustment, and complies with regulatory requirements (Burgess et al., 2003; Holmberg and Andersen, 2022; Archives, 2024). These features also prevent post-hoc bias and reduce confounding effects, particularly in observational studies where improper confounder selection could lead to over-adjustment bias (Vickers and Altman, 2001; van Zwieten et al., 2024).

LLMs have shown promise in clinical research, including extracting clinical information (Liu et al., 2021; Mulyar et al., 2021), summarizing CT descriptions (White et al., 2023), and comparing trial similarities (Wang and Sun, 2022). Recent advances in prompting strategies have expanded the LLM use cases in specific medical domains (Wang et al., 2023; Lee et al., 2024; Singhal et al., 2023). Several studies have explored using LLMs to aid in creating eligibility criteria for CTs (Yuan et al., 2019; Jin et al., 2023; Datta et al., 2024; Hamer et al., 2023), demonstrating the potential for reducing expert screen time and improving trial matching efficiency.

However, the task of automating the baseline feature selection for CTs remains largely unexplored. Baseline features have become increasingly complex in the last decade (Markey et al., 2024), leading to a need for approaches that can suggest or generate standardized sets of cohort demographics and features. Existing datasets are limited by small CT cohort sizes and/or rely on general clinical notes rather than CT-specific data (Koopman and Zuccon, 2016; Roberts et al., 2021). Furthermore, while `clinicaltrials.gov` provides extensive trial information, it often lacks comprehensive baseline feature data reported in final publications (Cahan and Anand, 2017).

To address these gaps, we present two novel

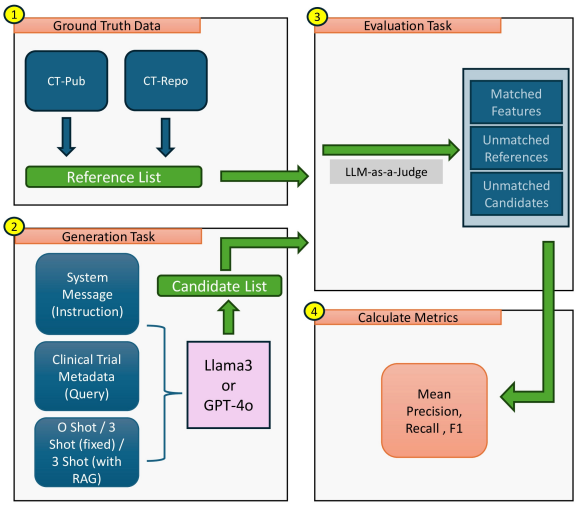

Figure 1: Overview of our process of Generating Baseline Features and Evaluating using LLMs

datasets: CT-Pub and CT-Repo, which include metadata and baseline features from 1,690 CTs and 100 manually curated trials, respectively. Our main contributions include 1) creating CT metadata datasets with associated baseline features for trials from a set of five chronic diseases; 2) evaluating LLM capabilities in CT baseline feature generation; 3) developing automated evaluation methods for comparing predicted and actual trial baseline features; and 4) demonstrating the use of robust prompt engineering techniques on several LLMs to generate baseline feature variables and evaluate their performance. We plan to release our data, code, and demo examples publicly to encourage further research in baseline feature selection and potentially improve the efficiency and robustness of clinical study design.

2 Baseline Feature Selection in Clinical Trial Design

The task of baseline feature selection is essential in clinical trial design to ensure balanced groups and valid outcomes. This process can be complex and requires expert judgment based on trial objectives and target populations. Common baseline features like age, gender, and race/ethnicity are prevalent across all types of trials, while trial-specific features such as 'ECOG Performance Status' or 'Response to Immunotherapy' are more relevant in cancer trials, and 'HBA1C levels' are commonly used in diabetes trials. To address this challenge, we propose a method that treats baseline feature selection as a *text-generation* task. In this approach, an LLM is provided with trial metadata and tasked with generating a list of appropriate baseline features. The goal of this task is to automate and assist trial designers in the initial feature selection process. See Table 3 in Appendix A for a sample clinical trail metadata and the corresponding baseline features.

2.1 Data

We used the public API of `clinicaltrials.gov` to collect data on interventional clinical trials for five common chronic diseases. Our selection criteria include completed trials with reported results and a minimum of six baseline features. After processing, we had a set of 1693 CTs, from which we created **CT-Repo** (consisting of 1690 trails). The remaining 3 trials were used as fixed few-shot examples in the LLM prompt. We chose a subset of 100 trials from CT-Repo and manually annotated baseline features from associated publications to create **CT-Pub**. The CT-Repo dataset showcases a wide spectrum of conditions within each disease category, ranging from those directly related to the primary disease (e.g., various cancer types) to associated conditions evaluated in these studies (e.g., depression, pain). This diversity broadens the applicability of our work across numerous clinical trials, enhancing its generalizability to the entire `clinicaltrials.gov` database. See Table 1 for a comprehensive breakdown of these health conditions and Table 3 for an overview of the structure of the data we collected.

The CT-Pub and CT-Repo datasets are designed as evaluation benchmarks, offering a balanced representation of various conditions while maintaining a manageable size for a thorough assessment. This allows for a comprehensive evaluation of model performance across diverse medical contexts without the computational overhead associated with larger datasets. The dataset includes a variety of trial metadata in free-text format, including titles, summaries, conditions, eligibility criteria, interventions, primary outcomes, and baseline features. The unstructured nature of this data presents challenges due to inconsistent terminology and lack of standardization.

2.2 Generation

In our study, we primarily assess the model performance using the state-of-the-art open-source LLM, LLaMa3-70B-Instruct (AI@Meta, 2024). To provide a broader context, we also evaluate the

Table 1: Description of CT-Pub and CT-Repo datasets

	CT-Pub		CT-Repo	
Trial Groups	**# Trials**	**# Unique Conditions**	**# Trials**	**# Unique Conditions**
Cancer	16	49	484	756
Chronic Kidney Disease	18	23	169	289
Diabetes	34	39	479	196
Hypertension	14	25	266	188
Obesity	18	20	292	205
Total	100	156	1690	1634

commercial GPT-4o (OpenAI, 2023) for comparison (see Figure 1). For LLaMa3-70B-Instruct, we utilized APIs from GROQ (Groq, 2023) and HuggingFace's serverless inference service (HuggingFace, 2023), while OpenAI's API was used for GPT-4o (OpenAI, 2021). Our investigation focused on two in-context learning settings for baseline feature generation: zero-shot and three-shot (Dong et al., 2022). In the three-shot prompting, we explored both a fixed set of three random trials and a Retrieval-Augmented Generation (RAG) approach based on selecting the most similar trials as three-shot examples. For the RAG approach, we created an indexed database of all trials in the CT-Repo dataset. When using a trial as a query, we retrieve the three most similar trials from this database to serve as examples in the few-shot prompt, providing more contextually relevant comparisons than random examples. See Appendix B for more details on RAG setting.

2.3 Evaluation

We evaluate the "candidate features" generated by each LLM with the "reference baseline features" sourced from the `clinicaltrials.gov` API for the CT-Repo dataset and the corresponding CT publications for the CT-Pub dataset. The goal is to assess each pair of features, one from the reference list and one from the candidate list, to determine if they are contextually and semantically similar. For example, 'BMI" should match "Body Mass Index". After identifying all matched pairs, we categorize the final results into three lists: *matched pairs*, *unmatched reference* features, and *unmatched candidate* features. We employ a 'LLM-as-a-Judge' approach for this evaluation, utilizing GPT-4o as our evaluator. For each study, the evaluator receives both the reference and candidate features as input, along with trial metadata (excluding actual baseline features) for context. The evaluator then identifies matched pairs and generates unmatched sets, returning the results as a JSON object. Once matched and unmatched items are identified, we calculate precision, recall, and F1 scores, reporting their mean values across all studies. We further validate the LLM-as-a-judge evaluator on the CT-Pub dataset using assessments from human experts in a clinical trial (see Appendix E). All hyperparameters, prompts, and other details pertaining to our generation and evaluation tasks are presented in the Appendix D. To ensure deterministic and reproducible outputs, we use a fixed seed and a temperature value of 0.0 across all generation and evaluation experiments (OpenAI, 2022).

2.4 Metric

We report mean Precision, mean Recall, and mean F1 scores across all studies for each dataset (see details in Appendix D)

$$\text{Precision} = \frac{\text{number of correctly matched features}}{\text{number of candidate features}}$$

$$\text{Recall} = \frac{\text{number of correctly matched features}}{\text{number of reference features}}$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

3 Results and Discussion

We prioritize the F1 score in our analysis as it provides a balanced measure of precision and recall. This balance is crucial in the context of clinical trial design, where both accuracy in suggesting relevant features and completeness in covering all necessary aspects are important. Additional experiments are available in Appendix E.

CT-Pub vs CT-Repo Dataset: In the CT-Pub dataset, we observe F1 scores ranging from 0.40 to 0.49 across different models and settings. GPT-4o with RAG-based examples demonstrates the best performance with an F1 score of 0.49. This sug-

Table 2: Performance Comparison for CT-Pub and CT-Repo datasets. **Bold** fonts indicate the best performance.

	CT Pub			CT Repo		
Model	F1	Precision	Recall	F1	Precision	Recall
GPT-4o (Zero Shot)	0.40	0.36	0.50	0.33	0.27	0.51
LLama3 (Zero Shot)	0.46	0.43	0.55	0.40	0.32	0.57
GPT-4o (Three Shot + Fixed Example)	0.43	0.39	0.56	0.46	0.40	0.58
LLama3 (Three Shot + Fixed Example)	0.46	0.42	**0.57**	0.45	0.41	0.56
GPT-4o (Three Shot + RAG Based Example)	**0.49**	0.53	0.54	0.49	0.45	**0.63**
LLama3 (Three Shot + RAG Based Example)	0.48	**0.54**	0.50	**0.51**	**0.48**	0.62

gests that providing contextually relevant examples significantly enhances the model's ability to generate appropriate baseline features. Interestingly, Llama3 shows competitive performance, especially in the zero-shot setting (F1: 0.46), indicating its strong inherent understanding of clinical trial contexts without additional prompting. The CT-Repo dataset shows a wider range of F1 scores, from 0.33 to 0.51. Llama3 with RAG-based examples achieves the highest F1 score of 0.51, outperforming GPT-4o in this larger, more diverse dataset. This indicates Llama3's strength in generalizing across a broader range of trial types. GPT-4o shows substantial improvement as more context is provided, with its F1 score increasing from 0.33 to 0.49, highlighting the importance of relevant examples in enhancing performance.

GPT-4o vs Llama3: GPT-4o generally performs better on the CT-Pub dataset, possibly due to its extensive pre-training on diverse medical literature. However, Llama3 shows stronger performance on the larger CT-Repo dataset, suggesting its robustness in handling a wider variety of trial types. Both models benefit from additional context, indicating that their performance in suggesting baseline features can be enhanced through strategic prompting and example selection.

Few-shot Prompting: Zero-shot performance serves as a baseline, with Llama3 outperforming GPT-4o in these scenarios. This suggests Llama3's strong inherent understanding of clinical trial contexts. The three-shot approach with fixed examples shows modest improvement over zero-shot for both models, indicating that even random examples can provide useful context. However, the RAG-based three-shot approach consistently yields the best overall performance for both models, with a significant boost in precision, especially for GPT-4o. This clearly shows that there is value in providing relevant trials as context for designing new trials.

RAG Approach: The RAG-based approach consistently yields the highest F1 scores across models and datasets by retrieving relevant context from similar trials in a vectorized database. Unlike fixed or random examples, this additional context from RAG closely aligns with the current trial's characteristics which helps the LLMs understand patterns of baseline features typically associated with specific trial types, enhancing their ability to generate appropriate baseline feature lists. While there's room for improvement, the RAG approach shows significant potential in streamlining the clinical trial design process. (See details in Appendix B, E)

4 Conclusion

In this study, we tackled the challenge of baseline feature generation in clinical trial design by leveraging state-of-the-art large language models . We evaluated GPT-4o and LLaMa3-70B-Instruct on two custom datasets, CT-Repo and CT-Pub, using zero-shot, three-shot with fixed examples, and three-shot with RAG-based approaches. Our "LLM-as-a-Judge" framework, utilizing GPT-4o as an evaluator, validated through human-in-the-loop assessments with clinical experts, revealed that while current state-of-the-art models can moderately identify baseline features, they often struggle without additional context, particularly in complex clinical settings. The RAG-based approach consistently outperformed other configurations by providing relevant, context-specific examples that improve the model's predictions. This work represents an important first step in developing AI tools for clinical trial design. Beyond clinical research, our work contributes to the broader NLP community by demonstrating how LLMs can be adapted to highly specialized, high-stakes domains like healthcare. Future work could focus on fine-tuning LLMs on clinical trial data, and expanding their integration to accelerate medical research and enhance patient outcomes.

5 Limitations

Focused Contribution and Scope of Dataset: Our study focuses on five major chronic diseases: cancer, chronic kidney disease, diabetes, hypertension, and obesity which are the most common and widely studied diseases in clinical trials. These broad-categories include hundreds of different unique health conditions (see Table 1). While this represents an important portion of the available trials on clinicaltrials.gov with high health impact, we acknowledge that this focus limits the scope of our findings to these conditions. This work should be viewed as a small, focused contribution designed to demonstrate the feasibility and potential of LLMs in automating baseline feature generation for high-prevalence diseases. Expanding this work to rare diseases or other less-studied conditions is part of our future research direction, and we believe that our pipeline can seamlessly incorporate more diverse clinical trial datasets without modification.

Scale of the Dataset: While evaluating language models on a dataset of 1,690 randomized clinical trials from clinicaltrials.gov may seem like a limitation given the availability of over 400,000 trials of many types, it is not truly restrictive in this context. We selected a diverse and representative set of trials across five major chronic disease categories (see Table 1) which covers a broad spectrum of trial designs and conditions, providing meaningful insights into the models' ability to predict baseline features. Importantly, our prompts and evaluation methods are fully adaptable to any number of additional trials without altering the existing pipeline, making the evaluation scalable. While this subset was chosen to ensure rigorous and reproducible results within current resource constraints, this work is a work in progress. Future expansions will seamlessly incorporate more trials, continuing to build on the robust evaluation framework established here.

Additional Methods for Generation and Evaluation: Our study evaluates two state-of-the-art models, LLaMa3-70B-Instruct and GPT-4o, using zero-shot and three-shot prompts, with resource constraints in mind. By comparing an open-source model (LLaMa3-70B-Instruct) with a closed-source model (GPT-4o), we aim to provide an initial assessment of leading technologies. While this comparison offers a valuable contrast between open-source and proprietary models in supporting clinical trial design, there are many other models worth exploring. For evaluation, we utilize GPT-4o as the judge, though alternatives like LLaMa3 or Mistral are viable options. In the future, we plan to expand our experiments to include additional models for both generation and evaluation tasks.

Randomness in experiments: In our experiments, both for text generation and evaluation API calls, we ensured consistency by using a fixed seed and setting the temperature parameter to 0.0. This choice follows OpenAI's guidelines (OpenAI, 2022), which suggest that a fixed seed and a temperature of 0.0 help produce reproducible and deterministic results. However, alternative approaches exist. For instance, running each API call multiple times with the same prompt and aggregating the responses could enhance the results, though we were unable to pursue this due to resource constraints. This is a work in progress, and as part of our future work, we plan to explore these alternative approaches to further improve the robustness of our evaluations.

Impact of Societal Bias: While our work demonstrates the potential of language models as tools to determine baseline features for clinical trials, it is important to consider potential biases, particularly in the handling of demographic features such as race, ethnicity, age, and gender. One challenge is that older trials in the benchmarks may not meet current requirements for definition of these features. Non-representative trials have been a problem historically and the National Institute Health for designing trials has evolved over time. Problematic trials in the dataset may introduce inappropriate examples in context sensitive learning and introduce bias in the evaluation process. As part of our future work, we plan to incorporate bias detection and mitigation strategies for these sensitive features, alongside human evaluation, to ensure that the generated baseline features are equitable and representative of diverse trial populations and that they meet current standards for trial design. Additionally, it is important to recognize that ethical considerations must be integrated when applying LLMs in high-stakes domains.

References

AI@Meta. 2024. Llama 3 model card. [Accessed 03-01-2024].

National Archives. 2024. 42 cfr § 11.48 - what constitutes clinical trial results information? https://www.ecfr.gov/current/title-42/chapter-I/subchapter-A/part-11/subpart-C/section-11.48. Accessed: Accessed 09/10/2024.

David C Burgess, Val J Gebski, and Anthony C Keech. 2003. Baseline data in clinical trials. *The Medical Journal of Australia*, 179(2):105–107.

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

Surabhi Datta, Kyeryoung Lee, Hunki Paek, Frank J Manion, Nneka Ofoegbu, Jingcheng Du, Ying Li, Liang-Chin Huang, Jingqi Wang, Bin Lin, et al. 2024. 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.

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

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

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

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

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

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

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

Bevan Koopman and Guido Zuccon. 2016. 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.

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. 2024. Seetrials: Leveraging large language models for safety and efficacy extraction in oncology clinical trials. *medRxiv*.

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

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

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

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

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

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

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

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. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Anita van Zwieten, Jiahui Dai, Fiona M Blyth, Germaine Wong, and Saman Khalatbari-Soltani. 2024. 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.

Andrew J Vickers and Douglas G Altman. 2001. Analysing controlled trials with baseline and follow up measurements. *Bmj*, 323(7321):1123–1124.

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

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

Renee White, Tristan Peng, Pann Sripitak, Alexander Rosenberg Johansen, and Michael Snyder. 2023. 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.

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

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

A CT Metadata and Corresponding Baseline Features

In Table 3, we present all the metadata for each clinical trial that we collect. All of them are in non-standardized free-text form which makes it very challenging to work with.

A.1 Example of Generation Response

Here are two examples of generated baseline features:

- **Generated by Llama3 in Three-Shot settings with fixed examples:** [‘Age’, ‘Sex’, ‘Race’, ‘Duration of diabetes’, ‘HbA1c’, ‘BMI’, ‘Weight’, ‘Waist circumference’, ‘Systolic blood pressure’, ‘Diastolic blood pressure’, ‘Fasting plasma glucose’, ‘Total cholesterol’, ‘LDL cholesterol’, ‘HDL cholesterol’, ‘Triglycerides’, ‘eGFR’, ‘Use of antihyperglycemic drugs’, ‘History of diabetic retinopathy’, ‘History of cardiovascular disease’]

- **Generated by GPT-4o in Three-Shot settings with fixed examples:** [‘Age’, ‘Sex’, ‘Race’, ‘Duration of diabetes’, ‘HbA1c’, ‘BMI’, ‘Fasting plasma glucose’, ‘Systolic blood pressure’, ‘Diastolic blood pressure’, ‘Total cholesterol’, ‘HDL cholesterol’, ‘LDL cholesterol’, ‘Triglycerides’, ‘Non-insulin antidiabetic therapy’]

A.2 Example of Evaluation by GPT-4o

Here is an example of identified matches between Llama3-generated features from Appendix A.1 and actual reference features:

- **Reference Features:** [‘Age‘, ‘Gender‘, ‘Racial Group‘, ‘Body Weight‘, ‘BMI‘, ‘Estimagted GFR‘, ‘Duration of Diabetes‘, ‘Duration of Basal Insulin‘, ‘Prior Basal Insulin Dose‘, ‘HbA1c‘, ‘Concomitant antihyperglycaemic medication use‘]

- **Candidate Features:** [‘Age’, ‘Sex’, ‘Race’, ‘Duration of diabetes’, ‘HbA1c’, ‘BMI’, ‘Weight’, ‘Waist circumference’, ‘Systolic blood pressure’, ‘Diastolic blood pressure’, ‘Fasting plasma glucose’, ‘Total cholesterol’, ‘LDL cholesterol’, ‘HDL cholesterol’, ‘Triglycerides’, ‘eGFR’, ‘Use of antihyperglycemic drugs’, ‘History of diabetic retinopathy’, ‘History of cardiovascular disease’]

- **Matched Features:** [["Age", "Age"], ["Gender", "Sex"], ["Racial Group", "Race"], ["BMI", "BMI"], ["Duration of Diabetes", "Duration of diabetes"], ["HbA1c", "HbA1c"], ["Estimagted GFR","eGFR"]]

- **Unmatched Reference Features:** ["Body Weight", "Estimagted GFR", "Duration of Basal Insulin", "Prior Basal Insulin Dose", "Concomitant antihyperglycaemic medication use"]

- **Unmatched Candidate Features:** ["Weight", "Waist circumference", "Fasting plasma glucose", "Systolic blood pressure", "Diastolic blood pressure", "Total cholesterol", "HDL cholesterol", "LDL cholesterol", "Triglycerides", "Use of antihyperglycemic drugs", "History of diabetic retinopathy", "History of cardiovascular disease"]

B RAG Method

In the context of our study, we implemented a Retrieval-Augmented Generation (RAG) method for the 3-shot settings to improve the performance of language models (LLMs) in predicting baseline features for clinical trials. Typically, 3-shot learning involves providing three fixed examples as context to guide the model’s generation. However, instead of using random or pre-selected examples, we enhanced this approach by dynamically retrieving the three most similar trials from an indexed vector database containing trial metadata.

Here’s how the RAG process works in our pipeline:

- Prompt Creation: A prompt is generated, which includes a specific query about the trial for which baseline features need to be predicted.

- Query the Vector Database: The query is sent to a retriever system that interacts with a vector database containing embeddings of all trials from the CT-Repo dataset.

- Retrieval of Similar Trials: The retriever identifies the three most similar trials to the query trial based on their metadata and retrieves these trials.

- Context Augmentation: The retrieved trials are then used as additional context for the prompt. These examples are more relevant

Table 3: A sample example of clinical trial metadata and corresponding baseline features

Field	Data
Trial ID	NCT01676220
Trial Title	Comparison of a New Formulation of Insulin Glargine With Lantus in Patients With Type 2 Diabetes on Non-insulin Antidiabetic Therapy
Brief Summary	Primary Objective: To compare the efficacy of a new formulation of insulin glargine and Lantus in terms of change of HbA1c from baseline to endpoint ...
Eligibility Criteria	*Inclusion Criteria:* *Adult participants with type 2 diabetes mellitus inadequately controlled with non-insulin antihyperglycemic drug(s); * Signed written informed consent. .... *Exclusion Criteria:* ...
Conditions	Type 2 Diabetes Mellitus, ...
Primary Outcomes	Change in HbA1c From Baseline to Month 6 Endpoint, ...
Interventions	HOE901-U300 (new formulation of insulin glargine), Lantus (insulin glargine) ...
Baseline Features	Age, Gender, Racial Group, Body Weight, BMI, Estimagted GFR, Duration of Diabetes, Duration of Basal Insulin, Prior Basal Insulin Dose, HbA1c, Concomitant antihyperglycaemic medication use

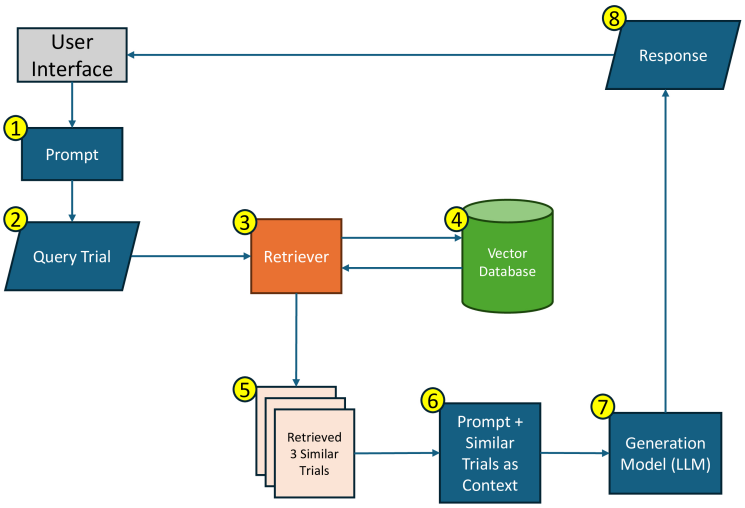

Figure 2: RAG-based approach to retrieve similar trials for 3-shot examples

than random selections, providing the model with similar cases to draw from.

- LLM Generation: The LLM (LLaMa3 or GPT-4o in our study) processes the query prompt alongside the retrieved similar trials as context to generate predictions for the baseline features.
- Response Evaluation: The generated baseline features are compared to the reference list from the trial metadata using our LLM based evaluation method to assess the accuracy.

By using RAG, we effectively provide the model with more contextually relevant examples, which improves its ability to predict accurate baseline features. This approach leverages the most similar trials to guide the model's learning, leading to better performance compared to using fixed examples. Additionally, this method allows for scalability, as the indexed database can expand to include more trials, providing ever-better contextual examples for future predictions.

C Prompts

C.1 Generation Prompt: Zero-shot

Figure 3 illustrates the full prompt used to generate LLM responses (i.e., baseline features) in a zero-shot setting. The system message includes detailed

instructions for the LLM, specifying the format and structure of the user query. Following this, the user query provides the trial information as context, serving as the question for the LLM.

C.2 Generation Prompt: Three-shot

Figure 4 shows the complete prompt used to generate LLM responses (i.e., baseline features) in a three-shot setting (both with fixed example and RAG based adaptive examples). The system message contains detailed instructions for the LLM, including the format and structure of the user query and instructions to expect three examples with their corresponding answers. Next, the user query provides example trial information and their answers as additional context, followed by the actual trial information serving as the question for the LLM.

C.3 Evaluation Prompt

Figure 5 displays the complete prompt used to evaluate LLM responses (i.e., candidate features) against a set of reference baseline features. The system message provides detailed instructions for the LLM on how to perform the matching and how to return the response in JSON format. Following this, the user query includes corresponding trial information, along with the list of reference features and candidate features, which serve as the question for the LLM to evaluate.

D Experimental Design

D.1 Hyperparameters

We present all our experimental hyperparameters for both generation and evaluation task in Table 4 in Appendix. We use a fixed seed and a temperature value of 0.0 across all experiments to ensure the outputs are deterministic and reproducible.

D.2 Computational Resources used

We spent around $400 throughout all of our experiments (both generation and evaluation in zero-shot and three-shot settings) using GPT-4o models. Besides that, we used around 250 compute units from Google Colab for GPU computations and around $100 in monthly subscription fees for HuggingFace Pro account for working with Llama3 models.

D.3 Metric Adjustment

Let's look at a hypothetical example -

- Reference Features = ['Age', 'Blood pressure', 'Height', 'Gender','Previous Medication']
- Candidate Features = ['Age', 'Systolic Blood pressure', 'Diastolic Blood Pressure', 'Body Mass Index', 'Race']

Let's assume, GPT-4o is asked to evaluate these two lists and find out matched features (a list of pairs, in each pair, the first element is from reference features and the second element is from candidate features) and features that are not matched. This is what it returns -

- Matched Features = [['Age', 'Age'], ['body mass index', 'Body Mass Index'], ['Blood Pressure', 'Systolic Blood Pressure'], ['Blood Pressure', 'Diastolic Blood Pressure'], ['Height', 'patient height']]
- Remaining Reference Features = ['Previous Medication']
- Remaining Candidate Features = ['Race']

So based on these outputs, we consider 3 types of possible errors -

- **Error type 1:** 'body mass index' and 'patient height' are not valid reference and candidates features respectively, but however are present during matching
- **Error type 2:** "Gender' was a valid reference features that was either supposed to be matched with some feature from candidate features list, or remain unmatched. But after the matching process, we find that 'Gender' does not appear anywhere.
- **Error type 3:** "Blood Pressure' is used in matching twice (once matched with 'Systolic Blood Pressure', and then again matched with 'Diastolic Blood Pressure') while we specifically instructed the evaluator to allow each feature to match only once.

So to adjust for these errors, in our Precision and Recall formulae, we only consider a number of matches that are correct. We define number of correctly matched features = number of total matches - number of type 1 errors - (number of type 3 errors - 1). This allows us to correctly penalize for this possible errors and gives us a conservative estimate of models' performance.

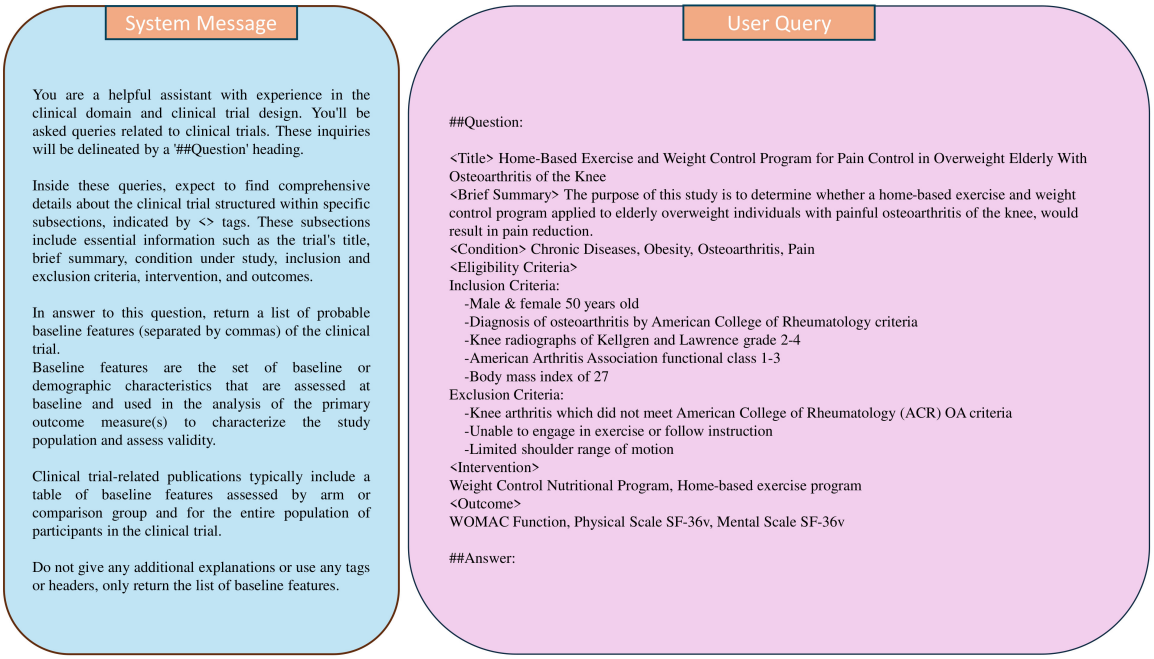

Figure 3: Full Prompt for Generation Task in Zero-Shot setting

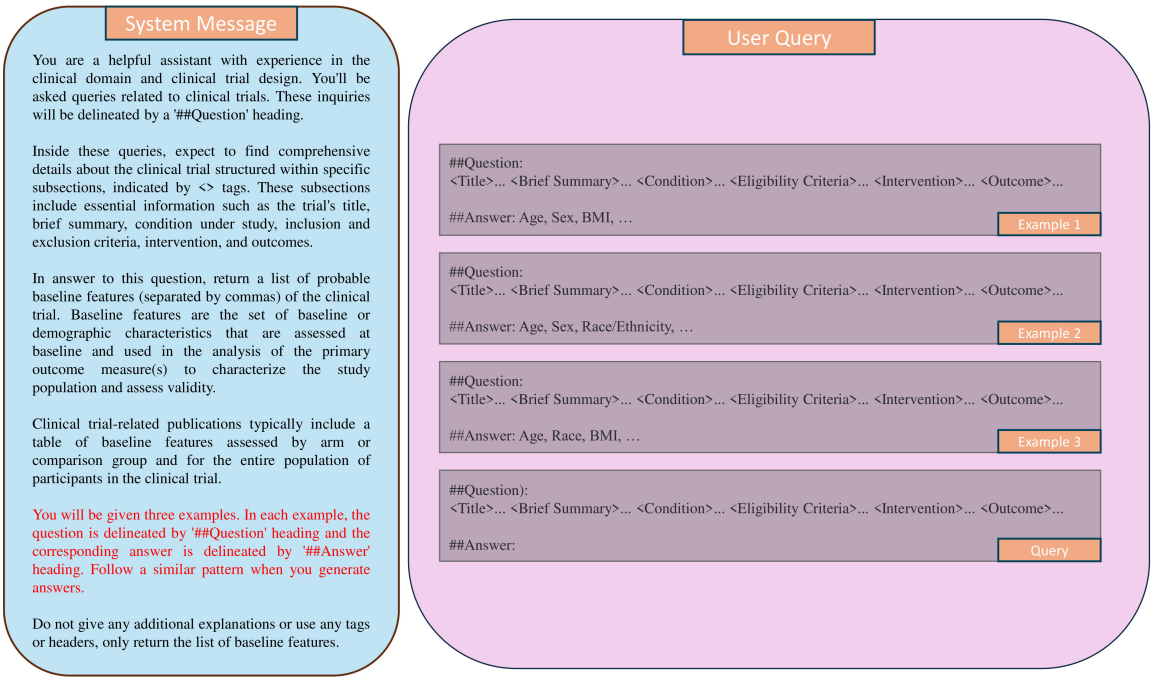

Figure 4: Full Prompt for Generation Task in Three-Shot setting

Table 4: Hyperparameters for experiment

Models	Seed	Temperature	Max Token	Message Format	Response Format
LLaMa-3-70B-Instruct (as generator)	42	0.0	1000	{"role": "system", "content": system_message} {"role": "user", "content": user_query}	Default
GPT-4o (as generator)	42	0.0	1000	{"role": "system", "content": system_message} {"role": "user", "content": user_query}	Default
GPT-4o (as evaluator)	42	0.0	1000	{"role": "system", "content": system_message} {"role": "user", "content": user_query}	JSON

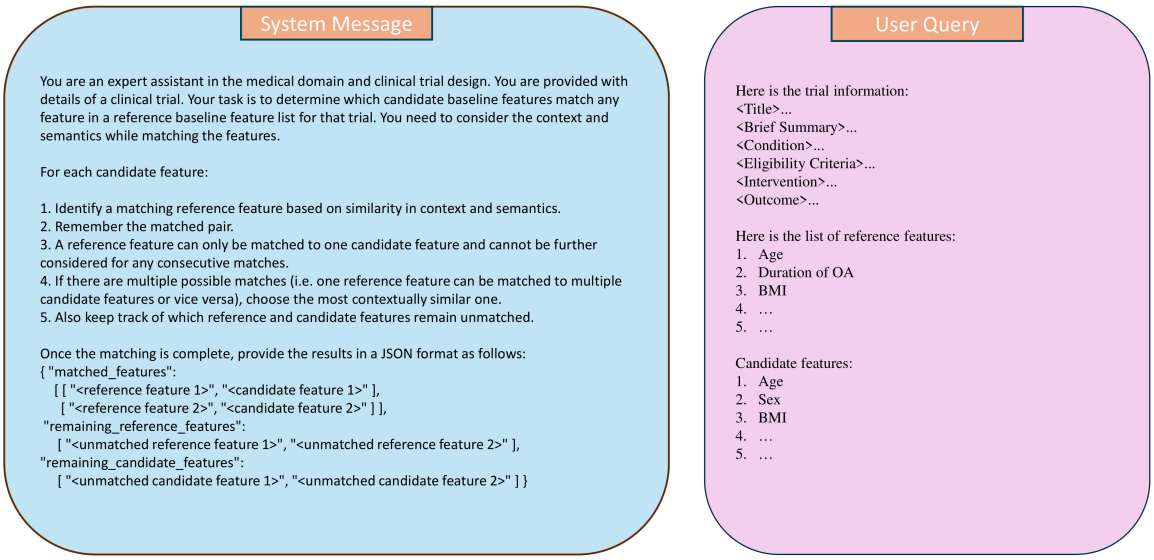

Figure 5: Full Prompt for Evaluation Task

E Additional Experiments

E.1 Analysis by Trial Group in CT-Pub

Table 5 compares the mean F1 scores of two models, GPT-4o and LLaMa3, across five trial groups (Cancer, Chronic Kidney Disease, Diabetes, Hypertension, and Obesity) within the CT-Pub dataset. Three different settings are evaluated for each model: Zero Shot, Three Shot with Fixed Examples, and Three Shot with RAG-based Examples.

Key observations include:

- Overall Best Performance: GPT-4o with the Three Shot + RAG-based Example setting consistently shows the best performance in most trial groups, achieving the highest F1 scores for Cancer (0.45), Chronic Kidney Disease (0.52), Diabetes (0.56), and Obesity (0.46).

- Impact of RAG: Both models perform better in the Three Shot + RAG-based Example setting compared to the Three Shot + Fixed Example and Zero Shot settings. This highlights the effectiveness of the RAG approach, where retrieving the most similar trials improves the prediction accuracy for baseline features.

- Model Comparison: While GPT-4o generally outperforms LLaMa3 in most categories, LLaMa3 performs closely in certain groups, especially in the Diabetes trial group, where its F1 score of 0.55 is nearly on par with GPT-4o (0.56). In the Obesity trial group, LLaMa3 also improves significantly with RAG, though GPT-4o still slightly outperforms it.

- Trial Group Variations: Performance varies across the trial groups. Chronic Kidney Disease and Diabetes have the highest F1 scores, suggesting that models are more successful at predicting baseline features in these groups. In contrast, the Cancer and Obesity groups have lower scores, indicating more difficulty in predicting features accurately in these trial types.

Overall, the results demonstrate the benefit of using RAG-based retrieval in the Three Shot setting for improving model performance across different clinical trial groups, with GPT-4o showing generally stronger results.

E.2 Analysis by Trial Group in CT-Repo

In this table, we see the comparison of mean F1 scores for GPT-4o and LLaMa3 across five trial groups in the CT-Repo dataset, under similar settings: Zero Shot, Three Shot with Fixed Examples, and Three Shot with RAG-based Examples.

Key observations:

- Overall Best Performance: LLaMa3 with Three Shot + RAG-based Examples achieves

Table 5: Comparison of mean F1 scores by Trial Groups in CT-Pub Dataset. **Bold** fonts indicate best performance.

	Trial Group (CT-Pub)				
Model	**Cancer**	**Chronic Kidney Disease**	**Diabetes**	**Hypertension**	**Obesity**
GPT-4o (Zero Shot)	0.31	0.44	0.45	0.39	0.36
LLama3 (Zero Shot)	0.42	0.49	0.50	0.48	0.36
GPT-4o (Three Shot + Fixed Example)	0.36	0.47	0.46	0.44	0.40
LLama3 (Three Shot + Fixed Example)	0.42	**0.52**	0.49	**0.45**	0.38
GPT-4o (Three Shot + RAG Based Example)	**0.45**	**0.52**	**0.56**	0.42	**0.46**
LLama3 (Three Shot + RAG Based Example)	0.40	0.51	0.55	0.43	0.43

Table 6: Comparison of mean F1 scores by Trial Groups in CT-Repo Dataset. **Bold** fonts indicate best performance.

	Trial Group (CT-Repo)				
Model	**Cancer**	**Chronic Kidney Disease**	**Diabetes**	**Hypertension**	**Obesity**
GPT-4o (Zero Shot)	0.29	0.34	0.36	0.36	0.34
LLama3 (Zero Shot)	0.35	0.42	0.42	0.44	0.37
GPT-4o (Three Shot + Fixed Example)	0.41	0.46	0.48	0.48	0.46
LLama3 (Three Shot + Fixed Example)	0.40	0.46	0.48	0.48	0.45
GPT-4o (Three Shot + RAG Based Example)	0.44	0.48	0.53	0.53	0.50
LLama3 (Three Shot + RAG Based Example)	**0.46**	**0.49**	**0.56**	**0.54**	**0.51**

the highest F1 scores in all trial groups, outperforming GPT-4o in Cancer (0.46), Chronic Kidney Disease (0.49), Diabetes (0.56), Hypertension (0.54), and Obesity (0.51). This highlights LLaMa3's advantage in this dataset when using the RAG-based example setting.

- Impact of RAG: As with CT-Pub, both models show improved performance in the Three Shot + RAG-based Example setting compared to Zero Shot and Three Shot + Fixed Example settings. The RAG approach, which retrieves the most similar trials, consistently enhances model performance across trial groups.

- Model Comparison: While GPT-4o generally performs well, particularly in the Diabetes group (0.53), LLaMa3 surpasses it in all trial groups under the RAG-based example setting, which differs from the results seen in the CT-Pub dataset. This suggests that LLaMa3 may have a performance advantage in the CT-Repo dataset with dynamic retrieval.

- Trial Group Variations: The Diabetes and Hypertension groups see the highest F1 scores, with Diabetes reaching 0.56 for LLaMa3 and 0.53 for GPT-4o, indicating that these trial types may have more consistent or predictable baseline features. In contrast, the Cancer group shows lower scores, particularly in the Zero Shot setting, where GPT-4o scores just 0.29.

Overall, the results in the CT-Repo dataset emphasize the effectiveness of the RAG-based example setting in boosting model performance, with LLaMa3 outperforming GPT-4o across all trial groups, particularly in the Three Shot + RAG setting.

E.3 Human-in-the-loop Evaluation of GPT-4o as a Judge

To assess GPT-4o's accuracy as an evaluator, we engaged clinical domain experts to identify matched pairs for 100 CT studies in the CT-Pub dataset. Focusing on GPT-4o's three-shot (fixed example) candidate responses, the experts used the same information and criteria as GPT-4o.

Table 7: Mean of Cohen's Kappa Score for each evaluator pair across all 100 CT-Pub studies

Evaluator Pair	**Mean Kappa Score**
Human 1 and Human 2	0.870561
Human 1 and Human 3	0.832767
Human 2 and Human 3	0.831810
Human 1 and GPT-4o	0.847636
Human 2 and GPT-4o	0.816234
Human 3 and GPT-4o	0.783869

We developed a web tool to collect and store their responses. We then compared the responses for the matched pairs from the human evaluators

and GPT-4o, creating an inter-rater agreement table and calculating pairwise Cohen's Kappa statistics. Cohen's Kappa measures the agreement level between two raters classifying items into categories. Our findings, presented in Table 7, show high agreement between the human evaluators and GPT-4o, underscoring GPT-4o's reliability in identifying nuanced feature similarities. The relevant code is available in the GitHub.

F Artifact Licenses

- ClinicalTrials.gov Data - Public and free to use https://clinicaltrials.gov/about-site/terms-conditions
- Meta Llama 3 - Meta Lllama 3 community license: https://www.llama.com/llama3/license/ (use of existing artifact(s) was consistent is their intended use)
- OpenAI GPT-4o - https://openai.com/policies/row-terms-of-use/
- CT-Pub and CT-Repo dataset - CC0 1.0 Universal