Retrieval-Reasoning Large Language Model-based Synthetic Clinical Trial Generation

Anonymous ACL submission

Abstract

Machine learning (ML) exhibits promise in the clinical domain. However, it is constrained by data scarcity and ethical considerations, as the generation of clinical trials presents significant challenges due to stringent privacy regulations, high costs, and the extended duration required for conducting studies with human participants. Despite the advancements of large language models (LLMs) in general generation tasks, their potential in facilitating the generation of synthetic clinical trials is under-explored. To address this gap, we introduce a novel Retrieval-Reasoning few-shot framework that leverages LLMs to generate artificial yet realistic and diverse clinical trials with binary success/failure labels. Experiments conducted on real clinical trials from the `ClinicalTrials.gov` database demonstrate that our synthetic data can effectively augment real datasets. Furthermore, by fine-tuning a pre-trained model as a binary classifier on synthetic clinical trial datasets, we demonstrate that this augmentation enhances model training for downstream tasks such as trial outcome prediction. Our findings suggest that LLMs for synthetic clinical trial generation hold promise for accelerating clinical research and upholding ethical standards for patient privacy. The code is publicly available at `https://anonymous.4open.science/r/Retrieval_Reasoning_Clinical_Trial_Generation-3EC4`.

1 Introduction

Large Language Models (LLMs) have advanced natural language processing (NLP) and computer vision (CV) by offering great linguistic and contextual understanding (Brown et al., 2020). Their capacity to generate high-quality, contextually relevant content has expanded their application across various domains, from chatbots and content creation to more specialized tasks like data augmentation (Vaswani et al., 2017). One particularly promising application of LLMs is in generating synthetic data, especially in fields where access to real-world data is constrained due to privacy concerns, high costs, or scarcity (Xu et al., 2020). The domain of medicine presents a prime example of these challenges. Clinical trials, which are critical for developing new treatments, drugs, and therapies, generate vast amounts of sensitive data (Bhatt et al., 2021). These trials undergo rigorous testing to ensure the safety and efficacy of medical interventions before public approval. However, the strict regulatory environment, sensitive nature of patient information, and high cost of data collection often impede the gathering of large, high-quality datasets necessary for training machine learning (ML) models (Chen and Asch, 2021). This lack of data poses a significant barrier for medical researchers seeking to leverage ML to improve clinical decision-making (Rajkomar et al., 2019).

With this in mind, synthetic data generated by LLMs offers a compelling solution for clinical trial studies. By simulating realistic clinical trial reports, researchers can create artificial datasets that mimic the complexity and structure of real trials without exposing sensitive patient information (Jordon et al., 2020). These synthetic datasets can be used for model training (e.g., augmentation), testing, and analysis, providing a valuable resource while adhering to strict privacy regulations (Tucker et al., 2020). However, generating synthetic data that faithfully captures the complexities of real clinical data is challenging. Ensuring the quality, validity, and representativeness of the synthetic data remains a critical area of research (Chen et al., 2021b; Wang et al., 2024).

One major challenge is the inherent discrepancy between synthetic and real data. While synthetic data can offer the necessary volume and diversity to train models, it often lacks the complex patterns and subtleties present in real-world datasets (Xu et al., 2020). If the synthetic data does not accurately replicate the real data, models trained on

synthetic datasets alone may perform poorly when applied to real-world clinical scenarios (Tucker et al., 2020). Furthermore, clinical text classification tasks—such as predicting outcomes from trial reports—often require a delicate balance between precision (minimizing false positives) and recall (capturing as many relevant cases as possible) (Rajkomar et al., 2019). Achieving this balance is critical in a medical context, where both false positives and false negatives can carry significant risks for patient safety and treatment efficacy (Chen and Asch, 2021).

To address these challenges, we propose a novel hybrid fine-tuning approach in this work. This method combines synthetic clinical trial data generated by LLMs with real clinical data to fine-tune a BioBERT model, a pre-trained transformer model specialized for biomedical text (Lee et al., 2020). This hybrid approach leverages the strengths of both synthetic and real datasets: synthetic data provides the diversity and volume necessary for robust model training, while real data ensures that the model is grounded in authentic clinical patterns (Jordon et al., 2020). Through this approach, we aim to mitigate the discrepancies between synthetic and real data, enhancing model performance for clinical text classification tasks.

Our contributions are three-fold. First, we developed an efficient pipeline for generating synthetic clinical trials with explicit success/failure labels, addressing a critical need for labeled data in clinical trial research. Second, we demonstrate that hybrid fine-tuning—utilizing a combination of synthetic and real data—yields significant improvements in model performance across key evaluation metrics, compared to models trained solely on either dataset for clinical text classification tasks such as clinical trial outcome prediction. This approach is particularly valuable in scenarios where access to real data is limited, providing a practical solution for advancing machine learning in healthcare while ensuring patient privacy is preserved. Third, we conduct an in-depth analysis of our synthetic data using t-SNE and cosine similarity to assess its quality and similarity to real data. Our findings indicate that the generated synthetic clinical trials serve as an effective augmentation to real trials by expanding the coverage of the data distribution and enhancing diversity, thereby improving overall model robustness.

2 Related Work

Synthetic Data Generation AI-driven synthetic data generation offers a transformative approach to overcoming data scarcity and privacy constraints by creating artificial yet realistic datasets. This technology enables the training of robust machine learning models without the need for extensive real-world data collection. In healthcare, patient information is often stored in electronic health records (EHR) format (Kruse et al., 2017; Wen et al., 2022; Fu et al., 2019a). Research in medicine has been greatly facilitated by the availability of information from electronic health records (Goncalves et al., 2020; Du et al., 2023). MedGAN, an adversarial network model for generating realistic synthetic patient records, has been proposed by Edward Choi and other colleagues. With the help of an autoencoder and generative adversarial networks, medGAN can generate high-dimensional discrete variables (e.g., binary and count features) based on real patient records (Choi et al., 2017). Based on their evaluations of medGAN's performance on a set of diverse tasks reported, including reporting distribution statistics, classification performance (Lu, 2018), and expert review, medGAN exhibits close-to-real-time performance (Bhanot et al., 2021; Fu et al., 2022, 2019b; Chen et al., 2021a; Choi et al., 2017). Using synthetic data can help reduce the regulatory barriers preventing the widespread sharing and integration of patient data across multiple organizations in the past (Eigenschink et al., 2021; Wu et al., 2022). Researchers across the globe would be able to request access to synthetic data from an institution to conduct their own research using the data. Such capabilities can increase both the efficiency and scope of the study as well as reduce the likelihood of biases being introduced into the results (Tucker et al., 2020; Wang and Qu, 2022; Du et al., 2022).

AI for Clinical Trial AI, particularly deep learning methods, holds significant potential for addressing numerous challenges within clinical trials, e.g., clinical trial outcome/approval prediction (Fu et al., 2022, 2023; Lu et al., 2024; Yue and Fu, 2024), duration prediction (Yue et al., 2024a), enrollment success prediction (Yue et al., 2024b), patient dropout rate prediction (Chen et al., 2024), and digital twin simulation (Wang et al., 2024). For example, conducting clinical trials is exceedingly costly (up to hundreds of millions of dollars (Martin et al., 2017)) and time-consuming (averaging

7-11 years (Peto, 1978; Ledford, 2011)), with a low success rate. Factors such as drug ineffectiveness, safety issues, and poorly designed eligibility criteria can lead to trial failures (Friedman et al., 2015). Improved prediction of trial outcomes could help avoid unsuccessful trials and allocate resources more efficiently (Lu et al., 2024, 2022).

3 Method

3.1 Preliminaries and Background

Data Source `ClinicalTrials.gov` is a comprehensive, publicly accessible database maintained by the U.S. National Library of Medicine (NLM) at the National Institutes of Health (NIH), which provides detailed information about clinical trials conducted globally. The database includes studies sponsored by a wide range of organizations, including federal agencies, pharmaceutical companies, universities, and other research institutions. Each entry in `ClinicalTrials.gov` typically contains information about the study's purpose, design, eligibility criteria, locations, and outcomes. Clinical trials span various phases (from early-phase to late-phase trials), therapeutic areas, and interventions, such as drugs, devices, and behavioral treatments. We use the complete base of clinical trials from `ClinicalTrials.gov` to construct our real dataset. These trial reports were originally in the form of XML, and we transform them into string texts, each denoted as S_i. Then the set of all trials in the complete base is denoted as $\mathbb{S}_{\text{total}}$. Thus $\mathbb{S}_{\text{total}} = \{S_i\}_{i=1}^{N_{\text{total}}}$, where $N_{\text{total}} = 494,290$ is the total number of clinical trials in the database. Within $\mathbb{S}_{\text{total}}$, some trials S_j are labeled manually by a team of IQVIA with their binary outcome $y_j \in \{0, 1\}$, of 0 being failure and 1 being success. We denote this set of trials by $\mathbb{S}_{\text{labeled}}$, where $\mathbb{S}_{\text{labeled}} \subsetneq \mathbb{S}_{\text{all}}$, and $\mathbb{X}_j = \{S_j\}_{j=1}^{N_{\text{labeled}}}$ with $N_{\text{labeled}} = 26,768$. We further construct all our datasets from the set of trial vector and label pairs $\{(S_j, y_j)\}_{j=1}^{N_{\text{labeled}}}$.

Few-Shot In-Context Generation with LLM The LLM $\mathcal{M}$ leverages a small number K of examples provided within the input prompt to learn without parameter updates. Formally, given a set of K input-output pairs of text $\{(I_i, O_i)\}_{i=1}^{K}$, and a new input I_{K+1}, the prompt $\mathcal{P}$ = Concat(format(I_1, O_1), format(I_2, O_2), ..., format(I_K, O_K), format(I_{K+1}) is constructed by concatenating these examples followed by I_{K+1}. $\mathcal{M}$ aims to generate the corresponding output O_{K+1} by modeling the conditional probability distribution $P_\theta(O_{K+1} \mid \mathcal{P}) = \prod_{t=1}^{T} P_\theta(o_t \mid \mathcal{P}, o_{<t})$, where $O_{K+1} = (o_1, o_2, \ldots, o_T)$ is the sequence of tokens in the output, and $o_{<t} = (o_1, o_2, \ldots, o_{t-1})$ denotes the tokens generated before the t_{th} token of the output O_{K+1}. In this way, $\mathcal{M}$ computes the probability of each next token o_t conditioned on the prompt $\mathcal{P}$ and the previously generated tokens o_{t-1}. This process effectively performs sequence modeling using the chain rule of probability. $\mathcal{M}$ utilizes its learned representations to identify patterns and relationships within the few-shot examples, enabling it to generalize and produce O_{K+1} that aligns with the task demonstrated in the prompt. This approach capitalizes on the model's capacity to perform meta-learning, adapting to new tasks through in-context information without updating its parameters.

Fine-Tuning for Clinical Trial Outcome Prediction To fine-tune a pretrained machine learning model as a classifier to predict clinical trial binary outcomes, we begin by defining a fine-tuning dataset $\mathbb{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector for the i-th clinical trial sample, and $y_i \in \{0, 1\}$ is the corresponding binary outcome, with 0 denoting failure and 1 denoting success. The goal is to learn a function $f_\theta : \mathbb{R}^d \to \{0, 1\}$, parameterized by θ, that maps the input features to the predicted binary outcomes. The model is trained by minimizing a loss function, typically the binary cross-entropy loss. During inference, for a new trial sample $\mathbf{x}_{\text{new}}$, the model computes the predicted probability $\hat{y} = P_\theta(y = 1 \mid \mathbf{x}_{\text{new}})$, and the final predicted outcome $\hat{y}_{\text{new}} = 1$ if $\hat{y} \geq 0.5$, and 0 otherwise. Thus, the classifier uses the learned parameters θ to predict binary outcomes for clinical trial data based on input features, where the threshold of 0.5 typically determines whether the predicted outcome is a success or failure. The performance of the classifier is usually evaluated using metrics such as accuracy, precision, recall, the area under the receiver operating characteristic curve (AUC-ROC), and the precision-recall area under the curve (PR-AUC).

3.2 Retrieval-Reasoning Few-shot Generation

3.2.1 Overview

We adopt the few-shot in-context generation with LLM approach described in Chapter 3.1, but with several modifications and specifications in the settings. As shown in Figure 1, our pipeline consists

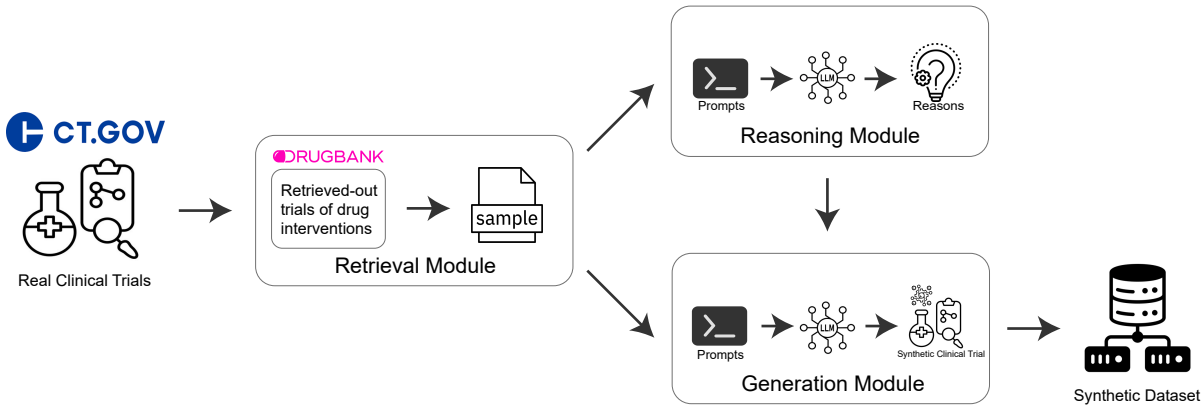

Figure 1: The overall pipeline of retrieval-reasoning clinical trial generation.

of three modules: retrieval module, reasoning module and generation module. For the few-shot example set $\{(S_i, y_i)\}_{i=1}^{K}$, the input S_i here is the string of the clinical trial reports from ClinicalTrials.gov. The output y_i here is the binary outcome of the given trial S_i. Furthermore, the generation setting in our case is to provide prompt $\mathcal{P}$ = Concat(Context Prompt, Reasoning Prompt($\mathcal{R}$), Example Prompt(S_i, y_i), Constraint Prompt, Generation Prompt(y_{K+1}), Diversity Prompt) to $\mathcal{M}$. Here $\mathcal{R}$ is the reasons that lead S_i to y_i, generated by the reasoning module. y_{K+1} is the label we would like the LLM-generated clinical trial to have, to generate S_{K+1}. So here, instead of providing input S_{K+1} and letting the LLM $\mathcal{M}$ generate output y_{K+1}, we are doing the reverse problem. In our settings, ChatGPT-4o-mini is used as LLM $\mathcal{M}$. The number K of few-shot examples we provide for each generation is set to be 3. This pipeline enables the model to adapt its generative behavior by leveraging real clinical trial reports, using the contextual information embedded within the prompt to simulate new clinical trial reports.

3.2.2 Retrieval Module

We first introduce a retrieval module which is designed to make our generation more reasonable and more informative throughout the process. It is designed to extract clinical trials involving drug interventions from the DrugBank database (https://www.drugbank.ca/). With clinical trials of interventions appearing in the drug vocabulary of DrugBank as examples, LLM is more capable of generating clinical trials operating on well-known and recognized drugs instead of treatments that would be unpopular and less informative. This retrieval strategy is beneficial to both the reasoning and generation modules, hence benefiting the whole generation process. Furthermore, the mechanism filters trials based on interventions with at least three positive (success) and three negative (failure) outcome labels. It is designed this way because it is more beneficial for $\mathcal{M}$ to discover reasons based on clinical trials of a specific intervention with the same binary outcome. A complete list of drug names filtered out by retrieval module can be found in Appendix A.2. At the end of this module, we fix a specific intervention and corresponding outcome label for each time, maintaining consistency across the process. As a result, the few-shot example set $\{(S_i, y_i)\}_{i=1}^{3}$ is sampled from the clinical trials corresponding to this list, and is equivalent to $\{(S_i, y)\}_{i=1}^{3}$ such that all 3 S_i shares the same intervention name and the same label y.

3.2.3 Reasoning Module

In this module, we utilize a prompt $\mathcal{P}'$, which consists of context prompt, example prompt with $\{(S_i, y)\}_{i=1}^{3}$ infused, constraint prompt, generation prompt and diversity prompt to ask $\mathcal{M}$ to generate 5 reasons that it summarized from S_i that lead to the outcome label y. See Table 1 for the exact prompts we use in this module.

Context Prompt Establishing a clear background framework is crucial for guiding the model to produce specialized and relevant responses. Positioning the model as a medical expert directs its analytical focus toward the complexities of the medical field. This ensures that the content aligns with clinical scenarios and reflects the technical precision typical of medical language. Providing specific examples and clearly defining the task further enhance the model's ability to generate accurate

Table 1: Prompts used for reasoning module

Category	Prompt
Context	You are now a medical expert in the clinical area. You are given information of a medical intervention, and three clinical trial reports of it, either all successful or all failed. You are asked to analyze these input and write reasons resulting in the trials' success/failure. Your writing style must be consistent within the clinical study. You must ensure that your language is precise, technical, and reasonable.
Example	(Successful/Failed) clinical trial example: ....
Constraint	Your output should strictly follow the following format: 1. (...) 2. (...) 3. (...) 4. (...) 5. (...), with (...) being the reasons you write.
Generation	Write 5 reasons leading (intervention name) to (succeed/fail) in these trials. Be creative and write unique reasons.
Diversity	Can you provide something more diverse compared to the previously generated reasons?

and contextually appropriate responses.

Example Prompt Incorporating real clinical trial reports as few-shot examples is key to improving the model's reasoning abilities. For each generation, the examples $\{(S_i, y)\}_{i=1}^{3}$ with a fixed intervention name and outcome label are randomly drawn and integrated into the prompt as strings. Given that GPT-4o mini has a maximum input capacity of 128K tokens, a check condition is applied to ensure that the total token count of the selected samples stays within this limit. The label y is explicitly stated each time before the associated content to remind the model of the outcome.

Constraint Prompt Imposing structural guidelines enhances the clarity and consistency of the model's output. A predefined format ensures organized and interpretable responses, facilitating comparative analysis and uniformity across outputs. This structure reduces irrelevant content and strengthens the coherence of the generated explanations, aligning them with the desired analytical framework. As a result, the model produces logically sound and contextually relevant responses within the clinical domain.

Generation Prompt Specifying a fixed number of reasons improves the precision and comprehensiveness of the model's output. By setting this limit, the system ensures responses are neither too brief nor overly detailed, balancing the analysis effectively. Encouraging originality in the generated reasons broadens the range of factors considered, enriching the analysis and reducing redundancy. This approach allows the model to capture the multifaceted nature of clinical trials, offering deeper insights for evaluation and decision-making.

Diversity Prompt Promoting diverse reasoning is crucial to avoid repetitive explanations and ensure a thorough exploration of factors influencing trial outcomes. Encouraging variability in the model's responses broadens the scope of analysis, identifying a wider range of influences and perspectives. This diversity-oriented approach enhances the quality and depth of the generated reasons, contributing to a more comprehensive and nuanced understanding of clinical interventions and their outcomes.

3.2.4 Generation Module

Finally, in the generation module, we provide the model with the five generated reasons along with the three clinical trials and instruct it to generate a simulated clinical trial for the fixed intervention and predetermined outcome label. The concatenated prompt contains context prompt, example prompt, constraint prompt, generation prompt, with a similar setting as the corresponding prompts described in the reasoning module. Apart from these prompts, a reasoning prompt is inserted before the example prompt, providing reasons leading to the common outcome of the three examples.

Reasoning Prompt The Reasoning Prompt ensures the model's output is precisely aligned with the clinical scenario by specifying the intervention and outcome y. This prompt directs the model's reasoning toward relevant factors. Incorporating the five previously generated reasons provides a consistent foundation for further explanation, reinforcing logical continuity and coherence.

Combined with retrieval-based selection, and reasoning, this module enables the model to produce plausible synthetic clinical trial reports grounded in real-world data. We generate 3358 synthetic clinical trial reports using GPT-4o-mini API with temperature 1, using a concatenated prompt. Also, we record the intervention names of each generated clinical trial. See Table 2 for the exact prompts we use in this module.

4 Experiments

We conducted a comprehensive set of experiments to evaluate the effectiveness of our generated synthetic clinical trials, both in distribution and the ability to generalize. This included fine-tuning a pre-trained model from the biological domain and employing it as a classifier to predict clinical trial outcomes. Additionally, we visualized the simi-

Table 2: Prompts used for generation module.

Category	Prompt
Context	You are now a medical expert in the clinical area. You are asked to write a report of a successful or failed clinical trial. Your writing style must be consistent within the clinical study. Ensure your language is precise, technical, and formal.
Reasoning	Here are five reasons that could lead to the (success/failure) of clinical trials of (intervention name): ...
Example	(Successful/Failed) clinical trial example: ....
Constraint	Your output style should strictly follow the XML-like format of the provided examples. You cannot simply modify or rewrite them. The intervention name must be (intervention name), and you must refer to these reasons when writing clinical trials.
Generation	Write a report of a (successful/failed) clinical trial of (intervention name). Ensure your language is precise, technical, and formal. Be creative and write unique reports.
Diversity	Can you provide something more diverse compared to the previously generated reports?

larities between the distributions of real and synthetic trials through t-SNE plots, and assessed the diversity within the respective distributions of real and synthetic trials by computing cosine similarity. These analyses provide insight into the alignment and variation between the synthetic and real datasets, demonstrating the potential of synthetic data to mimic real-world clinical trials.

4.1 Fine-Tuning BioBERT

As we recorded the intervention names of each generated clinical trials in the generation module, we created a list of intervention names of our generated clinical trials. We further splitted the IQVIA labeled real dataset $\{(S_j, y_j)\}_{j=1}^{N_{\text{labeled}}}$ as a set of clinical trials with intervention names lies in the list of intervention names of our generated clinical trials, denoted as $\mathbb{A}$, and a set of clinical trials with intervention names that do not lie in the list of intervention names of our generated clinical trials, denoted as $\mathbb{B}$. Note that $\mathbb{A} = \{(S_j, y_j)\}_{j=1}^{N_{\text{labeled}}} \backslash \mathbb{B}$, and $\mathbb{A} \cap \mathbb{B} = \varnothing$. $\mathbb{A}$ and $\mathbb{B}$ are complementary set of each other. $\mathbb{A}$ has a size of $6,056$, and $\mathbb{B}$ has a size of $20,712$. We conducted a two experiments to assess the efficacy of our generated synthetic clinical trials by fine-tuning BioBERT under three distinct settings:

- **Synthetic-Only:** Fine-tuning using solely synthetic data.
- **Real-Only:** Fine-tuning using only real clinical trial data.
- **Hybrid:** Fine-tuning using a combination of synthetic and real data.

For the in-distribution performance test of our synthetic trials, we constructed both the validation and test sets from $\mathbb{A}$, partitioned according to a 60%-20%-20% split. So that the real training set has 3,633 samples, and the validation and test sets consist of 1,211 and 1,212 real samples respectively. As a result, the synthetic-only training set consists 3,358 synthetic samples; the real-only training set consists 3,633 synthetic samples; the hybrid training set consists 6,991 real-synthetic combined samples.

In the ratio experiment, we control the size of the training sets to be 3,358, and the validation & testing sets to be 1,349. The training set have 6 compositions with different ratios of real and synthetic data, including 100% synthetic, 80% synthetic + 20% real, 60% synthetic + 40% real, 40% synthetic + 60% real, 20% synthetic + 80% real, and 100% real. The validation and testing sets are sampled from $\mathbb{A}$.

While for the generalization experiment, this time we constructed both the validation and test sets from $\mathbb{B}$, and the training sets from $\mathbb{A}$. Since $\mathbb{B}$ is an imbalanced dataset, we downsampled the major class to make it class-balanced. As a result, the validation and test sets both consists 7,546 samples from $\mathbb{B}$. The synthetic-only training set is still the 3,358 samples we generated. The real-only training set is the complete set $\mathbb{A}$ of 6,056 samples. The hybrid training set in this case is the combined set of 9,414 samples.

We used the pre-trained BioBERT model as the baseline method, and employed accuracy, precision, recall, ROC-AUC and PR-AUC as evaluation metrics to assess the quality of the generated synthetic clinical trials, and our hybrid training approach. These metrics provide a comprehensive comparison of model performance across different settings.

4.1.1 Implementation Details

Given the specific structure of our dataset, the labels are often embedded within the individual clinical trial texts. Specifically, the text typically includes the “overall status”, which is either “completed" or “terminated”. Additionally, in terminated trials, the “why stop" section details the reasons for the study’s termination. To avoid data leakage during the training and testing phases, we systematically remove these portions from the clinical trial texts throughout the experiment.

Moreover, we observed instances where the

Table 3: Performance metrics of BioBERT with different fine-tuning strategies tested on a real dataset. The upper section is the in-distribution test, the middle section is the ratio test, and the lower section is the generalization test. For each metric, we highlight the best method **in bold**.

Testing	Fine-Tuning	Accuracy	Precision	Recall	ROC-AUC	PR-AUC
In-distribution	No Fine-Tuning	0.527 ± 0.009	0.526 ± 0.008	**0.937 ± 0.090**	0.517 ± 0.035	0.531 ± 0.029
	Synthetic	0.534 ± 0.010	0.532 ± 0.005	0.873 ± 0.048	0.559 ± 0.009	0.584 ± 0.006
	Real	0.633 ± 0.003	**0.677 ± 0.006**	0.563 ± 0.009	0.689 ± 0.009	0.723 ± 0.013
	Hybrid	**0.642 ± 0.012**	0.662 ± 0.023	0.642 ± 0.027	**0.698 ± 0.012**	**0.725 ± 0.016**
Ratio	100% Synthetic	0.545 ± 0.009	0.543 ± 0.010	**0.825 ± 0.091**	0.566 ± 0.015	0.589 ± 0.017
	80% Syn + 20% Real	0.586 ± 0.015	0.594 ± 0.027	0.665 ± 0.068	0.620 ± 0.010	0.638 ± 0.011
	60% Syn + 40% Real	0.608 ± 0.013	0.620 ± 0.005	0.640 ± 0.082	0.648 ± 0.019	0.667 ± 0.012
	40% Syn + 60% Real	0.620 ± 0.016	0.623 ± 0.018	0.689 ± 0.092	0.670 ± 0.026	0.689 ± 0.028
	20% Syn + 80% Real	0.629 ± 0.010	0.642 ± 0.041	0.678 ± 0.091	0.692 ± 0.008	0.708 ± 0.001
	100% Real	**0.638 ± 0.011**	**0.695 ± 0.002**	0.542 ± 0.040	**0.702 ± 0.009**	**0.728 ± 0.005**
Generalization	No Fine-Tuning	0.504 ± 0.005	0.502 ± 0.003	**0.939 ± 0.086**	0.512 ± 0.022	0.510 ± 0.020
	Synthetic	0.530 ± 0.005	0.518 ± 0.004	0.883 ± 0.042	0.589 ± 0.020	0.574 ± 0.019
	Real	0.645 ± 0.008	0.616 ± 0.007	0.771 ± 0.034	0.709 ± 0.012	**0.697 ± 0.010**
	Hybrid	**0.658 ± 0.001**	**0.660 ± 0.006**	0.652 ± 0.020	**0.711 ± 0.001**	0.694 ± 0.004

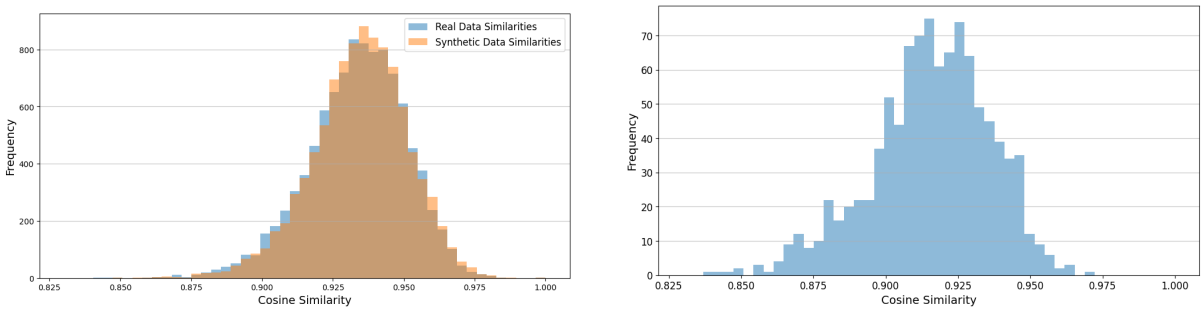

Figure 2: Distributions of cosine similarities between pairs of embeddings (a) within distributions (b) across distributions.

LLM exhibits hallucinations by including label terms such as “successful" or “failed" directly in the generated text. To further mitigate data leakage, we also exclude these terms when processing synthetic trials. The model’s learning rate was set to 1e-5, with a batch size of 8. We fine-tuned BioBERT under each experimental setting, training the model for 7 epochs. Each experiment is repeated 3 times with random seeds 40, 41, and 42 for reproducibility, and the results are reported by calculating the mean and standard deviation for each metric. BioBERT has 110M parameters, and we do these experiments on an Nvidia H100 80GB GPU, the time consumption for each run is around 25 minutes.

4.1.2 Results & Analysis

As Table 3 suggests, for the in-distribution test, without fine-tuning, the BioBERT model demonstrates nearly random performance, characterized by high recall but low precision, suggesting a tendency to over-predict the positive class. Fine-tuning on synthetic data leads to improvements across all evaluation metrics compared to the base model without fine-tuning, for both in-distribution and generalized tests. The model becomes more balanced, achieving a better trade-off between precision and recall, which indicates that the synthetic data contributes valuable information for the task. Fine-tuning on real data yields higher precision and ROC-AUC, suggesting that real data offers critical features for more precise classification. However, this comes at the expense of recall, indicating that the model becomes more conservative, potentially overlooking some positive cases. In contrast, hybrid fine-tuning, which utilizes both synthetic and real data, results in the best overall in-distribution performance, achieving the highest accuracy, ROC-AUC, and PR-AUC, and a compatible performance with real-only finetuning in the generalized test. This suggests that the combination of synthetic and real data enhances the model’s generalization capabilities by leveraging the complementary strengths of both datasets—diversity from the synthetic data and task-specific knowledge from the real data.

The ratio test results in the middle section of Table 3 indicate that increasing the proportion of real data in the training set consistently enhances per-

formance across all metrics. As more real data is incorporated, metrics like accuracy, precision, and ROC-AUC steadily improve. This trend suggests that real data is more effective for fine-tuning the model due to its higher quality compared to synthetic data. While synthetic data provides valuable initial training, the superior performance with more real data underscores its critical role in refining the model's predictive capabilities.

The generalization test results in the lower section show that models fine-tuned on synthetic data alone still generalize well to clinical trials with unseen interventions, outperforming the baseline. However, combining synthetic and real data leads to even better performance, closely matching the results of real-only fine-tuning. This suggests that synthetic data enhances the model's ability to generalize by introducing diverse examples, but real data remains essential for achieving optimal predictive accuracy in new, unseen scenarios.

4.2 Visualizations

We further assessed the diversity between the real and synthetic datasets through cosine similarity histogram comparisons (see Figure 2) by sampling 10,000 pairs of real-real, synthetic-synthetic and real-synthetic samples. Additionally, we analyzed the characteristics of the real and synthetic clinical trials by visualizing their distributions using a t-SNE plot (see Figure 3).

The cosine similarity (see Figure 2) histograms suggests that our generated dataset is consistent within its own distribution, which is comparable to the real dataset. The wider spread of similarities across-distributions indicates that some synthetic samples are closer to the real data (higher similarity scores), while others are more distant compared to the internal similarities within each dataset. This suggests that the distributions of our generated dataset is distinct to the real dataset, providing more diversity to training during hybrid training approach.

The t-SNE visualization (see Figure 3) demonstrates that the real and synthetic datasets form two distinct clusters, characterized by a clear separation. This separation indicates significant differences in the underlying distributions of the real and synthetic data. The minimal overlap between these clusters suggests that the synthetic clinical trials are not exact replicas of the real ones, but rather complement them by introducing additional diversity. This diversity can enhance model fine-tuning by providing a broader range of data variations. Moreover, the synthetic data cluster exhibits a wider spread compared to the real data cluster, indicating a potentially greater degree of internal diversity within the synthetic dataset, which may further contribute to more robust and generalized model training outcomes.

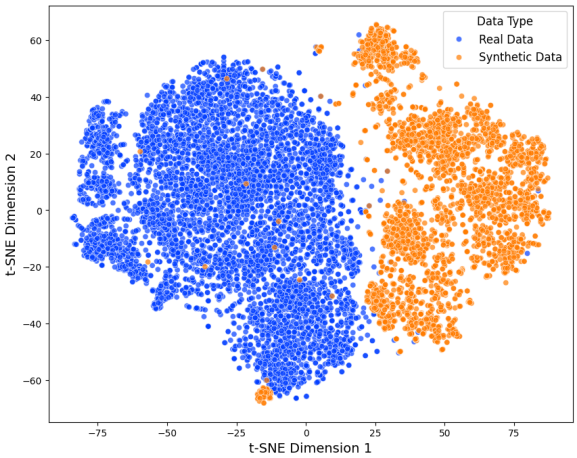

Figure 3: t-SNE visualization of real vs synthetic trials.

5 Conclusion

This paper introduced a pioneering framework that leverages retrieval-reasoning large language models (LLMs) to address the critical need for synthetic clinical trial generation. By generating artificial yet highly realistic clinical trials marked with binary outcomes, we not only mitigate the issues associated with data scarcity but also comply with stringent privacy regulations that govern the use of real-world patient data. Our empirical evaluation has demonstrated that the synthetic datasets produced by our framework can effectively augment existing real-world datasets, thereby enhancing the robustness and performance of machine learning models trained for clinical outcomes prediction. Furthermore, these synthetic trials facilitate faster model training and evaluation cycles, which can significantly accelerate the pace of clinical research. The successful application of our method underscores the potential of LLMs in advancing healthcare informatics, particularly in scenarios where traditional methods of data collection are constrained by practical or ethical limitations. In future work, we propose exploring the integration of multimodal information and extending the scope of synthetic data generation to include more complex clinical scenarios and endpoints.

Limitations

The quality of synthetic data generation in this study is intrinsically linked to the capabilities and limitations of the employed LLM API, making the generated data susceptible to the inherent biases and inaccuracies present in the LLM itself. In this work, we proposed a method for generating synthetic clinical trials. While these generated trials serve as valuable augmentations to the limited available real clinical trials, a significant distributional gap remains evident in the t-SNE visualization, indicating divergence between real and synthetic datasets. Additionally, quality disparities are observable through ratio tests, underscoring the need for further refinement to reduce this distribution gap and enhance the fidelity of synthetic clinical trials. Moreover, this study focuses exclusively on clinical trials involving drug interventions, representing only a narrow subset of the broader clinical trial database. Consequently, trials with alternative intervention types remain unexplored in this work.

Ethical Considerations

Potential Risks The use of LLMs for clinical data generation may introduce biases from the training data, which could lead to skewed or inaccurate synthetic trials that misrepresent certain patient populations or treatment outcomes. Finally, reliance on synthetic data might impact the robustness of downstream clinical research if real-world complexities are not adequately captured.

Use of AI Assistants We employed ChatGPT as a writing assistant for polishing the text after our initial composition, and for providing comments for our code.

References

Karan Bhanot, Miao Qi, John S Erickson, Isabelle Guyon, and Kristin P Bennett. 2021. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165.

Deepak L Bhatt et al. 2021. Clinical trials in the era of artificial intelligence. *Nature Reviews Drug Discovery*, 20(8):600–602.

Tom Brown et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jintai Chen, Yaojun Hu, Yue Wang, Yingzhou Lu, Xu Cao, Miao Lin, Hongxia Xu, Jian Wu, Cao Xiao, Jimeng Sun, et al. 2024. Trialbench: Multi-modal artificial intelligence-ready clinical trial datasets. *arXiv preprint arXiv:2407.00631*.

Jonathan H Chen and Steven M Asch. 2021. Ethical and regulatory considerations for using artificial intelligence in clinical trials. *JAMA*, 326(2):107–108.

Lulu Chen, Yingzhou Lu, Chiung-Ting Wu, Robert Clarke, Guoqiang Yu, Jennifer E Van Eyk, David M Herrington, and Yue Wang. 2021a. Data-driven detection of subtype-specific differentially expressed genes. *Scientific reports*, 11(1):1–12.

Rui Tao Chen et al. 2021b. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*. PMLR.

Dongping Du, Saurabh Bhardwaj, Sarah J Parker, Zuolin Cheng, Zhen Zhang, Yingzhou Lu, Jennifer E Van Eyk, Guoqiang Yu, Robert Clarke, David M Herrington, et al. 2023. Abds: tool suite for analyzing biologically diverse samples. *bioRxiv*, pages 2023–07.

Yuanqi Du, Tianfan Fu, Jimeng Sun, and Shengchao Liu. 2022. Molgensurvey: A systematic survey in machine learning models for molecule design. *arXiv preprint arXiv:2203.14500*.

Peter Eigenschink, Stefan Vamosi, Ralf Vamosi, Chang Sun, Thomas Reutterer, and Klaudius Kalcher. 2021. Deep generative models for synthetic data. *ACM Computing Surveys*.

Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. 2015. *Fundamentals of clinical trials*. Springer.

Tianfan Fu, Tian Gao, Cao Xiao, Tengfei Ma, and Jimeng Sun. 2019a. Pearl: Prototype learning via rule learning. In *ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 223–232.

Tianfan Fu, Trong Nghia Hoang, Cao Xiao, and Jimeng Sun. 2019b. DDL: Deep dictionary learning for predictive phenotyping. In *International Joint Conference on Artificial Intelligence*.

Tianfan Fu, Kexin Huang, and Jimeng Sun. 2023. Automated prediction of clinical trial outcome. US Patent App. 17/749,065.

Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. 2022. HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*, 3(4):100445.

Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. 2020. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20(1):1–40.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. 2020. Synthetic data for clinical research. *Nature Biomedical Engineering*, 4(9):870–872.

Clemens Scott Kruse, Brenna Smith, Hannah Vanderlinden, and Alexandra Nealand. 2017. Security techniques for the electronic health records. *Journal of medical systems*, 41(8):1–9.

Heidi Ledford. 2011. 4 ways to fix the clinical trial: clinical trials are crumbling under modern economic and scientific pressures. nature looks at ways they might be saved. *Nature*, 477(7366):526–529.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yingzhou Lu. 2018. *Multi-omics Data Integration for Identifying Disease Specific Biological Pathways*. Ph.D. thesis, Virginia Tech.

Yingzhou Lu, Tianyi Chen, Nan Hao, Capucine Van Rechem, Jintai Chen, and Tianfan Fu. 2024. Uncertainty quantification and interpretability for clinical trial approval prediction. *Health Data Science*, 4:0126.

Yingzhou Lu, Chiung-Ting Wu, Sarah J Parker, Zuolin Cheng, Georgia Saylor, Jennifer E Van Eyk, Guoqiang Yu, Robert Clarke, David M Herrington, and Yue Wang. 2022. Cot: an efficient and accurate method for detecting marker genes among many subtypes. *Bioinformatics Advances*, 2(1):vbac037.

Linda Martin, Melissa Hutchens, Conrad Hawkins, and Alaina Radnov. 2017. How much do clinical trials cost. *Nat Rev Drug Discov*, 16(6):381–382.

Richard Peto. 1978. Clinical trial methodology. *Nature*, 272(5648):15–16.

Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. 2019. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358.

Allan Tucker et al. 2020. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1):1–13.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*.

Ruyang Wang and Xiaodong Qu. 2022. Eeg daydreaming, a machine learning approach to detect daydreaming activities. In *Augmented Cognition: International Conference*. Springer.

Yue Wang, Yingzhou Lu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Honghao Gao, and Jian Wu. 2024. Twin-gpt: Digital twins for clinical trials via large language model. *arXiv preprint arXiv:2404.01273*.

Qianlong Wen, Zhongyu Ouyang, Jianfei Zhang, Yiyue Qian, Yanfang Ye, and Chuxu Zhang. 2022. Disentangled dynamic heterogeneous graph learning for opioid overdose prediction. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Chiung-Ting Wu, Minjie Shen, Dongping Du, Zuolin Cheng, Sarah J Parker, Yingzhou Lu, Jennifer E Van Eyk, Guoqiang Yu, Robert Clarke, David M Herrington, et al. 2022. Cosbin: cosine score-based iterative normalization of biologically diverse samples. *Bioinformatics Advances*, 2(1):vbac076.

Lei Xu et al. 2020. Generating synthetic data for medical research using generative adversarial networks. *Nature Communications*, 11(1):1–7.

Ling Yue and Tianfan Fu. 2024. Ct-agent: Clinical trial multi-agent with large language model-based reasoning. *arXiv preprint arXiv:2404.14777*.

Ling Yue, Jonathan Li, Md Zabirul Islam, Bolun Xia, Tianfan Fu, and Jintai Chen. 2024a. TrialDura: Hierarchical attention transformer for interpretable clinical trial duration prediction. *arXiv preprint arXiv:2404.13235*.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024b. Trialenroll: Predicting clinical trial enrollment success with deep & cross network and large language models. *arXiv preprint arXiv:2407.13115*.