Specific aims

Background and significance: The process of matching patients with appropriate clinical trials is a crucial step in the advancement of medical research and the delivery of optimal care. When patients are correctly matched with relevant clinical trials, patients might gain access to new, potentially life-saving treatments that are not yet available to the general public; on the other hand, considering that a clinical trial has a median estimate of US$19 million per trial, filling a clinical trial with appropriate participants can lead to cost savings, as prolonged or inappropriate trials can be expensive. It is crucial to identify solutions that guarantee precise and streamlined patient-trial matching. Nevertheless, contemporary methodologies encounter inherent challenges, encompassing data standardization, ethical deliberations, lack of cross-institutional collaboration features, and the absence of seamless integration between Electronic Health Records (EHRs) and clinical trial criteria. While recent studies have introduced solutions to these issues using black-boxed embedding matching, the effectiveness of AI-powered clinical trial matching services still faces challenges, particularly when facing the complexities of merging information from EHRs with the criteria outlined in clinical trials.

Proposed Innovation: This project will explore leveraging large language models (LLMs) to enhance the compatibility between EHRs and clinical trial descriptions, enabling a privacy-first and more accurate patient-trial matching. By leveraging current advances in natural language processing, new models are well suited to interpreting, understanding, and aligning the varied terminologies and ontologies found in EHRs with clinical trial inclusion and exclusion criteria. This improved integration is expected to simplify the matching procedure and heighten the precision in pinpointing appropriate patient trials while enabling collaborative tools among researchers, clinicians, and trial coordinators. As a result, LLM integration in clinical trial matching stands to be transformative, as it will enable improved patient results and advance medical research.

Preliminary Work: Leveraging insights from a wide range of applications, our preliminary work targeted the emerging potential of LLMs in clinical text mining, especially within the confines of biological named entity recognition and relation extraction. Initial results reveal that direct LLM use underperforms and raises data privacy issues related to patient information data handling. In response, we developed a unique training approach, leveraging LLMs to produce and label synthetic data sets while optimizing a locally trained model. Our preliminary work resulted in an improvement of the F1-score from 23.37% to 63.99% for named entity recognition and 75.86% to 83.59% for relation extraction. Showing that, when integrated effectively, LLMs can significantly enhance the accuracy and efficiency of clinical text-mining, all while presenting a promising venue for data privacy through effective dataset augmentation in patient-clinical trial matching.

Approach: In Phase I, we propose to further leverage the unique capabilities of LLMs for privacy-aware data augmentation and apply this to enhance cross-institutional clinical trial design through the following aims:

In Aim 1 Privacy-Aware Augmentation Techniques (Months 0-3), our team will leverage LLMs and limited seed data samples labeled by experts to generate a comprehensive synthetic dataset and ensure the data is reliable and adherent to stringent privacy standards, allowing for cross-institutional collaboration.

In Aim 2 Patient-Clinical Trial Matching Precision (Months 3-6), we will focus on the creation of a privacy-focused classification framework based on a Memory and Highway network approach that matches patients' EHR to clinical trials' inclusion and exclusion criteria effectively and ensures the matching process is finely tuned to optimize their potential for enrollment.

At the onset of Phase I, our team will develop a model prototype that provides a hands-on tool to critically assess and validate the efficiency, accuracy, and reliability of our patient-trial matching system using six different stroke clinical trials, namely NCT03263117, NCT03496883, NCT03545607, NCT03735979, NCT03805308, NCT03876457, using ClinicalTrials.gov as our source. Our proposed Phase I work is also tailored to give us a better understanding of the cross-collaboration needs of researchers, clinicians, and trial coordinators when undergoing a clinical trial, setting a robust foundation for our Phase II endeavors.

Team and long-term goal. This project brings together investigators with complementary expertise in Data Mining (Costilla), Machine Learning Automation (Hu), Biomedical Informatics (Liu), and Natural Language Processing (Xu) and collaborators for model validation and clinical trial implementation (Jiang). The successful completion of our phase I work will lay the foundation for a tool that sets new standards in clinical trial design to facilitate collaboration among medical professionals to enhance their ability to connect patients to potentially life-altering treatments and drive transformative outcomes in patient care and clinical research.

Research strategy

A. Significance

The landscape of clinical trials is rapidly evolving, with an impressive 19.1% annualized growth rate from 2017 to 2022, resulting in 462,645 registered studies, as reported by the US National Library of Medicine [1–3]. Those participating in these trials gain potential access to cutting-edge treatments, often unavailable in standard medical practice, and play a pivotal role in advancing medical science [4–6]. Clinical trials, while serving as essential conduits for medical advancements, come with significant financial implications as well. With an expenditure of US$19 million per trial or US$41,000 per patient [7], the stakes are high to ensure each participant is appropriately matched to a given trial; however, finding an ideal candidate for a clinical trial can be a complex and time-consuming task [8]. By 2022, records from the clinicaltrials.gov database indicate that 14.5% of interventional trials faced premature termination [9]. Moreover, only 7.9% of new drug applications resulted in an actual formal registration [10]. The financial impact is particularly more severe in later-phase trials, such as Phase 2 and 3, due to their larger sample sizes and prior investments [11,12]. Thus, optimizing trial design is key to the study's success, efficient resource use, and ethical participant involvement.

Current advances in artificial intelligence mostly focus on predicting trial outcomes [13–16] rather than on clinical trial design. Research studies centered on forecasting the success of clinical trials use machine learning techniques, drawing from structured and unstructured data derived from clinical studies. Such research predominantly aims at predicting the premature termination of clinical trials by integrating data about trial attributes with unstructured information. On the other hand, the limited exploration of a collaborative-focused design of clinical trials [17,18] still faces three crucial challenges: **(1)** the complexities of ensuring ethical practices in both patient involvement and data handling, **(2)** limited access to comprehensive data, and **(3)** the lack of uniform standards in data collection and interpretation [19–21]. As seen in Fig. 1 (top), a significant challenge is found in the mismatch of two key data components **(i)** the diversity of terminologies and ontologies found in patients' Electronic Health Records (EHR) and **(ii)** the intricate inclusion and exclusion Elegibility Criteria (EC) of clinical trials.

The evolution of large language models (LLMs) [22] has ushered in a wave of applications in healthcare. These models have showcased unmatched aptitude in diverse scenarios, from answering complex patient queries [23] to composing comprehensive medical notes [24]. Yet, when steered toward the intricate process of clinical trial design, their exact potential becomes a subject of investigation. Moreover, initial observations indicate that there might be a better solution than a straightforward implementation of LLMs, revealing performance deficits and triggering concerns over patient data privacy when interfacing with LLM APIs [25]. Despite the evident promise and transformative power of LLMs across different sectors, their application in clinical trial design requires a nuanced approach. Solutions tailored for clinical trial design should facilitate cross-institutional collaboration, efficiently navigate extensive medical lexicons, and uphold ethical standards, particularly when handling sensitive patient data. To truly harness LLM's potential, a balance must be struck between optimizing its language processing capabilities while ensuring strict data protection standards. This complex task underscores the need for a collaborative effort among data scientists, medical professionals, and ethical committees to create solutions where LLMs can function efficiently while upholding patient trust and the overarching goals of medical research.`

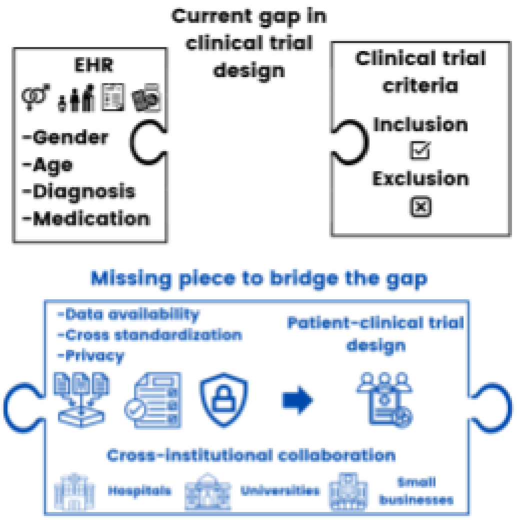


Fig 1. (Top) current gap in clinical trial design. (Bottom) Proposed innovation leveraging LLMs.

To address the challenges of deploying LLMs in clinical trial design, the work of our team at AI POW LLC prioritizes three methods to address the challenges mentioned above: **(1)** proper data de-identification and the need for only a limited amount of human-labeled examples as seeds, **(2)** the creation appropriate prompts to guide LLMs data augmentation task and verifying that it generates reliable datasets, and **(3)** the standardization between EHR data and clinical trial inclusion and exclusion criteria. We want to highlight that our Phase I proposal centers on using desensitized, rather than raw, patient data to guide LLMs to enhance their reliability in zero-shot tasks while minimizing potential privacy breaches, ensuring a privacy-aware cross-institutional collaboration. At the same time, our preliminary work has shown promising results that give us an early indication that our approach can be successfully implemented in clinical trial design. Particularly, our prior work excelled in extracting structured insights from unstructured healthcare records in biological named entity recognition (NER) and relation extraction (RE). By preserving the intrinsic value of the data, we have a high level of confidence that LLMs can be properly prompt-guided to produce high patient data quality.

Finally, in the broader landscape of medical research, the effective matching of patients to clinical trials is not just a matter of scientific precision but also of resource optimization, cost-efficiency, and enhancing public trust. Utilizing LLMs for privacy-conscious data augmentation provides the comprehensive computational tools necessary to revolutionize the design of clinical trials for better patient-trial matching. Beyond ensuring the success of individual trials, our approach can significantly streamline resources and reduce unnecessary financial expenditures. In an era marked by continuous concerns over data security, this approach presents an innovative solution that maintains the integrity of patient data while harnessing the unmatched capabilities of LLMs. By striking this delicate balance, we not only enhance the efficiency of medical research but also fortify public trust, setting the stage for a more collaborative and inclusive future in clinical trials.

B. Innovation

Clinical trials, which are pivotal in the advancement of innovative disease treatments, frequently encounter setbacks, with many being unable to start due to overwhelming recruitment obstacles [8]. The emergence of automated patient-trial matching introduces a groundbreaking and promising solution, poised to revolutionize the intricacies of cross-institutional collaboration for clinical trial design. The essence of this proposed approach lies in identifying eligible patients for clinical trials based on their EHR data and trial EC, which encompass both inclusion and exclusion criteria. If data limitations are solved, the problem can be framed as a classification problem with a given input consisting of a patient’s complete EHR data and a single trial’s EC, the output can be classified as either match, mismatch, or unknown.

The specific innovations of our proposed approach include the following:

Privacy-aware Data Augmentation: Data augmentation in natural language processing (NLP) traditionally involves diverse transformations of text data to bolster model training. In patient-trial matching, the challenge lies in the limited richness of the available training data. To address this, we introduce an effective data augmentation method designed to generate a more varied dataset, allowing machine learning models to better capture the intricacies of patient and eligibility criteria information. However, the vast potential of big data in healthcare is accompanied by significant privacy concerns. These concerns range from the ethical implications of data usage to the technical challenges of data de-identification and linking patient data from different sources. Our proposal's innovation lies in introducing privacy-conscious augmentation techniques using open-source components such as BERT [26]. The primary goal of these techniques is to enhance the accuracy of patient-trial match identification while ensuring no leakage of private-source data, making possible collaboration from multiple institutions to safely share data and models for comprehensive clinical trial design.

Leveraging LLMs for EHRs and Clinical Trial Descriptions: Our second innovation centers around the potential of LLMs to enhance the standardization and interoperability between EHRs and clinical trial descriptions. By utilizing the advanced natural language generation capabilities of LLMs, the approach aims to improve patient-trial matching by homogenizing the diverse terminologies and ontologies present in both EHRs and clinical trial criteria. The proposed approach not only enhances the matching process but also ensures the security and confidentiality of sensitive patient data.

Holistic Framework for Enhanced Medical Research: Finally, our proposed combined work presents a holistic framework where LLMs address multiple challenges in medical research – from text mining and data extraction to collaborative work in clinical trial design. The dual focus ensures efficient information extraction from clinical texts and optimizes patient-trial matching, all while maintaining strict adherence to data privacy.

C. Approach

Phase I Overview. This proposal will allow our team to perform a feasibility study of patient-clinical trial matching leveraging LLMs. In this phase, we will leverage our long-term partnerships, experience, and expertise gained from developing a framework for synthetic data generation using LLMs to help in clinical text mining and clinical trial design.

Preliminary Work:

Considering privacy and confidentiality of patient information are of utmost importance, AI POW LLC has developed a mechanism to ensure robust privacy protections that prevent unauthorized access to sensitive information. In our preliminary work, we developed a framework for measuring and improving the reliability of LLM for zero-shot tasks aimed at Biomedical NER and Biomedical RE while mitigating the privacy risk.

First, to assess the zero-shot performance of current LLM models for healthcare tasks, we conducted experiments on Generative Pre-trained Transformer (GPT) to investigate its ability to extract structured information from unstructured healthcare texts, specifically for biological NER and RE tasks. Our findings suggest that GPT directly only yields poor performance compared to SOTA models trained on the dataset for precision (P), Recall (R), and F-1 score (F). This result highlights that while GPT has demonstrated impressive inference and reasoning abilities in various classic natural language understanding (NLU) tasks, it is not adequate to apply GPT alone to healthcare tasks since it doesn't ensure both the required performance and the privacy requirements needed for cross-institutional collaboration for this domain [27,28].

Design: Our preliminary study was geared towards assessing the efficacy of LLMs in creating a large volume of superior synthetic data with labels, using LLMs, and fine-tuning a local model for Biomedical NER and RE. As illustrated in Fig. 2, our primary work involved the creation of an innovative training paradigm to address the challenges of using LLMs for healthcare tasks.

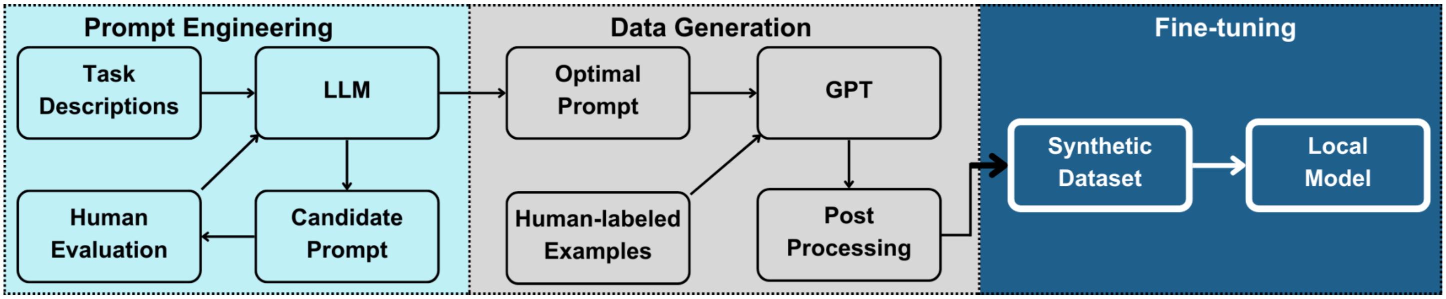
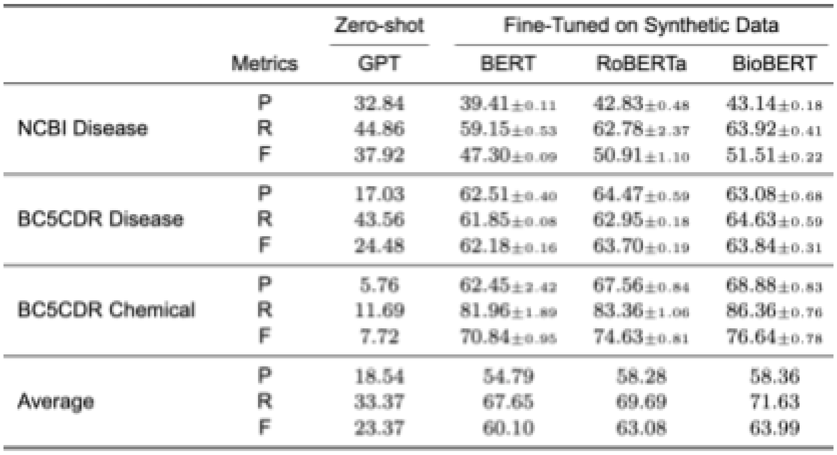


Figure 2: An overview of the workflow for synthetic data generation using GPT.

Contrary to the direct application of LLMs in a zero-shot setting, we generated a substantial amount of synthetic data with labels using LLMs. To enhance the quality and diversity of the synthetic data, we utilized a limited number of human-labeled examples as seeds and developed suitable prompts to guide LLMs in creating a range of examples with diverse sentence structures and linguistic patterns. A post-processing step was employed to eliminate low-quality or duplicated samples generated by LLMs. Finally, we used synthetic data to fine-tune a local pre-trained language model. **<u>This design is particularly beneficial for multi-center clinical trials, where synthetic data can be easily shared among different local sites. Each of these sites can then use this data to fine-tune their individual models. Given that online LLMs like GPT4 are not sharable, this approach ensures that the performance of all sites can be significantly improved.</u>** As demonstrated in the Results section below, our experiments on four representative datasets showed that our proposed pipeline substantially improved the performance of the local model compared to the zero-shot performance of LLMs. Moreover, the approach

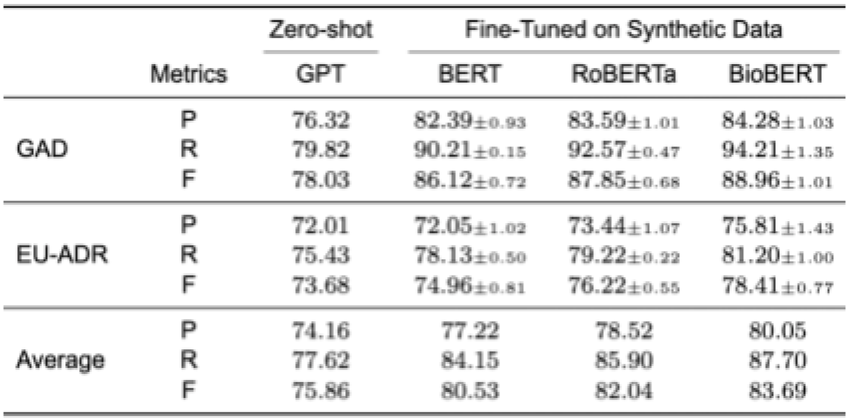


		Zero-shot	Fine-Tuned on Synthetic Data		
	Metrics	GPT	BERT	RoBERTa	BioBERT
NCBI Disease	P	32.84	$39.41_{\pm 0.11}$	$42.83_{\pm 0.48}$	$43.14_{\pm 0.18}$
	R	44.86	$59.15_{\pm 0.53}$	$62.78_{\pm 2.37}$	$63.92_{\pm 0.41}$
	F	37.92	$47.30_{\pm 0.09}$	$50.91_{\pm 1.10}$	$51.51_{\pm 0.22}$
BC5CDR Disease	P	17.03	$62.51_{\pm 0.40}$	$64.47_{\pm 0.59}$	$63.08_{\pm 0.68}$
	R	43.56	$61.85_{\pm 0.08}$	$62.95_{\pm 0.18}$	$64.63_{\pm 0.59}$
	F	24.48	$62.18_{\pm 0.16}$	$63.70_{\pm 0.19}$	$63.84_{\pm 0.31}$
BC5CDR Chemical	P	5.76	$62.45_{\pm 2.42}$	$67.56_{\pm 0.84}$	$68.88_{\pm 0.83}$
	R	11.69	$81.96_{\pm 1.89}$	$83.36_{\pm 1.06}$	$86.36_{\pm 0.76}$
	F	7.72	$70.84_{\pm 0.95}$	$74.63_{\pm 0.81}$	$76.64_{\pm 0.78}$
Average	P	18.54	54.79	58.28	58.36
	R	33.37	67.65	69.69	71.63
	F	23.37	60.10	63.08	63.99

Table 1. Results for biomedical NER tests are presented. We report Precision (P), Recall (R), and F1 (F) scores for each dataset. All values are given as percentages derived from an average of 3 trials. Datasets include the National Center for Biotechnology Information disease corpus (NCBI) and the BioCreative V CDR corpus (BC5CDR) [29].

effectively addressed data privacy concerns by minimizing the need for uploading patient data to an LLM API.

Results. Our work used LLMs to produce high-quality synthetic data, subsequently serving as the bedrock for fine-tuning dedicated models. Our approach yielded a significant performance boost, with F1-scores increasing from 23.37% to 63.99% in NER (Table 1) and from 75.86% to 83.59% in RE (Table 2), but also minimized data acquisition timelines and expenses. Furthermore, generating data using LLMs can significantly reduce the time and effort required for data collection and labeling and mitigate data privacy concerns. Our proposed framework sets a strong precedent to present this promising solution to enhance the applicability of LLM models to solve the critical clinical text mining challenge for efficient clinical trial design.



		Zero-shot	Fine-Tuned on Synthetic Data		
	Metrics	GPT	BERT	RoBERTa	BioBERT
GAD	P	76.32	$82.39_{\pm 0.93}$	$83.59_{\pm 1.01}$	$84.28_{\pm 1.03}$
	R	79.82	$90.21_{\pm 0.15}$	$92.57_{\pm 0.47}$	$94.21_{\pm 1.35}$
	F	78.03	$86.12_{\pm 0.72}$	$87.85_{\pm 0.68}$	$88.96_{\pm 1.01}$
EU-ADR	P	72.01	$72.05_{\pm 1.02}$	$73.44_{\pm 1.07}$	$75.81_{\pm 1.43}$
	R	75.43	$78.13_{\pm 0.50}$	$79.22_{\pm 0.22}$	$81.20_{\pm 1.00}$
	F	73.68	$74.96_{\pm 0.81}$	$76.22_{\pm 0.55}$	$78.41_{\pm 0.77}$
Average	P	74.16	77.22	78.52	80.05
	R	77.62	84.15	85.90	87.70
	F	75.86	80.53	82.04	83.69

Table 2. Results for biomedical RE tests. For each dataset, we detail the Precision (P), Recall (R), and F1 (F) scores. All figures are in percentages, averaged over 3 trials. The dataset includes the Gene Associations Database (GAD) [30], a corpus of gene-disease associations curated from genetic association studies with 5,330 annotations, and the EU-ADR [31]corpus, a biomedical relation extraction dataset that contains 100 abstracts with relations between drugs, disorders, and targets.

Experimental Approach to Leverage LLMs for Clinical Trial Design. As stated earlier, patient-trial matching involves finding appropriate patients for a specific clinical trial using their EHRs. These records reside in patients' medical databases. On the other hand, clinical trials are characterized by their descriptions, eligibility requirements, and other relevant details. Below, we elaborate on our problem setting and specific aims.

Problem setting overview: *Patient records (input 1)* are extracted from raw clinical documentation and converted into structured tables highlighting diagnoses, medications, and procedures as character strings. Alongside this, *clinical trials (input 2)* are considered, specifically focusing on eligibility criteria derived from raw trial documents. *The first task*, Patient-Criteria Matching, is approached as a multi-class classification with three categories: "Match" where patient records align with trial criteria, "Mismatch" indicating no alignment, and "Unknown" for ambiguous alignments. *The second task*, Patient-Trial Matching, determines a match if a patient satisfies all inclusion criteria and does not meet any exclusion criteria.

Aim 1: Privacy-aware augmentation techniques

To enable cross-institutional collaboration in patient-trial matching, extensive data availability is key. On the other hand, acquiring comprehensive and high-quality data will present significant challenges, especially with concerns about costs and potential privacy infringements. To address this, our team will introduce a data augmentation technique that harnesses the capabilities of LLMs. This approach will aim to generate supplementary data points while preserving the semantic coherence of the original trial's inclusion and exclusion criteria. Initially, we will employ the Chain-of-Thought method to guide the LLMs in gradually generating prompts. These prompts will be specifically crafted to ensure that the output data will be easily interpreted by machine learning models without losing its semantic essence. After establishing these prompts, we will use desensitized patient data combined with clinical trial data to execute a data augmentation process, thus guaranteeing the safeguarding of privacy. To illustrate this, Fig. 3 provides examples of the augmentation component of our proposed Large Language Model for Patient Trial Matching (LLM-PTM). On the technical front, given the criteria of a clinical trial represented as $T = [i_1, i_2, ..., i_{ni}, e_1, e_2, ..., e_{ne}]$, (where *i* and *e* are the inclusion and exclusion criteria respectively) we will task the LLM with producing augmented data points T that conform to specific constraints. For every criterion i_k and e_l in T, we will formulate input strings $i'_k$and e'_l using the equation $i'_k = o \oplus i_k$, $e'_l \oplus e_l$, where o is the designed prompt and $\oplus$ signify the concatenation operation. These strings will then be processed by the LLM to generate augmented data points A_{ik} and A_{el}, respectively. By viewing the LLM as a function, the expressions $A_{ik} = LLM(i'_k)$ and $A_{el} = LLM(e'_l)$ are derived, leading to the creation of the complete augmented trial dataset τ represented as: $\tau = \bigcup_{k=1}^{n} A_{i_k} \cup \bigcup_{l=1}^{m} A_{e_l}$

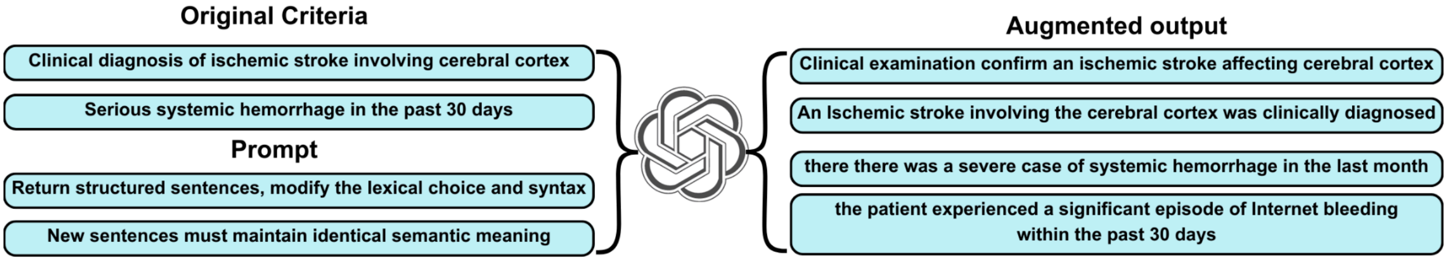

Original Criteria
Clinical diagnosis of ischemic stroke involving cerebral cortex
Serious systemic hemorrhage in the past 30 days
Prompt
Return structured sentences, modify the lexical choice and syntax
New sentences must maintain identical semantic meaning
Augmented output
Clinical examination confirm an ischemic stroke affecting cerebral cortex
An Ischemic stroke involving the cerebral cortex was clinically diagnosed
there there was a severe case of systemic hemorrhage in the last month
the patient experienced a significant episode of Internet bleeding within the past 30 days

Figure 3: Illustration of LLM-PTM augmented criteria.

Patient and Criteria Embedding. Once the text data has been preprocessed, it will undergo embedding through the taxonomy-guided deep learning method. We anticipate that latent representations of both a patient's visit record and trial criteria will be obtainable via LLMs. For the primary text encoding, our team will adopt the pretrained BERT [26]. When focusing on patient embedding, we will leverage a memory network [32], denoted as $Mem(\cdot)$. This approach will be instrumental in effectively preserving the sequential nature of visit data within the embedding space. Formally, the patient record embedding, denoted here as x_P, will be derived from the encoding function $f_P(\cdot)$, and can be expressed as $x_P = f_P(P) = Mem(BERT(a_1), BERT(a_2), ..., BERT(a_n))$, where P represents the patient's records.

Regarding the embedding of criteria, to capture essential features within the embedding space, we propose a method that intertwines a convolutional neural network (CNN) with a highway layer [33]. This combination is envisioned to distill patterns across various levels for the semantic matching task [34]. Formally, the encoding function $f_c(\cdot)$ will be employed to produce an EC embedding x_c, defined as $x_{i/e} = f_c(c) = Highway(BERT\ (c))$, where $c \in T$. Conclusively, the outputs from the highway network will be determined by the equation $Highway(\cdot) = Sigmoid(\cdot)Conv(\cdot) + Conv(\cdot)(1 - Sigmoid(Conv(\cdot)))$.

Potential Problems and Alternatives. We anticipate that GPT will demonstrate its capabilities in generating high-quality synthetic data, a capacity evidenced by our previous preliminary work. However, an emerging concern will be GPT's training on publicly available datasets. This raises the concern that the model might have already been exposed to the datasets we will use in our Phase I feasibility study, potentially leading to unintentional disclosures from the primary dataset. To counteract this potential pitfall, we will employ the sentence transformer to secure embeddings for both the original and the synthetic data. Subsequently, these embeddings will be projected using t-distributed stochastic neighbor embedding. Identifying distinct patterns between the synthetic and original data will be crucial, serving as indicators to ensure that GPT hasn't merely reproduced the dataset from its internal memory.

Aim 2: Patient-clinical Trial Matching

Throughout the model optimization process, our primary objective is to enhance patient-trial matching. A pivotal part of this endeavor involves addressing the distinct differences between inclusion and exclusion criteria. To reach our goal, we will introduce a composite loss function. This function will incorporate several loss terms, starting with the classification loss. As part of our strategy to optimize the classification performance, we intend to use a cross-entropy loss term $L_{cla} = - y^T log(\hat{y}) - (1 - y)^T log(1 - \hat{y})$ which will measure the discrepancy between the predicted outcome ŷ and the ground truth y.

Our optimization process will also involve the creation of an inclusion/exclusion contrastive loss term. By constructing this term, we aim to directly address the alignment between the patient embedding and the EC embedding, for both inclusion and exclusion stipulations. The essence of this loss term lies in its ability to enable the model to pinpoint specific features, such as negation words, within the inclusion/exclusion criteria. This, in turn, facilitates the decision-making process regarding whether to include or exclude a patient. From a mathematical perspective, our goal will be to heighten the similarity between the retrieved patient memory and the inclusion criteria embedding, represented as (x_P, x_i), while concurrently reducing the similarity with the exclusion criteria embedding, denoted as (x_P, x_i). The formulation of this contrastive loss term will hinge on the following pairwise distance loss:

$$L_{con} = \prod_{a=1,...,n_i} (1 - s(x_{ia}, x_P)) \cdot \prod_{b=1,...,n_e} max(0, s(x_{eb}, x_P) - \epsilon)$$

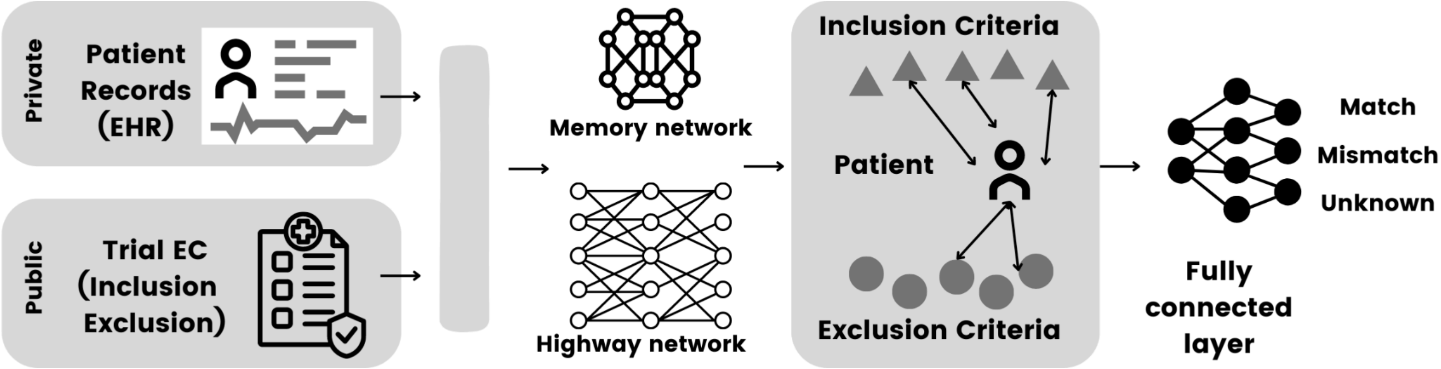


Figure 4: Overall model framework

where s(·, ·) represents the similarity function between two vectors. Our team plans to opt for the cosine similarity function, anticipating its efficacy in measuring the distance between two data modalities. The hyperparameter ε will be introduced to define the minimum allowed similarity between the exclusion criteria embedding and the patient's memory. When a patient's data aligns with an inclusion criterion, the model's design will aim to amplify the cosine similarity between the associated embeddings, making the value of $1 - s(x_i, x_P)$ converge to zero. Conversely, if a patient's profile leads to an exclusion based on a specific criterion, the model will strive to diminish the similarity between the corresponding embeddings. In this scenario, the value of $max(0, s(x_e, x_{P,}) - ε$ will be driven downwards, but it will always remain above ε, ensuring that x_i and x_e maintaining distinguishable distances relative to x_P in the latent space. Ultimately, the strategy will encompass the joint minimization of the loss functions via backpropagation, executed in an end-to-end manner as depicted by $L = \alpha \cdot L_{cla} + (1 - \alpha) \cdot L_{con}$, where α will serve as a crucial parameter adjusting the weight of classification loss. A detailed representation of this model's architecture and workflow will be available in Fig. 4.

Proposed Work: Experimental**.** We will collect data from six different stroke clinical trials, namely NCT03735979, NCT03805308, NCT03263117, NCT03496883, NCT03876457, and NCT03545607, using ClinicalTrials.gov as our source. Our focus will be on both the inclusion and exclusion criteria, which will result in at least 150 sentence-level statements extracted.

Potential Problems and Alternatives: There is a potential negative impact of the number of synthetic sentences on the effectiveness of our proposed method. To address this issue, we will conduct experiments with varying numbers of synthetic sentences and ratios of seed examples. We propose a series of tests that uses a range of synthetic data to train our local model and to vary the pool size of our seed examples. The results of these examinations could indicate, for example, that increasing the number of synthetic sentences can improve model performance up to a certain point, beyond which the improvement becomes marginal; or that using a larger number of seed examples can increase the quality and diversity of the generated data.

Long-term goal. In Phase I, our team will showcase the effectiveness of our LLM-driven platform in allowing cross-organizational collaboration for clinical trial design. Our solution will ensure data de-identification with minimal human-labeled data, guide LLMs in reliable data augmentation through precise prompts, and align EHR data with clinical trial criteria. In Phase II, AI POW LLC aims to refine and scale the platform, integrating feedback from Phase I to enhance accuracy and user experience. We'll also expand partnerships with leading healthcare institutions to ultimately drive forward the vision of AI-powered, collaborative clinical trial design.

References

1. Trends, Charts, and Maps. [cited 17 Aug 2023]. Available: https://classic.clinicaltrials.gov/ct2/resources/trends

2. McCray AT, Ide NC. Design and implementation of a national clinical trials registry. J Am Med Inform Assoc. 2000;7: 313–323. doi:10.1136/jamia.2000.0070313

3. Laine C, Horton R, DeAngelis CD, Drazen JM, Frizelle FA, Godlee F, et al. Clinical trial registration: looking back and moving ahead. Croat Med J. 2007;48: 289–291. Available: https://www.ncbi.nlm.nih.gov/pubmed/17589970

4. Bardyn TP, Patridge EF, Moore MT, Koh JJ. Health Sciences Libraries Advancing Collaborative Clinical Research Data Management in Universities. J Escience Librariansh. 2018;7. doi:10.7191/jeslib.2018.1130

5. Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB. Fundamentals of Clinical Trials. Springer International Publishing; doi:10.1007/978-3-319-18539-2

6. Inan OT, Tenaerts P, Prindiville SA, Reynolds HR, Dizon DS, Cooper-Arnold K, et al. Digitizing clinical trials. NPJ Digit Med. 2020;3: 101. doi:10.1038/s41746-020-0302-y

7. Moore TJ, Heyward J, Anderson G, Alexander GC. Variation in the estimated costs of pivotal clinical benefit trials supporting the US approval of new therapeutic agents, 2015-2017: a cross-sectional study. BMJ Open. 2020;10: e038863. doi:10.1136/bmjopen-2020-038863

8. Hargreaves B. Clinical trials and their patients: the rising costs and how to stem the loss. In: Pharmafile [Internet]. [cited Aug 2023]. Available: http://www pharmafile com/news/511225/clinical-trials-and-their-patients-rising-costs-and-how-stem-loss

9. CTG labs - NCBI. [cited 18 Aug 2023]. Available: https://clinicaltrials.gov/ct2/about-site/background

10. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014;32: 40–51. doi:10.1038/nbt.2786

11. Glick HA, Doshi JA, Sonnad SS, Polsky D. Economic Evaluation in Clinical Trials. OUP Oxford; 2014. Available: https://play.google.com/store/books/details?id=Xqi1BAAAQBAJ

12. Drummond MF, Stoddart GL. Economic analysis and clinical trials. Control Clin Trials. 1984;5: 115–128. doi:10.1016/0197-2456(84)90118-1

13. Beacher FD, Mujica-Parodi LR, Gupta S, Ancora LA. Machine Learning Predicts Outcomes of Phase III Clinical Trials for Prostate Cancer. Algorithms. 2021;14: 147. doi:10.3390/a14050147

14. Gayvert KM, Madhukar NS, Elemento O. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. Cell Chem Biol. 2016;23: 1294–1301. doi:10.1016/j.chembiol.2016.07.023

15. Follett L, Geletta S, Laugerman M. Quantifying risk associated with clinical trial termination: A text mining approach. Inf Process Manag. 2019;56: 516–525. doi:10.1016/j.ipm.2018.11.009

16. Elkin ME, Zhu X. Predictive modeling of clinical trial terminations using feature engineering and embedding learning. Sci Rep. 2021;11: 3446. doi:10.1038/s41598-021-82840-x

17. Hassanzadeh H, Karimi S, Nguyen A. Matching patients to clinical trials using semantically enriched document representation. J Biomed Inform. 2020;105: 103406. doi:10.1016/j.jbi.2020.103406

18. Alexander M, Solomon B, Ball DL, Sheerin M, Dankwa-Mullan I, Preininger AM, et al. Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients. JAMIA Open. 2020;3: 209–215. doi:10.1093/jamiaopen/ooaa002

19. Bhatt A. Artificial intelligence in managing clinical trial design and conduct: Man and machine still on the learning curve? Perspect Clin Res. 2021;12: 1–3. doi:10.4103/picr.PICR_312_20

20. Gao J, Xiao C, Glass LM, Sun J. COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery; 2020. pp. 803–812. doi:10.1145/3394486.3403123

21. Zhang X, Xiao C, Glass LM, Sun J. DeepEnroll: Patient-Trial Matching with Deep Embedding and Entailment Prediction. Proceedings of The Web Conference 2020. New York, NY, USA: Association for Computing Machinery; 2020. pp. 1029–1037. doi:10.1145/3366423.3380181

22. Sanderson K. GPT-4 is here: what scientists think. Nature. 2023;615: 773. doi:10.1038/d41586-023-00816-5

23. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2305.09617

24. Hosseini M, Gao CA, Liebovitz DM, Carvalho AM, Ahmad FS, Luo Y, et al. An exploratory survey about using ChatGPT in education, healthcare, and research. medRxiv. 2023. doi:10.1101/2023.03.31.23287979

25. Aziz HA. A review of the role of public health informatics in healthcare. Journal of Taibah University Medical Sciences. 2017;12: 78–81. doi:10.1016/j.jtumed.2016.08.011

26. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv [cs.CL]. 2018. Available: http://arxiv.org/abs/1810.04805

27. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33: 1877–1901. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_medium=email&utm_source=transaction

28. Qin C, Zhang A, Zhang Z, Chen J, Yasunaga M, Yang D. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2302.06476

29. Li J, Sun Y, Johnson RJ, Sciaky D, Wei C-H, Leaman R, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database . 2016;2016. doi:10.1093/database/baw068

30. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database . 2016;2016. doi:10.1093/database/baw100

31. van Mulligen EM, Fourrier-Reglat A, Gurwitz D, Molokhia M, Nieto A, Trifiro G, et al. The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. J Biomed Inform. 2012;45: 879–884. doi:10.1016/j.jbi.2012.04.004

32. Weston J, Chopra S, Bordes A. Memory Networks. arXiv [cs.AI]. 2014. Available: http://arxiv.org/abs/1410.3916v11

33. Srivastava RK, Greff K, Schmidhuber J. Highway Networks. arXiv [cs.LG]. 2015. Available: http://arxiv.org/abs/1505.00387

34. You Q, Luo J, Zhang Z. End-to-end convolutional semantic embeddings. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2018. pp. 5735–5744. doi:10.1109/cvpr.2018.00601

Other:

Title: Leveraging Large Language Models for Privacy-Aware Data Augmentation to Enhance Patient-Clinical Trial Matching

Abstract. The process of matching patients with appropriate clinical trials is a crucial step in the advancement of medical research and the delivery of optimal care. When patients are correctly matched with relevant clinical trials, they might gain access to new, potentially life-saving treatments not yet available to the general public. Conversely, considering that a clinical trial has a median estimate of US$19 million per trial, filling a clinical trial with appropriate participants can lead to cost savings, as prolonged or inappropriate trials can be expensive. Identifying solutions that ensure precise and streamlined patient-trial matching is vital. However, contemporary methodologies face challenges, including data standardization, ethical considerations, a lack of cross-institutional collaboration features, and the absence of seamless integration between Electronic Health Records (EHRs) and clinical trial criteria. While some solutions using black-boxed embedding matching have been proposed, the efficacy of AI-powered clinical trial matching services remains a challenge, especially when merging information from EHRs with clinical trial criteria. This project aims to leverage large language models (LLMs) to enhance the compatibility between EHRs and clinical trial descriptions, enabling a privacy-first and more accurate patient-trial matching. By utilizing advances in natural language processing, new models can interpret, understand, and align the varied terminologies and ontologies in EHRs with clinical trial inclusion and exclusion criteria. Such integration is expected to simplify the matching process and increase the precision of finding suitable patient trials, fostering collaboration tools among researchers, clinicians, and trial coordinators. In Phase I, the team plans to harness the capabilities of LLMs for privacy-aware data augmentation to improve cross-institutional clinical trial design. The aim is to use LLMs and limited seed data samples to generate a comprehensive synthetic dataset, ensuring data reliability and adherence to strict privacy standards for cross-institutional collaboration. A subsequent goal is to create a privacy-focused classification framework based on a Memory and Highway network approach that effectively matches patients' EHR to clinical trials' criteria, optimizing their potential for enrollment. At the onset of Phase I, a model prototype will be developed to assess and validate the patient-trial matching system using six different stroke clinical trials from ClinicalTrials.gov. This phase is also designed to better understand the collaboration needs of researchers, clinicians, and trial coordinators during a clinical trial. The project team comprises experts in Data Mining, Machine Learning Automation, Biomedical Informatics, and Natural Language Processing, along with collaborators for model validation and clinical trial implementation. Completing Phase I work will pave the way for a tool that redefines clinical trial design, fostering collaboration among medical professionals to connect patients to potentially transformative treatments and driving groundbreaking outcomes in patient care and clinical research.

Project Narrative.

Matching patients with appropriate clinical trials is pivotal for medical research and patient care, yet contemporary methodologies face challenges like data standardization, integration between Electronic Health Records (EHRs), and trial criteria. This project aims to leverage large language models (LLMs) to enhance the compatibility between EHRs and trial descriptions, simplifying the matching process and fostering collaboration among researchers, clinicians, and coordinators. By harnessing advances in natural language processing and privacy-aware data augmentation, the initiative seeks to redefine trial design, potentially transforming patient outcomes and advancing medical research.