Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial Matching

Jiayi Yuan[1], Ruixiang Tang[1], Xiaoqian Jiang, PhD[2], Xia Hu, PhD[1]
[1] **Rice University, Houston, TX;** [2]**University of Texas Health Science Center, Houston, TX.**

Abstract

The process of matching patients with suitable clinical trials is essential for advancing medical research and providing optimal care. However, current approaches face challenges such as data standardization, ethical considerations, and a lack of interoperability between Electronic Health Records (EHRs) and clinical trial criteria. In this paper, we explore the potential of large language models (LLMs) to address these challenges by leveraging their advanced natural language generation capabilities to improve compatibility between EHRs and clinical trial descriptions. We propose an innovative privacy-aware data augmentation approach for LLM-based patient-trial matching (LLM-PTM), which balances the benefits of LLMs while ensuring the security and confidentiality of sensitive patient data. Our experiments demonstrate a 7.32% average improvement in performance using the proposed LLM-PTM method, and the generalizability to new data is improved by 12.12%. Additionally, we present case studies to further illustrate the effectiveness of our approach and provide a deeper understanding of its underlying principles.

1 Introduction

Identifying suitable clinical trials for patients is crucial for the reliable assessment and examination of medical treatments. This process not only enables patients to obtain the best possible care but also supports advancements in medical research. Models that facilitate clinical trial matching compare patient profiles with the eligibility requirements of ongoing studies to pinpoint potential matches [1]. Nonetheless, discovering the ideal clinical trial for a patient can be a complex and lengthy procedure. This is where advanced clinical trial matching methods come into play. By employing artificial intelligence (AI) or specific algorithms, researchers can pair patients with fitting clinical trials [2, 3], a task that is growing increasingly vital as the number of studies continues to expand. Please refer to our preliminary section for more details.

In spite of the potential benefits, integrating AI into clinical trial processes encounters various obstacles, such as issues pertaining to data availability, standardization, and ethical considerations [4]. A significant challenge lies in the discordance between the ontology and terminology utilized in Electronic Health Records (EHRs) and those implemented in clinical trial inclusion and exclusion criteria. Although recent research has produced solutions to address these concerns through black-boxed embedding matching [5, 6], the efficacy of AI-driven clinical trial matching services may still be impeded by the difficulty in interoperating the information from the two disparate sources.

The advent of large language models (LLM) [7] presents an opportunity to enhance the compatibility between EHRs and clinical trial descriptions, promoting more accurate patient-trial matching. By capitalizing on their sophisticated natural language processing abilities, these models can efficiently decipher, comprehend, and harmonize the diverse terminologies and ontologies present in both EHRs and clinical trial inclusion and exclusion criteria. This enhanced interoperability not only streamlines the matching process but also ensures greater accuracy in identifying suitable trials for patients. Consequently, large language models hold the potential to transform clinical trial matching, enabling better patient outcomes and contributing to more efficient medical research.

The foremost challenge when implementing LLMs for clinical trial matching lies in managing privacy concerns that arise due to the handling of sensitive patient data, which may include personal and health-related information. Ensuring the security and confidentiality of this data is crucial to maintain patient trust and adhere to legal and ethical standards. To overcome this challenge, we propose an innovative data augmentation that prioritizes data privacy while maintaining the benefits offered by LLMs. During the actual implementation, instead of feeding original patient data directly into LLMs, we use desensitized patient data as a prompt to guide the LLM in the augmentation process of the trial data. Through comprehensive experimentation, our proposed LLM-PTM demonstrates an average improvement in performance by 7.32%, it also improves generalizability by 12.12%. Moreover, we conduct case studies to elucidate the efficacy of the method and provide a broader insight into its underlying principles.
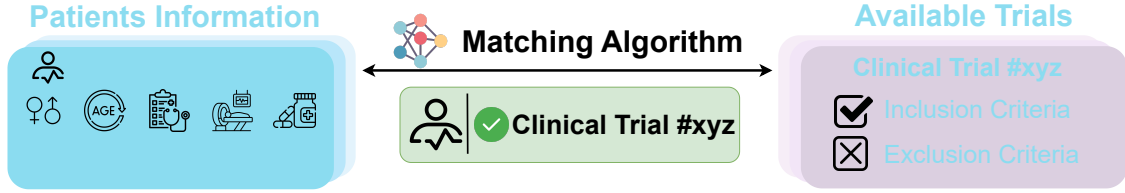
Figure 1: An illustration of patient-trial matching.

2 Preliminary

2.1 Patient Trial Matching

In the realm of clinical research, clinical trials serve as the solely established methodology for the development of novel disease treatments. However, these trials frequently encounter challenges such as costly, imprecise, and inadequate patient recruitment. A substantial number of trials grapple with obtaining the necessary patient population, with 50% experiencing delays due to recruitment issues, while others fail to secure a sufficient number of participants to initiate the trial [8]. The advent of automated patient-trial matching offers a promising avenue for optimizing the trial recruitment process. The crux of this approach lies in identifying eligible patients for clinical trials based on their longitudinal electronic health records (EHR) and trial eligibility criteria (EC), which encompass both inclusion and exclusion criteria as depicted in Fig. 1. The problem can be framed as a classification problem. Given input consisting of a patient's complete Electronic Health Record (EHR) data and a single trial's eligibility criteria (EC), the output can be classified as either match, mismatch, or unknown. A match between a patient and a trial occurs if and only if the patient satisfies all of the trial's ECs. Mathematically, let P represent the patient's EHR data and $T = [t_1, ..., t_n]$ represent the trial's eligibility criteria. The matching function $m(P, t) = \{match, mismatch, unknown\}$ and $M(P, T) = \{match, mismatch\}$.

2.2 Data Augmentation

Text data augmentation, involving diverse transformations, is widely utilized to improve model training for text classification tasks within the domain of natural language processing (NLP). Contemporary data augmentation techniques in NLP function at distinct levels of granularity, including characters, words, sentences, and documents. The objective of data augmentation is to generate intelligible and varied additional instances that uphold semantic congruity. In the context of patient-trial matching, data augmentation assumes particular importance due to the limited richness of the available training data. The implementation of an effective augmentation method can facilitate the generation of a more varied text dataset, allowing machine learning models to better capture the intricacies of patient and eligibility criteria information. Consequently, this leads to enhanced training of classifiers, resulting in more accurate and reliable outcomes for patient-trial matching tasks.

2.3 Data Privacy in Healthcare

The potential of big data to revolutionize healthcare is acknowledged [9], but privacy concerns remain. Privacy issues lead to both consequentialist and deontological concerns in data usage [9]. Restricting access to patient data can impede data-driven innovation [10], while data deidentification complicates linking patient data from different sources [11, 12]. Alternative privacy-preserving approaches, such as pseudonymized data or differential privacy techniques, can be applied in certain contexts [13, 14, 15]. Privacy audits and security standards can also help, with data holders acting as stewards rather than privacy-agnostic intermediaries. Yet, a privacy-innovation tradeoff may still persist in many situations. In the context patient-trial matching, privacy concerns also arise due to the inherent nature of big data. This study explores privacy-conscious augmentation techniques that interact with open-source components, aiming to improve match identification accuracy while preventing the leakage of private-source data.

3 Methodology

3.1 Problem Setting

As previously elucidated, the process of patient-trial matching entails the identification of suitable patients for a specific clinical trial based on their Electronic Health Records (EHR). These records contain comprehensive medical information about the patient, while clinical trials are characterized by detailed descriptions, eligibility criteria, and other pertinent data. In the following section, we will systematically introduce the problem setting.

Input 1: Patient Records. We extract patient records from raw clinical documentation and subsequently transform them into a structured tabular format. We represent patient records by the symbol P and formally define it as: $P = [d_1, d_2, \ldots, d_{n_d}, m_1, m_2, \ldots, m_{n_m}, p_1, p_2, \ldots, p_{n_p}]$. Here, d_i refers to a diagnosis belonging to the set $\mathcal{D}$ of diagnoses, m_i represents a medication from the set $\mathcal{M}$ of medications, and p_i denotes a procedure within the set $\mathcal{P}$ of procedures. All elements, i.e., diagnoses, medications, and procedures, are character strings. The quantities $n_d, n_m,$ and n_p are the total numbers of diagnoses, medications, and procedures in the electronic health record (EHR) categories, respectively.

Input 2: Clinical Trials. In our problem setting, we focus on the criteria of clinical trials. Analogous to the extraction of patient notes, we retrieve eligibility criteria from unprocessed patient trial documents. Let us denote clinical trials by the symbol T and express it as: $T = [i_1, i_2, \ldots, i_{n_i}, e_1, e_2, \ldots, e_{n_e}]$. In this representation, i_i and e_i symbolize the inclusion and exclusion criteria, respectively, both of which are character strings. The variables n_i and n_e indicate the total number of inclusion and exclusion criteria, correspondingly.

Task 1: Patient-Criteria Matching. Given a patient's visit records P and a collection of inclusion or exclusion criteria, we frame the patient-criteria matching task as a multi-class classification problem. The objective is to categorize the matching outcomes between patients and Eligibility Criteria (ECs) into three distinct classes: "match", "mismatch", and "unknown", according to the similarity between patient records and trial criteria. We can represent this as: $\hat{y}(c, P) \in \{match, \ mismatch, \ unknown\}$, where $c \in T$.

Task 2: Patient-Trial Matching. Given a patient's visit records P and a clinical trial C consisting of a collection of inclusion and exclusion criteria, we assert that a patient and a trial constitute a match only if the patient satisfies all inclusion criteria and contradicts all exclusion criteria in the trial. We can represent this condition as: $Match(C, P) \Leftrightarrow (\forall i \in \{i_1, \ldots, i_{n_i}\}, : \hat{y}(i, P) = match) \wedge (\forall e \in \{e_1, \ldots, e_{n_e}\}, : \hat{y}(e, P) = match)$.

3.2 Proposed Pipeline

Trial Eligibility Criteria Augmentation

As previously discussed, the acquisition of comprehensive and high-quality data in patient-trial matching presents significant challenges, including considerable expenses and potential privacy infringement. Recognizing the necessity for data augmentation in this context, we have introduced LLM-PTM. In this work, we put forth a data augmentation technique utilizing Language Models (LLMs) to create supplementary data points while preserving the semantic coherence of the original trial's inclusion (i) and exclusion (e) criteria. We first employ Chain-of-Thought to direct the LLMs to gradually generate the prompts. These prompts incorporate our requirement that the output data be more comprehensible to machine learning models, while preserving the exact semantic content. Subsequently, we use desensitized patient data, clinical trial data, and the previously generated prompts to execute a data augmentation process that preserves privacy.

The illustration and examples of the augmentation part of LLM-PTM can be found in Fig. 2. Given the criteria of a clinical trial: $T = [i_1, i_2, \ldots, i_{n_i}, e_1, e_2, \ldots, e_{n_e}]$, we aim to utilize an LLM to generate a set of augmented data points $\mathcal{T}$ that adhere to these constraints. Formally, for each criteria i_k and $e_l \in T$, we construct input strings i'_k and e'_l as follows: $i'_k = o \oplus i_k, e'_l = o \oplus e_l$ where o is the designed prompt, $\oplus$ denotes the concatenation operation. Subsequently, we feed the input strings i'_k and e'_l into the LLM to generate a set of augmented data points A_{i_k} and A_{e_l}, respectively. Regard the LLM as a function $LLM()$: $A_{i_k} = LLM(i'_k), A_{e_l} = LLM(e'_l)$. The final augmented trial dataset $\mathcal{T}$ can then be represented as:
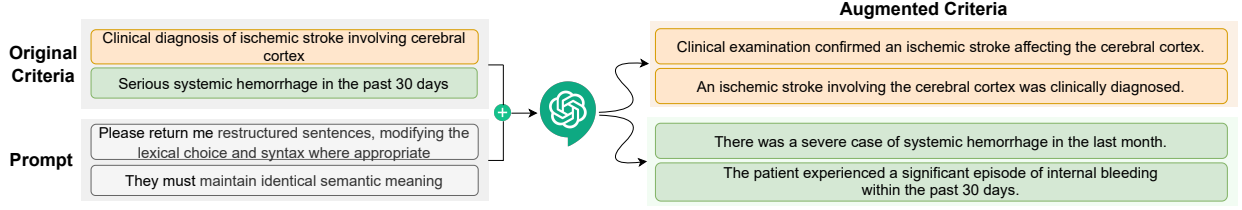
Figure 2: Illustration of LLM-PTM augmented criteria.

$$\mathcal{T} = \bigcup_{k=1}^{n} A_{i_k} \cup \bigcup_{l=1}^{m} A_{e_l} \tag{1}$$

Patient and Criteria Embedding

After getting preprocessed text data, we then do embedding with the taxonomy-guided deep learning method. Latent representations of a patient's visit record and trial criteria can be acquired through the utilization of extensive language models (LLMs). In this work, we employ the pretrained BERT [16] as our text encoder. For patient embedding, a memory network [17], $Mem(\cdot)$, is utilized to manage the patient records, thereby effectively maintaining the sequence of visit data within the embedding space. Formally, the patient record embedding, represented as x_P, is derived from the encoding function $f_P(\cdot)$, which can be expressed as:

$$x_P = f_P(P) = Mem(BERT(a_1), BERT(a_2), \ldots, BERT(a_n)), \; a_i \in \{\mathcal{D}, \mathcal{M}, \mathcal{P}\}. \tag{2}$$

In order to capture and encode these essential features within the embedding space, we implement a prior approach that employs a convolutional neural network (CNN) and a highway layer [18] to extract patterns at various levels for the semantic matching task [19]. Formally, the encoding function $f_c(\cdot)$ is employed to encode an EC embedding x_c, which can be described as:

$$x_{i/e} = f_c(c) = Highway(BERT(c)), \tag{3}$$

where $c \in \mathcal{T}$. The outputs of highway networks are calculated as $Highway(\cdot) = Sigmoid(\cdot)Conv(\cdot) + Conv(\cdot)(1 - Sigmoid(Conv(\cdot)))$.

Prediction and Embedding Learning

Throughout the model optimization process, we aim to maximize patient-trial matching while explicitly addressing the differences between inclusion and exclusion criteria. To achieve this, we design a composite loss function that comprises the following loss terms. The first loss is classification loss. To optimize the classification performance between the predicted outcome $\hat{y}$ and the ground truth y, we employ a cross-entropy loss term:

$$\mathcal{L}_{cla} = -y^T \log(\hat{y}) - (1 - y)^T \log(1 - \hat{y}). \tag{4}$$

Furthermore, we construct inclusion/exclusion contrastive loss term to explicitly address the match between patient embedding and EC embedding for both inclusion and exclusion criteria. This loss term enables the model to extract distinct features (e.g., negation words) within the inclusion/exclusion criteria, thereby aiding the decision to include or exclude a patient. Mathematically, this involves maximizing the similarity between the retrieved patient memory and the embedding for inclusion criteria (i.e., (x_P, x_i)) while minimizing the similarity between the memory and the embedding for exclusion criteria (i.e., (x_P, x_e)). We formulate the loss term using the following pairwise distance loss:

$$\mathcal{L}_{con} = \prod_{a=1,\ldots,n_i} (1 - s(x_{i_a}, x_P)) \cdot \prod_{b=1,\ldots,n_e} \max(0, s(x_{e_b}, x_P) - \varepsilon), \tag{5}$$
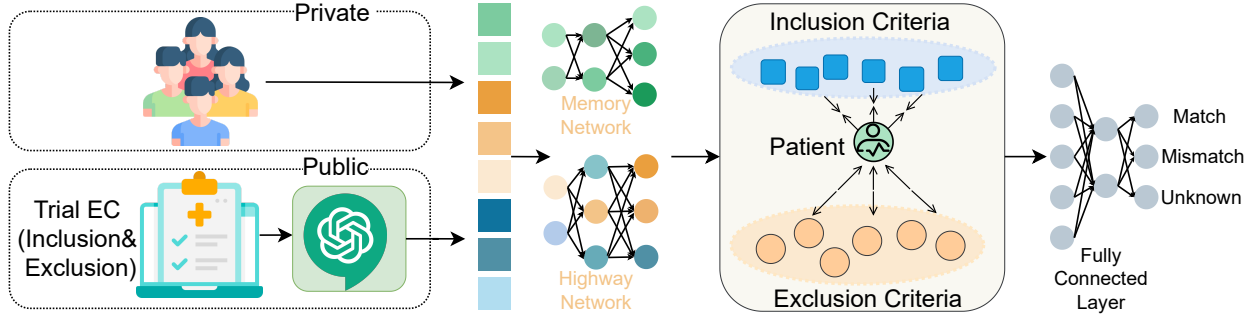
Figure 3: Overall model framework

where $s(\cdot, \cdot)$ represents the similarity function between two vectors. In our work, we utilize the cosine similarity function to determine the distance between two data modalities. The hyperparameter ε signifies the minimum similarity between the exclusion criteria embedding and patient memory. If a patient matches an inclusion criterion, the model maximize the cosine similarity between the two embeddings, making $1 - s(x_i, x_P)$ approach 0. If a patient is excluded due to an exclusion criterion, the similarity between the two embeddings (i.e., $\max(0, s(x_e, x_P) - \varepsilon)$) is minimized and must be no less than ε. This ensures that x_i and x_e have distinct distances to x_P in the latent embedding space. Finally, we minimize the loss functions jointly through backpropagation in an end-to-end manner in Eq.6,

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{cla} + (1 - \alpha) \cdot \mathcal{L}_{con}, \tag{6}$$

where α controls the strength of classification loss. The overall framework can be found in Fig. 3.

4 Experiment

4.1 Dataset

Clinical Trial Data. We collected data from six different stroke clinical trials, namely NCT03735979, NCT03805308, NCT03263117, NCT03496883, NCT03876457, and NCT03545607, using *ClinicalTrials.gov* as our source. Our focus was on both the inclusion and exclusion criteria, resulting in 150 sentence-level statements extracted.

Patient EHR Data. The project was approved by the UTHealth Institutional Review Board (IRB) under HSC-SBMI-21-0529 - "Re-admission Risk Estimation for Stroke Patients". Using the stroke patient database, we gathered patient claims data for 825 patients who were enrolled in at least one of the six stroke trials. This data included longitudinal prescription and medical claims data for each patient, encompassing their diagnoses, procedures, and medications.

Augment Data. In the development of the baseline methods, we utilized open-source libraries to generate augmented data. For our LLM-PTM, which employed a LLM-based prompt augmentation, we incorporated OpenAI's ChatGPT from the version released on March 5th.

We then matched each inclusion/exclusion criteria with its corresponding patient EHR, labeling them as either "match" or "mismatch". To add an element of uncertainty, we also included "unknown" labels for one inclusion criterion and one exclusion criterion randomly selected from another trial, leading to a total of 100,000 labeled pairs.

4.2 Model and Setting

Model Configuration In the process of text embedding, we employed the Clinical BERT embeddings as our primary method, as delineated by [20]. This specific model was pre-trained on a substantial dataset comprising 2 million clinical notes, which were extracted from the MIMIC-III v1.4 database [21]. To optimize the performance of the model, a total of 150,000 iterations were executed during the training phase. The resultant embedding produced by the Clinical BERT model featured a dimensionality of 768. As for the architectural design of the highway network,

it consisted of a two-layered convolutional structure with each layer containing 128 channels. This configuration was strategically chosen to enhance the learning capacity of the model while maintaining computational efficiency. The machine learning models are built upon PyTorch.

Baselines. We benchmarked the proposed data augmentation approach in comparison to original data and the standard data augmentation technique on the dataset. Specifically, we tested 1) Swap Word Augmentation, a method that randomly swaps words within the text and is a subset of the Easy Data Augmentation (EDA) [22] proposed by Wei et al; 2) Context Word Augmentation [23, 24], which employs BERT to insert contextually relevant words by adding a mask token at a random position in the input text and allowing BERT to predict the appropriate word; 3) Back Translation Augmentation. This technique [25] involves translating the text to another language and then back to English, yielding a distinct text that maintains the original meaning. In this work, we employed Google Translate in deep translator as a translation framework and utilize German, French, and Spanish as intermediate languages.

Training Settings. In the process of developing our experimental framework, we established our training configurations in accordance with the prevailing settings that are widely accepted within the field. To elaborate, we commenced the training of the entire neural networks from scratch for a predetermined duration of 12 epochs. Throughout this stage, we employed the Adam optimizer [26], accompanied by a batch size of 128 and a learning rate of 1e-4. The hyperparameter α was set to 0.5 and ε was set to 0.01.

Evaluation Metrics. The outcomes of our study are presented at two distinct levels: the patient-criteria level and the patient-trial level. To comprehensively evaluate the performance of the model, we employed a variety of widely recognized evaluation metrics, including Precision, Recall, and F1 score.

4.3   Experimental Results Analysis

4.3.1   Overall Performance

In this section, we aggregate the results from both patient criteria and patient-trial matching across all trials under investigation. A comparative analysis between our proposed methods and the established baseline methods is presented in Tab. 1. The experimental results are obtained by calculating the average of three separate tests. Upon examining the experimental outcomes, we discerned that the implementation of our LLM-based augmentation approach yielded a substantial enhancement in the performance of machine learning models across all evaluation metrics.

Table 1: Overall performance

Performance	Patient-Criteria Level			Patient-Trial Level		
	Prec	Rec	F1	Prec	Rec	F1
Vanilla (No Augmentation)	$0.858_{\pm1.1e-2}$	$0.841_{\pm5.7e-3}$	$0.850_{\pm8.1e-3}$	$0.715_{\pm2.0e-2}$	$0.748_{\pm1.7e-2}$	$0.731_{\pm1.9e-2}$
Swap Word Augmentation	$0.854_{\pm1.8e-2}$	$0.814_{\pm1.6e-2}$	$0.833_{\pm1.7e-2}$	$0.707_{\pm1.2e-2}$	$0.759_{\pm1.4e-2}$	$0.732_{\pm1.3e-2}$
Context Word Augmentation	$0.855_{\pm1.0e-2}$	$0.835_{\pm1.4e-2}$	$0.844_{\pm1.2e-2}$	$0.752_{\pm1.7e-2}$	$0.789_{\pm1.5e-2}$	$0.770_{\pm1.6e-2}$
Back Translation Augmentation	$0.919_{\pm1.1e-2}$	$0.845_{\pm1.3e-2}$	$0.880_{\pm1.2e-2}$	$0.777_{\pm1.4e-2}$	$0.796_{\pm1.3e-2}$	$0.787_{\pm1.3e-2}$
LLM-PTM	$\mathbf{0.964}_{\pm8.0e-3}$	$\mathbf{0.862}_{\pm1.2e-2}$	$\mathbf{0.910}_{\pm1.0e-2}$	$\mathbf{0.801}_{\pm1.3e-2}$	$\mathbf{0.830}_{\pm1.1e-2}$	$\mathbf{0.815}_{\pm1.1e-2}$

More specifically, the average performance improvements for Precision, Recall, and F1 scores at the patient-criteria level are more than 10.6%, 2.1%, and 6.0%, respectively. A more pronounced performance gain is observed at the patient-trial level, with respective improvements of 8.6%, 8.2%, and 8.4%. These performance gains can be attributed to the inclusion of a more diverse dataset (in comparison to the original dataset) and the incorporation of more precise semantic information (relative to other augmentation techniques). In the forthcoming case study, we will delve deeper into the implications and nuances of these findings, providing a more comprehensive understanding of the impact of our proposed LLM-based augmentation method on the performance of machine learning models in this domain.

4.3.2 Performance Throughout Different Trials

In this section, we aim to evaluate the performance of our proposed model at the patient criteria level across various trials. The primary objective of this analysis is to determine whether criteria in certain trials prove to be more challenging than others and to assess the extent to which the models can achieve satisfactory results. The outcomes of this examination are presented in Tab. 2. Our findings indicate that the performance of the baseline model does indeed falter in some trials, whereas our proposed method demonstrates a more consistent and robust capacity to tackle these challenging tasks. For instance, in Trial 1, the baseline model only managed to attain a precision of 40.2%, while our method exhibited an improvement of 24.0%. This suggests that our approach, which incorporates enriched and varied semantic information, can significantly enhance the model's performance, particularly when faced with difficult data. In the subsequent case study, we will delve further into the nuances of these observations to better understand the factors contributing to the success of our proposed method.

Table 2: Performance throughout different trials

Performance	Vanilla			LLM-PTM		
	Prec	Rec	F1	Prec	Rec	F1
Trial1 (NCT03263117)	0.402	0.382	0.391	0.642	0.502	0.563
Trial2 (NCT03496883)	1.000	0.925	0.961	1.000	0.976	0.987
Trial3 (NCT03545607)	0.788	0.749	0.768	0.882	0.801	0.839
Trial4 (NCT03735979)	0.601	0.557	0.578	0.633	0.557	0.592
Trial5 (NCT03805308)	0.915	0.902	0.908	0.933	0.900	0.916
Trial6 (NCT03876457)	0.498	0.496	0.496	0.593	0.405	0.481

4.3.3 Generalizability

In this section, we focus on assessing the generalizability of our proposed model in comparison to the baseline (vanilla) model. Our objective is to determine the effectiveness of both models when trained on certain trials and subsequently tested on unseen trials. To conduct this analysis, we designed several experimental scenarios: 1) Training on easy trials and testing on easy trials; 2) Training on easy trials and testing on hard trials; 3) Training on mixed trials and testing on hard trials. The level of difficulty is determined by the performance demonstrated in Section 4.3.2. The results of these experiments are presented in Tab. 3.

Table 3: Performance of generalizability

Performance	Vanilla			LLM-PTM		
	Prec	Rec	F1	Prec	Rec	F1
Case 1	0.793	0.734	0.762	0.855	0.803	0.828
Case 2	0.525	0.452	0.485	0.673	0.604	0.636
Case 3	0.566	0.535	0.550	0.742	0.656	0.696

Upon analyzing the outcomes, it becomes evident that the baseline model exhibits weak performance with respect to generalizability. In contrast, our proposed method demonstrates a notable enhancement in accuracy, achieving an average boost of 12.12% across the different scenarios. In an effort to understand the factors contributing to the observed generalizability when training on easy and hard trials, we consider various scenarios. Both the vanilla model and our method perform similarly when training and testing on easy trials, with our method showing a slight accuracy improvement due to its enriched semantic information. When training on easy trials and testing on hard ones, the baseline model struggles to generalize, while our method excels thanks to its strong adaptability using varied semantic information. Even when training on mixed trials, our method outperforms the vanilla model during testing on hard trials, as it can better comprehend the problem domain. Overall, our proposed method demonstrates the potential to enhance machine learning model performance across different contexts and problem domains by efficiently utilizing enriched semantic information.

4.4 Case Study

In this section, we present a case study to illustrate the strengths of our proposed method through the use of concrete examples. Specifically, we discuss two examples that demonstrate augmentation's capacity to alleviate challenges posed by hard data and to enrich the semantic information of the given problem domain. By showcasing these examples, we aim to provide a deeper understanding of the benefits of our method in practice.

Case 1: Hard Data Easing

Consider a challenging case where traditional machine learning models struggle to accurately classify patients based on their medical records. In this scenario, our augmentation method generates additional criteria by leveraging the enriched semantic information derived from the language model. These newly created criteria assist the model in capturing nuanced relationships and patterns that may have been overlooked by the baseline model. Consequently, the model is better equipped to handle the complexity of hard data, resulting in improved performance and more accurate classification of patients. As shown in Tab. 4, the initial data cause erroneous predictions, whereas the model trained on augmented data displayed accurate classification. This improved performance indicates that augmented datasets can enhance the efficacy of machine learning models.

Table 4: Case study: ease of hard data. We highlight the clear expressions.

Criteria-Level Matching	Criteria	Prediction	Ground Truth
Vanilla	Positive urine or serum pregnancy test for women of child bearing potential.	Match	
LLM-PTM	Women of reproductive age have received a positive result on their urine or serum pregnancy test.	Mismatch	Mismatch
	Women who are capable of bearing children have a positive pregnancy test in their urine or serum.	Mismatch	

Case 2: Semantic Enrichment

In a second example, let us consider a situation where the initial dataset contains limited semantic information, which may hinder the model's ability to make accurate predictions. Our augmentation method addresses this issue by generating a more diverse set of criteria based on the LLM's understanding of the problem domain. By doing so, the method enriches the semantic information available to the model, allowing it to uncover previously undiscovered relationships and patterns within the data. This enriched representation of the problem domain subsequently contributes to better performance. As shown in Tab. 5, the vanilla model exhibited inaccuracies in patient-trial level matching; however, when trained on augmented data, the model demonstrated improved prediction accuracy. This evidence suggests that the incorporation of augmented datasets can enhance the machine learning model's capacity to comprehend specific criteria and establish associations with other related criteria.

Table 5: Case study: semantic enrichment. We highlight the diverse expressions.

Trial-Level Matching	Criteria within one trial	Prediction	Ground Truth
Vanilla	Acute ischemic stroke patients.	Mismatch	
LLM-PTM	Patients suffering from a sudden blockage of blood flow to the brain due to ischemia. Individuals experiencing a sudden onset of neurological deficits resulting from a lack of blood supply to the brain. People with an abrupt interruption of blood flow to the brain caused by an ischemic event.	Match	Match

Through the exploration of these examples, the case study highlights the potential of our proposed method to tackle challenges associated with hard data and enrich semantic information. This, in turn, contributes to a more robust and comprehensive understanding of the problem domain, ultimately leading to improved performance of machine learning models in various contexts.

5 Related Works

Patient Trial Matching. Existing patient-trial matching methodologies can be classified into rule-based systems and deep embedding-based models. Rule-based systems endeavor to extract named entities and relationships for trial eligibility criteria (ECs) and devise rules for patient identification. These systems rely either on extensive human annotations such as EliXR [27], supervised learning classifiers for rule extraction [28], or a combination of machine learning and rule-based approaches such as Criteria2Query [29] to establish rules for ECs. In recent years, deep embedding-based models, such as DeepEnroll [6], have emerged. These models jointly embed patient records and trial ECs in a shared latent space, subsequently aligning them using attentive inference. COMPOSE[5] represents the state-of-the-art method, it used a fine-grained model design to differentiate between inclusion and exclusion criteria. In this work, we follow COMPOSE as our baseline model structure.

Data Augmentation. In natural language processing (NLP), data augmentation techniques operate at character, word, sentence, and document levels. Character-level methods, such as random insertion, exchange, replacement, or deletion, improve model resilience against textual noise [30, 31]. Word-level techniques, like random swap, deletion [22], synonym augmentation [32], and word embeddings-based approaches [33, 34], enhance text classification tasks while maintaining semantic consistency. Contextual augmentation uses MLMs such as BERT [16] and RoBERTa [35] for generating relevant new text [36]. Sentence and document-level methods include back translation augmentation [25]. Patient-trial matching faces similar issues in obtaining training data, our proposed method shows that LLMs can help acquire high-quality data.

6 Conclusion

In summary, this study has successfully developed and implemented a novel data augmentation method called LLM-PTM, designed to enhance the patient-trial matching process while preserving patient privacy. By leveraging the power of LLMs for data augmentation, our approach demonstrated a significant performance improvement, with a 7.32% increase over the baseline model. Moreover, LLM-PTM proved to be highly generalizable, exhibiting a substantial improvement of 12.12% when applied to new datasets. The promising results obtained from the implementation of LLM-PTM highlight the effectiveness of using LLM-empowered data augmentation techniques in the patient-trial matching pipeline. The success of this method has the potential to revolutionize patient-trial matching processes, contribute to improved clinical trial outcomes, and expedite the development of novel therapeutic interventions. The study also paves the way for future research exploring the extension of LLM-PTM to other domains in the healthcare industry, ultimately optimizing data-driven decision-making processes.

References

1. Mattson ME. Patient-Treatment Matching: Rationale and Results. Alcohol health and research world. 1994;18(4):287-95.
2. Hassanzadeh H, Karimi S, Nguyen A. Matching patients to clinical trials using semantically enriched document representation. Journal of biomedical informatics. 2020 May;105:103406.
3. 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 Jul;3(2):209-15.
4. Bhatt A. Artificial intelligence in managing clinical trial design and conduct: Man and machine still on the learning curve? Perspectives in clinical research. 2021 Jan;12(1):1-3.
5. Gao J, Xiao C, Glass LM, Sun J. COMPOSE: cross-modal pseudo-siamese network for patient trial matching. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining; 2020. p. 803-12.
6. Zhang X, Xiao C, Glass LM, Sun J. DeepEnroll: patient-trial matching with deep embedding and entailment prediction. In: Proceedings of The Web Conference 2020; 2020. p. 1029-37.
7. OpenAI. GPT-4 Technical Report; 2023.
8. Hargreaves B. Clinical trials and their patients: the rising costs and how to stem the loss. Pharmafile URL http://www pharmafile com/news/511225/clinical-trials-and-their-patients-rising-costs-and-how-stem-loss. 2016.
9. Price WN, Cohen IG. Privacy in the age of medical big data. Nature medicine. 2019;25(1):37-43.
10. Price W, Nicholson I. Drug Approval in a Learning Health System. Minn L Rev. 2017;102:2413.

11. Eisenberg RS, Price WN. Promoting healthcare innovation on the demand side. Journal of Law and the Biosciences. 2017;4(1):3-49.
12. Evans B. Big data and individual autonomy in a crowd. Big Data, Health Law, and Bioethics. 2018.
13. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-preserving generative deep neural networks support clinical data sharing. Circulation: Cardiovascular Quality and Outcomes. 2019;12(7):e005122.
14. Dwork C, Roth A, et al. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science. 2014;9(3–4):211-407.
15. Moussa M, Demurjian SA. Differential privacy approach for big data privacy in healthcare. In: Privacy and Security Policies in Big Data. IGI Global; 2017. p. 191-213.
16. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.
17. Weston J, Chopra S, Bordes A. Memory networks. arXiv preprint arXiv:14103916. 2014.
18. Srivastava RK, Greff K, Schmidhuber J. Highway networks. arXiv preprint arXiv:150500387. 2015.
19. You Q, Zhang Z, Luo J. End-to-end convolutional semantic embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 5735-44.
20. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv preprint arXiv:190403323. 2019.
21. Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3(1):1-9.
22. Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:190111196. 2019.
23. Kobayashi S. Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:180506201. 2018.
24. Kumar V, Choudhary A, Cho E. Data augmentation using pre-trained transformer models. arXiv preprint arXiv:200302245. 2020.
25. Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. arXiv preprint arXiv:151106709. 2015.
26. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.
27. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. Journal of the American Medical Informatics Association. 2011;18(Supplement_1):i116-24.
28. Bustos A, Pertusa A. Learning eligibility in cancer clinical trials using deep neural networks. Applied Sciences. 2018;8(7):1206.
29. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association. 2019;26(4):294-305.
30. Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine translation. arXiv preprint arXiv:171102173. 2017.
31. Liu Z, Jin H, Wang TH, Zhou K, Hu X. DivAug: plug-in automated data augmentation with explicit diversity maximization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 4762-70.
32. Niu T, Bansal M. Adversarial over-sensitivity and over-stability strategies for dialogue models. arXiv preprint arXiv:180902079. 2018.
33. Wang WY, Yang D. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In: Proceedings of the 2015 conference on empirical methods in natural language processing; 2015. p. 2557-63.
34. Mrkšić N, Séaghdha DO, Thomson B, Gašić M, Rojas-Barahona L, Su PH, et al. Counter-fitting word vectors to linguistic constraints. arXiv preprint arXiv:160300892. 2016.
35. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019.
36. Tang R, Han X, Jiang X, Hu X. Does Synthetic Data Generation of LLMs Help Clinical Text Mining? arXiv preprint arXiv:230304360. 2023.