

Applications of Federated Large Language Model for Adverse Drug Reactions Prediction: A Scoping Review

David Guo, Kim-Kwang Raymond Choo

Submitted to: Journal of Medical Internet Research
on: November 01, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 4

Supplementary Files..... 29

..... 29

Figures 30

Figure 1..... 31

Multimedia Appendixes 32

Multimedia Appendix 0..... 33

Applications of Federated Large Language Model for Adverse Drug Reactions Prediction: A Scoping Review

David Guo¹ PhD; Kim-Kwang Raymond Choo¹ PhD

¹The University of Texas at San Antonio San Antonio US

Corresponding Author:

Kim-Kwang Raymond Choo PhD
The University of Texas at San Antonio
1 UTSA Circle
San Antonio
US

Abstract

Background: Adverse drug reactions (ADRs) pose significant challenges in healthcare, where early prevention is vital for effective treatment and patient safety.

Objective: Traditional supervised learning methods are limited in addressing healthcare data, which is often unstructured, heavily regulated, and involves restricted access to sensitive personal information.

Methods: The integration of Federated Learning (FL) and Large Language Model (LLM) offers a promising solution to these challenges since FL supports the distributed training on edge device with limited resources and the capability of LLM to deal with unstructured healthcare data. Additionally, client models trained on the edge device can be merged into a global model on the server, preserving data privacy.

Results: Natural Language Processing (NLP) technologies underpinning LLM provide a full set of tools that can readily be used to process unstructured ADR as input, enabling LLM to predict ADR outcome effectively. The ADR output space can be discrete labels, unstructured texts, or both.

Conclusions: This review presents a scoping review following the PRISMA protocol on the applications of Federated Large Language Model (FedLLM) in ADR prediction, aiming to explore future research venue on ADR applications

(JMIR Preprints 01/11/2024:68291)

DOI: <https://doi.org/10.2196/preprints.68291>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a JMIR journal, I will be able to make the full manuscript visible to all users.
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Applications of Federated Large Language Model for Adverse Drug Reactions Prediction: A Scoping Review

David Guo and Kim-Kwang Raymond Choo

Department of Information Systems and Cyber Security
The University of Texas at San Antonio

Abstract

Background: Adverse drug reactions (ADR) present significant challenges in healthcare, where early prevention is vital for effective treatment and patient safety. Traditional supervised learning methods struggle to address heterogeneous healthcare data due to their unstructured nature, regulatory constraints and restricted access to sensitive personal identifiable information.

Objective: This review aims to explore the potential of Federated Learning (FL) combined with Natural Language Processing (NLP) and Large Language Models (LLMs) to enhance ADR prediction. FL enables decentralized training across client clusters with limited resources, while LLMs effectively process unstructured healthcare data. By aggregating client-trained models into a global model, FL ensures broader data inclusion while maintaining privacy.

Methods: A scoping review was conducted on peer-reviewed publications retrieved from Google Scholar and Semantic Scholar between 2019 and 2024.

Results: Following the PRISMA protocol, 145 articles from PubMed, arXiv, IEEE, and ACL Anthology met the inclusion criteria. Of these, 12 articles were selected for an in-depth review to examine use cases in ADR prediction. We synthesized ADR data sources on structured and unstructured data types, use cases of FL integrated with NLP, and open-source frameworks for ADR identifications and predictions. Special attention is given to unstructured ADR prediction using FL with LLMs (FedLLM), including development and deployment strategies and evaluation metrics.

Conclusions: Given the recent emergence of LLM, the integration of FL and LLM for ADR prediction remains in its early stage with limited documented use cases. This review explored the potential applications and highlighted the advancements of FedLLM in healthcare research, particularly in ADR prediction. Key focus areas include fine-tuning and merging algorithms, fairness and unbiasedness, implementation challenges, and real-world deployment strategies. By synthesizing current insights, this review aims to lay the groundwork for future research in privacy-preserving and scalable ADR prediction systems.

Keywords

adverse drug reactions; [distributed systems](#); [federated learning](#); scoping review; large language model; open-source; fine-tune

Introduction

Adverse drug reactions (ADR) are a major concern in the field of medicine, posing serious risks to patient safety. It is estimated that as many as 98,000 people die each year from medical errors that occur in hospitals[1]. This figure surpassing deaths reported from motor vehicle accidents, breast cancer, or AIDS [2], [3]. In the US, ADR have been reported to Adverse Event Reporting System (FAERS) which maintained by the U.S. Food and Drug Administration (FDA) since 2004. In 2022 alone, over 1.25 million serious adverse events were reported with nearly 175,000 fatalities.

ADR can range from mild symptoms, such as nausea and dizziness, to severe and life-threatening conditions like anaphylaxis and cardiac arrest[4]. As the number of available drugs on the market continues to grow, early identification of potential ADR is crucial to minimize adverse risks and improving patient safety. Reliable prediction of ADR can play a critical role in preventing harmful medical outcomes. However, healthcare data are highly regulated due to privacy protection laws such HIPPA in the US and GDPR in Europe. Given the distribution of vast amount of sensitive healthcare data across different locations, Federated learning (FL) has been emerged and proved as a promising solution to address data privacy concerns by enabling the training of machine learning models on decentralized data sources while maintaining data privacy and security[5], [6], [7]. Research on FL in healthcare, particularly for ADR prediction, is still in its early stages. Most current studies focus on structured outcome prediction. Comparisons between centralized learning and distributed learning like FL, some research have shown FL for ADR prediction with electronic health records (HER) can achieve higher accuracy than centralized learning approaches[8], [9], [10].

A major challenge in healthcare data utilization is that most of the data exists in unstructured and heterogeneous formats that limit its full potential. Recent advancements in natural language processing (NLP) techniques coupled with enhanced computational power from the latest GPU hardware have demonstration the abilities to uncover hidden patterns and relationships in public datasets[11]. One of the most successful applications of NLP is the Large Language Model (LLM), a type of artificial intelligent model pre-trained on vast amounts of text data to perform NLP tasks with high accuracy and adaptability. However, no open-source LLMs have ever been pre-trained on public or private ADR datasets given the high costs and technical barriers. Fine-tuning the open-source LLM on domain-specific data has emerged as a practical approach to gain new domain knowledge on ADR. By learning linguistic patterns and the relationship between inputs and outcomes of ADR training datasets, an adapted LLM can improve its predictive capabilities for ADR in both existing or newly developed drugs[12], [13].

Most FL applications have been applied to parameterized, gradient, or transformer based machine learning models[14], [15], [16]. The integration of FL with LLM, known as FedLLM, has gained traction since the emergence of LLMs in 2022. By leveraging FL techniques to train LLMs on distributed data, FedLLM offers several advantages over traditional centralized training procedures[17], [18]. This approach inherently enhances data privacy and security by training the model on decentralized data sources. It therefore mitigates the risks associated with centralized data repositories. FedLLM maximizes the potential of LLMs in processing unstructured data without requiring feature engineering, as needed in traditional ML methods. Additionally, it enables the inclusion of diverse and heterogeneous ADR data from both public and private sources, leading to more accurate ADR predictions with broader data access. Furthermore, the FedLLM is highly scalable for the integration of multimodal data beyond text and adapt to evolving healthcare needs. Altogether, FedLLM make it a promising solution for future ADR prediction and analysis.

This paper focuses on the application of FedLLM for ADR prediction, highlighting its advantages and broader potential applications in the healthcare industry. We also explore how FedLLM can facilitate collaboration among researchers, institutions, and the healthcare sector, enabling more comprehensive and dynamic ADR research with innovative technologies. A key benefit of FedLLM applications is their independence from

original training data, allowing for real-time and personalized ADR risk assessment based on unstructured individual patient characteristics. Additionally, we examine the challenges associated with the implementation of FedLLM in healthcare and outline future directions for this promising area of research.

This scope review follows the guidelines of the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) framework[19]. Our objective is to assess the progress of FedLLM in healthcare, with a particular focus on the potential for Adverse Event (AE) or ADR research. By evaluating its effectiveness in handling sensitive, heterogeneous, and unstructured data, we aim to provide a deeper understanding of its applications in ADR identification and prediction. As a result, we hope this review offers valuable insights for various stakeholders, including health care providers, practitioners, and policy makers, into the limitations and challenges associated with using FedLLM beyond ADR prediction. These insights will help informed decision-making regarding its broader application in healthcare initiatives. The remainder of this paper is structured as follows. **Section 2** discusses review protocol, **Section 3** outlines the findings on the FedLLM relevant to ADR research, and **Section 4** Discussions and Conclusion.

Methods

Information sources and search criteria

Our literature search was conducted using 2 search engines, google scholar and semantic scholar, for peer-reviewed articles published since 2019. In both cases, an initial query statement was conducted with keyword combinations as "language model" ("adverse event" OR "adverse drug") "federated learning" on google scholar. The query statement on semantic scholar is "language model federated learning adverse events drug reactions prediction".

Inclusion and Exclusion Criteria

This study focuses on the review of applications of FedLLM on a broader topic of Adverse Events, in particular Adverse Drug Reactions. Specifically, it examines how FedLLM could or has been used to train a language model on public and private data with restricted access within the client storage space. As such, articles regarding the general FL techniques with no applications on ADR or pre-defined features for structured ADR outcome prediction are excluded. Additionally, only articles published in peer-reviewed journals and conference proceedings with PDF available are considered, whereas other types of publications, such as dissertation paper, conference abstracts, book chapters, editorials, and commentaries were excluded. Also, the articles referenced in this review are limited to publication in English.

Data Extraction

A comprehensive search on journal articles was conducted as of July 15, 2024, focusing on publications from 2019 onward. After removing non-relevant and duplicates articles from two search engines mentioned Section 2.1, only those meeting the inclusion criteria were selected: 1) using unstructured data, 2) incorporating FL, and 3) having full text or PDF copy available.

Figure 1 demonstrates the data sources and search flowchart. Google scholar (GS) returns a total of 83 articles, including 28 review papers. Together with 62 articles returned from Semantic Scholar (SS) search platform, a total of 145 articles were used as raw inputs. Each full-text paper was reviewed thoroughly to support quantitative and qualitative synthesis of the data.

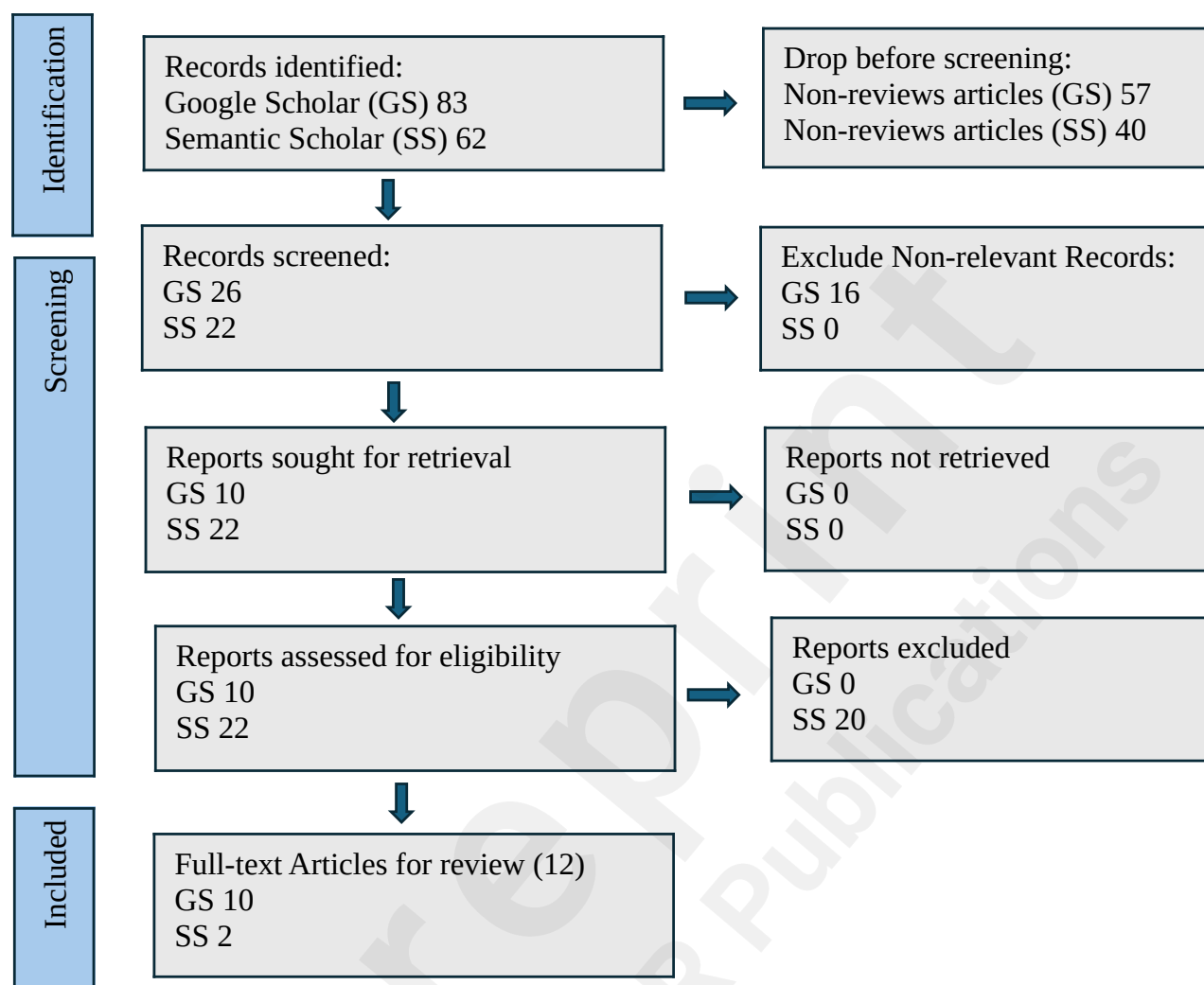


Figure 1. Literature Search Strategy for FedLLM in ADR Prediction

Characteristics of Sources of Evidence

The included articles have a mix of diverse ranges of informatics and biomedical literature, including systematic reviews, meta-analyses, and original research articles sourced from PubMed, arXiv, IEEE, ACL Anthology, and other public databases. To minimize selection bias, we employed a 2-stage screening process involving title/abstract review followed by full-text assessment, ensuring transparent and reproducible study selection.

This study synthesizes findings from 12 key review articles identified through systemic search of Semantic Scholar and Google Scholar. Selected articles focused on ADR-related research with primary emphasis on either language model, federated learning, or both. We extracted and compared 4 critical dimensions from these reviews: 1) primary research tasks, 2) scope of ADR coverage, 3) Methodological approaches focusing on techniques used, and 4) reported limitations or challenges as shown in **Table 1**.

Table 1: Comparative Summary of Review Articles

Review Articles	Tasks	ADR Coverage	Techniques used	Limitations or Challenges
DeepHealth: Review and challenges of artificial intelligence in health informatics (2020)[20]	1) Imputation methods for missing data 2) Annotation 3) Clinical events 4) prediction by ML, LR, RF, NN	Named entity recognition of ADR	1) Word embedding and unsupervised learning 2) Skip-gram representation 3) CNN with word2vec 4) DeepCare framework 5) LDA for topic modeling 6) SDA for generalizing clinical notes	1) accessibility of health data is limited 2) no continuous chronic illness data 3) unstructured online data 4) data collection is not consistent
A Comprehensive Review and Application of Interpretable Deep Learning Model for ADR Prediction (2022)[21]	1) query-based search 2) data integration 3) data charting 4) summarization	1) pharmacovigilance centers 2) Spontaneous reporting system 3) Drug-ADR association 4) ADR prediction 5) geographic distribution of ADR research	1) dis-proportionality analysis 2) fuzzy decision making 3) temporal association mining 4) ML and DL 5) NLP	1) limited to drug-drug interaction and genetic interaction of drugs 2) limited to clinical research
A Review of Federated Learning: Algorithms, Frameworks and Applications (2023)[22]	1) joint training 2) integration of training datasets 3) prevent data leakage 4) transfer learning for small dataset	ADR and mortality rate	1) Vertical FL 2) differential privacy 3) secure aggregation 4) homomorphic encryption 5) Federate transfer learning	1) data leakage during transfer 2) small data fail to train global model 3) data share between deep networks 4) statistical heterogeneity of model 5) limited overlap of data
Adverse drug event detection using natural language processing: A scoping review of supervised learning methods (2023)[23]	1) ensemble method-based ADR extraction 2) identify ADR in clinic notes 3) detect ADR from EHR	ADR detection methods by supervised NLP	supervised NLP	1) lack of assessment tool for clinic NLP 2) limited options to match ontology 3) lack of data preparation and deployment for NLP
AI in Health: State of the Art, Challenges, and Future Directions (2019)[24]	1) NLP for clinic narrative 2) multi-view ensemble classification 3) PubMed tools 4) interpretable ML 5) CLAMP clinic NLP pipeline	1) Data mining on FAERS 2) drug similarity graph for drug-drug interaction prediction 3) map drug names and outcomes	1) multi-view ensemble learning for classification 2) PubMed Phrases, Labs, LitVar 3) scalable and accurate DL 4) Graph-CNN 5) EHR and Genomics (eMERGE) 6) topological analysis 7) model interpretation	1) NLP algorithms are error-prone 2) Ultra-high sparsity of EHR data 3) RNN-based models weak on long-term dependencies 4) NLP needs well-defined rules/regular expressions
Causality Mining in Natural Languages Using Machine and Deep Learning Techniques: A Survey (2021)[25]	1) Mining valuable relational information, 2) Automated knowledge extraction, 3) corpus associated,	causality relations of ADR	1) DL 2) MXNet 3) Caffe, Theano 4) CNTK 5) Neon DL 6) Gluon	1) extract causation knowledge from text 2) relationship of syntax, semantics, vocabulary, and context in textual data

	4) Exploring the causality effect,			
Handling temporality of clinical events with application to Adverse Drug Event detection in Electronic Health Records: A scoping review (2024)[26]	1) Data mining 2) Temporal abstraction of EHR Series 3) Time Series Classification	Temporality of EHR data for ADE identification and detection	1) Temporal abstraction 2) ML prediction 3) Data mining	1) Limited to ADR on EHR data 2) limited to classification of multivariate TS and temporal abstraction features
Machine and cognitive intelligence for human health: systematic review (2022)[27]	1) public health surveillance 2) AI in healthcare 3) Web-based UI 4) Ontology development	ADR information extraction with data mining	1) NLP 2) web-based semantic system 3) healthcare support system 4) Real-time case finding algorithm 5) topic modeling	1) Lack of standardized procedure 2) Data quality concern 3) Scalability and generalizability issues 4) Ethical consideration 5) Lack of transparency
Differentially Private Federated Learning: A Systematic Review (2024)[28]	Differentially private FL for healthcare applications – classification, topic modeling and recommender systems	differential privacy in FL with EHR for ADR prediction	1) Poisson binomial RDP 2) LWE RDP 3) discrete Gaussian RDP 4) advanced sequential composition theory 5) basic sequential composition theory	1) limited to a few specific techniques 2) limited to subset of applications 3) insufficient discussion of challenges 4) few concrete case studies
The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review (2024)[29]	Summarize and report the applications of LLMs in ER by case studies	ADR identification and audits	1) ChatGPT 2) PaLM and its variants 3) BERT and its variants	1) lack of discussion on FL w/ data security concern 2) lack of coverage on open-source LLMs 3) lack of comprehensive evaluation on studies
Assessing the performance of large language models in literature screening for pharmacovigilance: a comparative study (2024)[30]	Application of LLMs for literature screening in healthcare.	Systematic monitoring of ADR and the detection of potential safety concerns related to drugs.	1) prompt engineering of LLM 2) Fine-tuning LLMs	1) less coverage on open-source LLMs 2) less discussion on the LLM applications
The Use of Artificial Intelligence in Pharmacovigilance: A Systematic Review of the Literature (2022)[31]	Artificial Intelligence in patient safety and Pharmacovigilance	Modeling algorithms for ADR signal identification, characterization, assessment, and management	1) classification algorithms 2) text mining for safety narratives	1) more experimentation studies than concrete case studies 2) lack of discussion on the integration of social media
This review paper	1) FL w/ LLM 2) Performance matrix 3) ADR data sources 4) Implementation pipeline 5) Development and deployment options	Prediction of structured and unstructured ADR outcomes by FedLLM	1) BERT-based encoder language models 2) BERT-based embedding models 3) Open-source LLMs 4) RAG 5) Fine-tuning language models 6) open-source framework	

Results

Language Model for ADR Prediction

Given the nature of unstructured healthcare data including patient demographic data, medication administration, treatment purpose, leveraging the power of language models is a promising approach for ADR prediction. Many current efforts utilize pretrained BERT as the base model to fine-tune on a customized dataset for downstream task[32]. Through fine-tuning on a large amount of unstructured medical domain data, these models can acquire domain-specific medical knowledge also known as transfer learning. However, BERT by itself, as an encoder focused on representation learning, is not sufficient to fulfill a prediction task on a defined target. To address this issue, additional decoder layers should be developed on top of encoder, such as bidirectional LSTM can be incorporated on the top of BERT to fulfill the prediction task of structured targets. Thus, BERT framework usually contains 2 steps: pre-training and fine-tuning, the first step is unsupervised trained on the unlabeled data, the second step will use the model parameters from the first step to fine-tune the labeled data, such as binary outcomes ADR (1) and Non-ADR (0)[33], [34].

Recently, Generative Per-trained Transformers (GPT), a type of LLM represented by ChatGTP from OpenAI, have been pre-trained on huge amounts of unannotated and unstructured public data using thousands of high profile Nvidia GPUs over several months. These models are designed to generate human-like responses in interactive conversations with strong ability to understand and follow user instructions or prompts[35]. This generative approach expands possibilities for extracting insights and performing complex ADR-related tasks and enables more efficient and adaptive analysis in response to specific queries.

Summarization of ADR Types

According to the International Conference on Harmonization, ADR is defined as “A response to a drug which is noxious and unintended, and which occurs at doses normally used for prophylaxis, diagnosis, or therapy of disease or the modification of physiologic function.”[36], Based on symptoms, and onset time, and nature, ADR are commonly categorized into three primary types. Type 1 reactions can be referred to as augmented reactions, which are dose-dependent and predictable according to the pharmacology of the drug. Type 2 reactions can be referred to as time dependent reactions. This type of ADE usually shows chronic and delayed effects. Type 3 reactions are unexpected reactions other the type 1 and type 2, making them idiosyncratic and unpredictable from the perspective of drug pharmacology. Beyond dosage, additional factors can also contribute to the ADR such as the duration of the reaction and susceptibility factors, including genetic, pathological, drug-to-drug interactions, and other demographic and biological differences. Taking all these factors into consideration will help mitigate any possible ADR and reduce the risk of harm to the patient in clinic practice.

In view of reported outcomes, ADR can be classified into 2 classes, 1) unstructured outcome in text, such as the reaction files of FAERS quarterly release from FDA report ADR in preferred term to describe the event by using the Medical Dictionary for Regulatory Activities (MedDRA)[37]. 2) structured outcome summarized in discrete codes, such as the outcome codes in FAERS quarterly release defining discrete outcome in medical code for every primary id, e.g. “DE” for death; “LT” for life-threatening; “HO” for hospitalization – initial or prolonged; “DS” for disability; “CA” for congenital anomaly; “RI” for required intervention to prevent or permanent impairment/damage; and “OT” for serious medical events. Similarly, the adverse events are graded by severity from 1 to 5 according to the Common Terminology Criteria for Adverse Events (CTCAE) version 5.0 (see details in **Table 2**).

ADR Benchmark Data

Numerous ADR benchmark data are publicly available from the surveillance practice enforced by regulatory agencies and some independent research institutes. Regulatory agencies such as FDA in US, the Therapeutic Goods Administration (TGA) in Australia, the European Medicine Agency (EMA) in EU, and the Canada Vigilance Program in Canada, collect volunteer reports of adverse drug side effects, investigate the reports, and issue safety alerts if a drug is suspected of causing an adverse effect. A summary of publicly available ADR benchmark data is summarized in **Table 2**.

Table 2 Sources of ADR Data ^a

ADR Data Base	Data Type	Data Sources and number of records
Adverse Drug Event Corpus (Cadec)[38]	Unstructured Structured	and AskaPatient, a web-based medical forum where patients can share their experiences with medications, including drug, adverse effect, disease, symptom, finding. The text annotated with ADR are mapped to the corresponding MedDRA term.
European Medicine Agency (EMA) [39]	Unstructured Structured	and The European Medicines Agency (EMA) coordinates the European Union (EU) pharmacovigilance system and operates services and processes to support pharmacovigilance in the EU. The database is named as Union Pharmacovigilance Database (UPD)
FDA Adverse Event Reporting System (FAERS)[40]	Unstructured Structured	and FDA's post-marketing safety surveillance program collects and reports medication errors and product quality complaints resulting in adverse events, and release cases on a quarterly basis
Canada Vigilance adverse reaction online database (CAERS)[41]	Unstructured Structured	and The Canada Vigilance Adverse Reaction Online Database contains information about suspected adverse reactions (also known as side effects) since 1965. It is grouped by 26 system organ classes and reported by: 1) consumers and health professionals, who submit reports voluntarily. 2) manufacturers and distributors (also known as market authorization holders), who are required to submit reports according to the Food and Drugs Act.
Database of Adverse Event Notifications (DAEN) - medicines[42]	Unstructured Structured	and The Database of Adverse Event Notifications (DAEN) - medicines contains information from reports of adverse events (also known as side effects or adverse reactions) received in relation to medicines, vaccines and biological therapies used in Australia since 1971.
Social Media Mining for Health (SMM4H)[43]	Structured	Ten tasks include English and Spanish posts on Twitter/X, Reddit, and WebMD.
National NLP Clinical Challenges (n2c2)[44]	Unstructured Structured	and shared task on adverse drug events and medication extraction in electronic health records primarily utilized RNN consisting of bidirectional LSTM units
ADE_Corpus_V2[45]	Unstructured Structured	and This is a dataset contains Adverse Drug Reaction Data for Classification if a sentence is ADE-related (True) or not (False) and Relation Extraction between Adverse Drug Event and Drug
Common Terminology Criteria for Adverse Events (CTCAE) v5.0[46]	Unstructured Structured	and The NCI Common Terminology Criteria for Adverse Events is a descriptive terminology which can be utilized for Adverse Event (AE) reporting. A grading (severity) scale is provided for each AE term.

^a Information Accurate as of July 15, 2024

Evaluation Matrices on Prediction Accuracy

For structured ADR outcome evaluation, metrics such as F-1 score, recall, and precision are commonly applied. However, given the predominantly unstructured nature of healthcare data, this review focuses on evaluating unstructured prediction outcome leveraging the power of NLP technologies. The predicted outcome will emulate the responses by different individual professional or non-professional to a specific ADR event with different yet semantically similar texts. The core evaluation of NLP based ADR prediction involves either leveraging another language model with domain knowledge to evaluate the performance of ADR prediction or incorporating a human-in-the-loop approach, where healthcare professionals review and validate the predictions. A summary of their strengths and limitations of each evaluation approach is provided in **Table 3**.

Table 3 Evaluation Metrics on Unstructured Predictions

Matrix	Description	Strength	Limitation	Equation
G-Eval[47]	This metric uses LLMs with chain-of-thoughts (CoT) to evaluate LLM outputs based on any custom criteria	It is easy to compute and does not require any additional data.	It may not capture the quality of the generated instructions beyond the accuracy of the predicted ADR outcome	$Score = \sum_{i=0}^n p(s_i) * s_i$ where S_i are the predefined scores, and $p(S_i)$ are calculated probabilities by the LLM
BLEU_N[48]	This metric measures the n-gram precision based on the similarity between the generated text and the ground truth instructions	It can capture the quality of the generated text beyond the accuracy of the predicted text	It may not be suitable for evaluating the accuracy of the predicted adverse drug reaction as it rewards fluent but incorrect text generation	$p_n = \frac{\sum_c \sum_{n-gram} Count(n-gram)}{\sum_{c'} \sum_{n-gram'} Count(n-gram')}$ where C denotes the reference count, C' denotes the reference count, n-gram denotes the clip count.
ROUGE_N[49]	This metric measures the n-gram recall between candidate and reference set based on the overlap between the generated text and the ground truth instructions	It can capture the quality of the generated text beyond the accuracy of the predicted text,	It may not be suitable for evaluating the accuracy of the predicted text as it rewards overlapping but incorrect text generation.	$R_n = \frac{\sum_{Se} \sum_{n-gram} Count_{match}(gram_n)}{\sum_{Se} \sum_{n-gram} Count(gram_n)}$ Where Se denotes reference summary, $Count_{match}(gram_n)$ denotes the maximum number of n-grams co-occurring in candidate and reference summaries
BERTScore[50]	This metric measures a similarity score for each token in the candidate sentence with each token in the reference sentence. Instead of focusing on exact matches, token similarity between the sentences is computed using BERT contextual embeddings.	It can capture the semantic similarity between the generated text and observed text.	BERT model is not trained on ADR data; thus, the BERT embeddings may not fully represent the tokens from ADR domain.	$R_{Bert} = \frac{1}{ X } \sum_{x_i} \frac{\max}{\hat{x}_j} x_i^T \hat{x}_j,$ $P_{Bert} = \frac{1}{ \hat{X} } \sum_{\hat{x}_i} \frac{\max}{x_j} x_i^T \hat{x}_j,$ $F_{Bert} = 2 \frac{P_{Bert} * R_{Bert}}{P_{Bert} + R_{Bert}}$ Where R_{Bert} denotes the recall of tokens, P_{Bert} denotes the precision of tokens, F_{Bert} denotes the F1 measure.
SEMScore[51]	This metric computes the semantic textual similarity between embeddings of model and target. It consists of two steps, 1) it embeds model and target responses separately using the chosen sentence transformer. 2) it computes the cosine similarity of the respective embeddings as the value of SEMScore.	It can capture the semantic similarity between the generated text and observed text based on the cosine similarity of embeddings.	Like the BERT score, it also depends on the BERT model with domain knowledge.	$\cos(\theta) = \frac{A \cdot B}{\ A\ \ B\ } = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$ Where A denotes the embedding vector of predicted texts, B denotes the embedding of embedding of observed texts
ANLS*[52]	This metric computes the Average Normalized Levenshtein Similarity (ANLS) with penalty on unanswerable questions for a wide variety of tasks, including information extraction and classification tasks.	It is general metric applicable for a wide variety of tasks with varied output formats, either structured or unstructured.	It can only deal with strings and lists but not dictionaries or any combination of types.	$ANLS^*(g, p) = \frac{s(g, p)}{l(g, p)}$ Where, g denotes ground truth, p denotes prediction, s is the score between the ground truth and the prediction and l is the size of the trees g and p such that $ANLS^*(g, p) \in [0, 1]$

FL and LLM for ADR Research

The research and application of FL and LLM for ADR have demonstrated promising results in enhancing the accuracy of ADR detection and summarization. For instance, FL has been applied to ADR detection tasks by leveraging distributed data from multiple sources, such as EHR and claims data. By federating the learning process, FL can improve the accuracy of supervised ADR detection

without compromising data privacy and security[53]. Additionally the integration of BERT-based transfer learning with unique molecular embeddings allows for the development of multimodal models that effectively combine language understanding and molecular property prediction, thereby enhancing adverse drug event (ADE) classification[54]. Recently, LLMs have been used to summarize and extract ADR-related text, such as clinical notes and medical articles. By using pre-trained language models to encode text, LLMs can capture the semantic meaning of ADR-related phrases and generate accurate summaries. A multi-task learning approach that combined question summarization and RQE (relevance-based question answering) was used to improve the performance of pre-trained language models in ADR summarization[55].

Given the fact that most distributed ADR data are imbalanced, data augmentation-based methods have been proposed to improve the performance of LLMs in ADR summarization. These approaches generate additional training data inspired by a data augmentation method in NLP that generates synthetic sentences for training by replacing some words with syntactically and semantically similar words[56].

Zero-shot-based methods can adapt LLMs to ADR-related tasks without the need for additional labeled training data. These methods typically use transfer learning to leverage pre-trained language models and fine-tune them for ADR-related tasks. This technique is particularly useful for detecting rare or unexpected ADR, as well as for identifying potential drug-drug interactions that may not be captured by traditional ADR detection methods[57].

Discussion

Strategies for FedLLM on ADR

Most reported use cases of FL with language models leverage the knowledge acquired by the pre-trained model on a large dataset and apply it to a domain specific task with a smaller dataset through transfer learning represented by customized fine-tuning process [58], [59], [60]. This process involves taking a pre-trained model and further adapting it to a new task by adjusting its parameters. To facilitate the fine-tuning of language models on client data while reducing computational demands on limited client infrastructure, several strategies should be considered: 1) configure a high drop-out rate on client model; 2) reduce model precision through quantization, such as using 4-bit instead of 16-bit precision; 3) segment each parameter matrix of the large model into sub-matrices with varying reduction ratios; 4) transfer only distilled federated knowledge back to server, for example, at the end of the training process on client data, only the student model from a teacher-student model framework is shared to minimize communication costs.

Before fine-tuning, it is crucial to optimize the process to compromise the low-profile computation capability of client facilities to achieve the best training performance, including a systematic evaluation of options among inclusive FL, exclusive FL, and Hetero FL is necessary to achieve fairness across client and global models[61]. For example, using the distilled model can allow more client data to be included in the training process. Taking Named Entity Recognition (NER) for ADR as an example, adoption of distillation models such as DistilBERT to fine-tune ADR data achieve even better performance than fine-tuned large language models xlm-RoBERTa and GPT-Neo-125M in all 4 metrics, Accuracy, Precision, Recall and F1-score[62]. In an evaluation study for ADE extraction, PubMedBERT, distilled from GPT-3.5 with merely 0.1% of model size, outperformed its parent model by over 6 absolute points in F1 score, and a distilled model from GPT-4 achieved an improvement of 5 absolute points[63].

Fine-tuning Framework for ADR FedLLM

There are 2 types of LLMs, closed source and open-source. The closed source models like ChatGPT and Gemini, which are pre-trained on public data without ADR domain knowledge. The weights of models are not publicly available, we therefore discuss the fine-tuning strategies limited to open-source pre-trained models shared in Hugging Face website with Apache-2.0 license, such as Llama-3, Gemma, Mistral/ Mixtral, Falcon, DeepSeek, Qwen[64]. The fine-tuning process involves the preparation of ADR instruction data as new training data for a domain specific task, in this case, the prediction of structured and unstructured ADR outcomes. A widespread practice of fine-tuning LLM usually applies Low-Rank Adaptation (LoRA) which freezes the weights of the pre-trained model and injects trainable rank decomposition matrices into each layer of the LLM transformer architecture. This process significantly reduces the number of trainable parameters down to less than 1% of original parameters given the selection of rank[65], [66].

The implementation of FedLLM can utilize frameworks and libraries that support FedLLM, such as TensorFlow Federated (TFF) a framework for implementing federated learning algorithms using TensorFlow, PySyft for PyTorch developed by OpenMined, Federated AI Technology Enabler (FATE) a federated learning framework by WeBank, and Federated Learning Application Runtime Environment (FLARE) by Nvidia[67]. In all cases, the general workflow includes: 1) declare LLM and define FL parameters; 2) train client data on client infrastructure with shared parameters from global model and return the model weights to server which will be used to update the global model with aggregated weight. This process could be an interactive process; and 3) model evaluation on validation dataset and deployment of global model for low latency and scalable inference.

After a client model being fine-tuned, only model weights or gradients instead of private training data will be shared to the server and merged with global model into an updated new model. The straightforward method to merge client model parameters is to average the weights of client and global models[68]. Other advanced

merging process includes SLERP that takes weight of each model into consideration[69], DARE that can effectively eliminate 90% or even 99% delta parameters of multiple supervised fine-tuned models[70], and Model Stock approach that approximates a center-close weight using only two fine-tuned models[71].

Federated Learning on Embedding Models

Ideally, a FedLLM pre-trained or fine-tuned ADR data would be preferable for direct ADR prediction. As of the time of this review, none of the open-source LLMs have been pre-trained for ADR use cases. Alternatively, for Retrieval-Augmented Generation (RAG) based ADR prediction, any open-source LLM without ADR domain knowledge can indirectly predict ADR outcome by leveraging the retrieved evidence from similarity search between the prompt embedding and pre-populated embeddings in vector database. This approach allows any LLM to function as a summarization engine. Compared with the LLM, fine-tuning embedding model is more manageable, as transformer-based embedding model usually is typically much smaller than LLM by an order of 10 to 100 magnitudes. This reduced model size translates to less computation resource requirement for fine-tuning an embedding model on a small set of private ADR data, thereby gaining distributed ADR domain knowledge.

With embedding model, the distributed ADR specific terms and concepts, such as diseases, symptoms, treatments can be transformed into dense vectors also known for embeddings which capture the semantic relationships between medical entities, allowing AI models to work within a mathematical framework[32], [72], [73]. Since most embedding models are rooted on BERT model, the length of embeddings is typically 512 or more, the longer the embedding, the better ADR information will be represented. These embeddings can be indexed and hosted in a vector database for rapid and accurate content search within a RAG pipeline. Most open-source vector database options use Hierarchical Navigable Small World (HNSW) as vector index type. This graph-based algorithm efficiently identify nearest neighbors among large, multi-dimensional vectors without requiring a full scan the entire vector database[74].

Like the fine-tuning FedLLM, federated learning on ADR embedding model can also be accomplished by fine-tuning a selected foundation embedding model on client data within resource constrained infrastructure with customized settings. Some embedding foundation models are not BERT based models such as Word2Vec in Hugging face website, which is trained on biomedical text to learn the vector representations of medical terms based on their co-occurrence patterns in large text corpora like PubMed articles. Most foundation embedding models, however, are BERT-based. For example, BioBERT is specifically fine-tuned on biomedical text to provide contextual embeddings for biomedical entities[75]. It can enhance performance on tasks like named entity recognition and biomedical question answering. ClinicalBERT is pretrained on clinical notes and EHR, capturing nuances specific to medical contexts[76]. It supports varied tasks such as clinical entity recognition, medical coding, and patient phenotype extraction. All these embedding models capture entire sentences, paragraphs, or even longer texts. This distinction capability allows any LLMs to understand medical terminology at a contextually rich level.

Implementation of FedLLM for ADR Prediction

The implementation of FedLLM inference for ADR prediction is usually not a topic in academic research. Its importance should never be underestimated. It is the last mile of a FedLLM modeling effort to connect to the end users. It requires a GPU-supported backend to meet the significant

computational demands of FedLLM processing. GPUs provide the parallel processing capabilities necessary for handling complex model architectures, enabling faster inference and efficient handling of large-scale datasets. This is particularly critical in both federated learning process and afterward model inference, where distributed data processing and model inference must be performed seamlessly to ensure scalability, accuracy, and real-time responsiveness in ADR prediction tasks.

The interaction between end user and FedLLM relies on prompt engineering techniques, which help transform a user query into a FedLLM-compatible prompt by incorporating the query as a part of context. This process is facilitated through a user interface (UI) serving as the front end, typically designed as an interactive chatbot. Such a design enhances user engagement, particularly in healthcare applications, where individualized, real-time, and professional responses are essential. During inference, FedLLM can be hosted either on a third-party cloud or an internal on-premises cloud. Opting for a third-party cloud implementation solution could be more practical for the institute when addressing technical debt. In fact, integration of prompt engineering techniques with other settings has demonstrated improved effectiveness in detecting AE or ADR[30].

There are 2 ways to use FedLLM as the back end for ADR prediction. The first method involves directly submitting the prompt to a fine-tuned FedLLM without a vector database search. This approach can significantly reduce response latency. However, its limitation is clear: newly acquired ADR information cannot be immediately incorporated into the ADR prediction process after FedLLM has been fine-tuned. Additionally, the cost of fine-tuning is non-trivial. A more flexible approach involves integrating the FedLLM into the RAG pipeline. In this setup, the search engine will calculate the similarity of embedding of user input against pre-populated vector database to identify the top K relevant ADR outcomes, which are then processed into a customized prompt before being submitted to FedLLM for ADR prediction. For example, the Telehealth system for virtual consultations and remote monitoring can help reduce the workload of healthcare providers while improving patient outcomes[77].

Several open-source Python-based tools facilitate lightweight UI development. With Streamlit being a notable example, this open-source framework enable data scientists and AI/ML engineers to create and deploy an interactive web application[78]. It also allows the integration of chatbots powered by LLM through open-source framework such as Llama Index[79] and LangChain[80].

To facilitate the development and deployment of UI in pilot test (also known for user acceptance test, or UAT) and in production, options include web-based UI deployment and container-based UI deployment. However, both methods should be implemented under tight access control and physician supervision to prevent misuse of the tool. Running a lightweight pilot test is designed to evaluate the model performance through small scale user acceptance testing. To achieve this goal, both web-based and container-based deployments present practical solutions. Depending on the readiness of client facility for Kubernetes-based continuous integration and continuous delivery/continuous deployment (CI/CD) pipeline, the CI process can ensure the incremental code changes are made frequently and reliably. This setup allows Python development code to be quickly and seamlessly deployed as a part of the CD process directly from development to production. Alternatively, packaging the entire development environment into a portal image for local container-based deployment is becoming popular due to its portability and scalability, either using a daemon dependent Docker engine[81] or a daemon independent Podman engine[82].

Challenges and Limitations of FedLLM in ADR Research

Numerous limitations of FedLLM for ADR research and application have been highlighted in the selected

articles. The primary limitation is the lack of fine-grained control over the data collection process. While open-source scraping and crowdsourcing are two common methods for collecting large textual datasets, both methods face challenges related to observational bias, scalability, and the abilities to exert fine control over the data collection process. Although, the crowd sourcing platforms can be more expensive than open-source alternatives, they do provide greater control over the data collection process[83]. Therefore, the main limitation of FedLLM for ADR research and application is the need for more controlled and diverse data sources to improve the accuracy and generalizability of the models.

The inherent server-client coupling of FL systems also increases the complexity of overall architecture and compromises the integrity of client environments. Additionally, the limited reusability of FL models may lead to suboptimal performance in some specific scenarios. Most FL systems are non-public, meaning that the data used for training is not available for public scrutiny. The lack of transparency restricts the reproducibility of results and limits the applicability of the model in certain situations. In addition, the performance of global model can be sensitive to the number of clients participating in the federated averaging process, and a sudden drop in performance in some scenarios may be attributed to smaller data partitions and inconsistencies among the participating models[84].

Security of FedLLM

FedLLM algorithms, combined with security and privacy techniques, enable federated training. However, post-training processes of FedLLM exhibit varied vulnerabilities, including but not limited to adversarial attacks, model inversion, backdoor, data extraction. Inversion attacks aim to reverse engineer and reconstruct the sensitive input data from model output, such as Personally Identifiable Information (PII)[85]. Gradient inversion attack poses an emerging threat by attempting to recreate original data from model gradient, highlighting potential security and privacy breaches in federated learning. It is important to note that some models may memorize some specific data points rather than general patterns due to the possible over-fitting on a small training data[86]. In contrast, adversarial attacks seek to deceive models by injecting carefully curated adversarial samples designed to induce incorrect predictions or unexpected behaviors [87]. This poses substantial risks to critical applications such as ADR prediction which has critical consequences such as injuries and fatalities. Mitigation strategies to defend against adversarial attacks require joint approaches, including: 1) adversarial training with adversarial samples to enhance the model's resilience against adversarial perturbations. 2) input validation before training can prevent potential adversarial modification by identifying and rejecting suspicious data from inputs. 3) gradient masking which limits the information available to adversarial attacks, making it more challenging for attackers to deceive the model.

Various strategies have been applied to enhance model security against potential threats including corpus curation, application of penalty on the training loss, instruction-based fine-tuning, classification of sensitivity data sample, and constrained direct preference optimization. Among these efforts, instruction fine-tuning with representative examples proved to be a promising solution[88].

Interpretability of FedLLM

Given the heterogeneity nature of client data and complexity of client models due to diverse client infrastructure, the explanation and interpretability of FedLLM remain complex topics. However, various techniques and tools can enhance the interpretability of FedLLM. Transformer architecture, the foundation of the most modern LLMs, employs an attention mechanism to weigh the importance of different parts of the input sequence when generating outputs, a process known as self-attention. Beyond self-attention, multi-head attention enhances the model's ability to capture diverse contextual

information by simultaneously attending to different parts of the input sequence. It achieves this by performing multiple parallel self-attention operations, each with its own set of learned query, key, and value transformations[89].

Understanding how attention mechanisms operate and interpret their output can be challenging. Techniques like attention attribution help identify the contributions of individual tokens to model predictions, thereby improving explainability. Methods for visualizing attention weights and interpreting their significance have been developed to enhance the interpretability of multi-head self-attention in transformer-based language representation models[90]

In contrast to the fixed relationship established by word embedding regardless of its context within a sentence or document, attention mechanism solves this challenge by enabling models to selectively focus on relevant parts of input sequences, thereby incorporating context sensitivity into the representation learning process in tasks requiring a nuanced understanding of language especially when words carry different contextual meanings[91]. Therefore, a fine-tuned LLM with ADR domain knowledge can be the answer to the explainable AI when trained on sufficient public and private ADR data.

With RAG technologies, the prediction of ADR outcome becomes a joint effort of FedLLM and RAG, initiated by prompt query on the ADR vector database. With no exception, all ADR data used to build the vector database should be free of PII which can be achieved through NER and information masking techniques. The top K ($K \geq 1$) relevant ADR cases returned from similar search will be used as context and incorporated into prompt for ADR prediction by FedLLM. In this model, the fine-tuned FedLLM does not work alone to predict the ADR outcome, rather, it serves primarily as an information summarization tool for ADR prediction. The retrieved ADR cases are not only used for ADR prediction but also presented to the user as source citations, making the decision-making process of FedLLM more transparent and interpretable.

Computation Costs and Environmental Impact

Unlike the pre-training LLM that requires orders of magnitude more data and computation power, fine-tuning an ADR LLM uses much less data and computational power to enhance performance on a specific downstream task. However, due to data privacy concern, fine-tuning LLM still requires an on-site training process with access to high-quality and low bias data, as well as technical readiness, construction of GPU infrastructure, and energy consumption. With the emergence of AI regulations, such as the EU AI Act[92], requiring simple data lineage, minimizing low data bias, and enhancing energy efficiency have become a priority for everyone offering AI services. This can be a major concern for organizations that are looking to reduce their environmental impact and improve the compliance and sustainability of their operations.

Running inference on LLM for ADR prediction also requires an environment with GPU or TPU to minimize response latency. It is not a cost-efficient way to set up a private inference infrastructure due to its high maintenance cost and low utilization of GPU, which typically operate at only 20%-25% of GPU memory is in use at any given time on average[93]. Currently, cloud computing services for enterprise LLM inference are offered by providers such as Google Vertex AI, Amazon AWS Bedrock, Nvidia AI Inference platform, and Snowflake Cortex. All these services operate as Infrastructure as a Service (IaaS), providing centralized computation, data storage, and network resources on a pay-as-you-go basis. Allowing customers to pay for the service they use by token counts.

Bias and Fairness

The performance of Federated Learning algorithms with LLM has been evaluated across various ADR related tasks, including Named Entity Recognition and Relation Extraction[94], [95]. However, both open-source and closed source foundation LLMs are typically pre-trained on public data with no specific focus on ADR. Since the weights of closed source LLMs are not publicly accessible, only open-source LLMs can be used as the foundation model. Most open-source foundation models have been trained to reduce bias before being released for public use. It is important to note that FedLLM may be vulnerable to undesired biases present in the fine-tuning data distributed across clients, which can lead to unfair outcomes particularly given the fact the training data should be non-IID (independent and identically distributed). These biases may come from demographic biases, data collection biases, and data sampling biases. To mitigate these issues, researchers should take extra steps in data preprocessing to remove biases and ensure that it is representative of the target population. Regularization techniques, such as dropout or weight decay, can help reduce the impact of biased features on the model's predictions.

The fairness of FedLLM involves ensuring the model responses do not disadvantage any group and it does not generate harmful or discriminatory outputs to end user. To achieve this goal, researchers can design and develop optimized algorithms that incorporate fairness metrics, such as demographic parity or equalized odds regulation, along with fairness-aware gradient descent, to promote fairness and reduce bias[24], [96]. Additionally, inclusion of adversarial examples that simulate the biases in the fine-tuning data can help the model learn to recognize and correct these biases[97].

Conclusion and Future Work

Unlike centralized model training process that brings the client data to code in server, FL systems adopt a distinct approach by bringing the server code to client data. This strategy minimizes data movement, thereby preserving the data privacy. FL has been widely used for classification tasks on structured healthcare outcomes. However, the complexity of the overall architecture and concerns about the integrity of client data remain challenges, as most FL use cases are showcased on non-public data lack reproducibility. Therefore, leveraging the FedLLM for synthetic healthcare data generation will provide a new source to meet a growing demand for unbiased training data[98].

With the prevalence of LLM, which can seamlessly process unstructured data without the need for additional data pre-processing. This approach will make it easy to leverage distributed private data in heterogeneous formats that was previously difficult to utilize. With fine-grained control over the data collection and preparation process of high-quality training data, more private and public healthcare data will be accessible for FedLLM fine-tuning tasks.

Furthermore, healthcare data extend beyond textual information. Medical image data such as computed tomography (CT) and magnetic resonance imaging (MRI) have significantly increased the volume and complexity of patient information available for multimodal LLM fine-tuning[99]. A key limitation of FedLLM for ADR research and application is the need for more controlled and diverse data sources to improve the prediction accuracy and generalizability. The latest multimodal open-source LLMs with vision abilities such as Llama 3.2 vision from Meta and MVLm from Nvidia are poised to address this demand. Meanwhile, as computational power becomes more affordable and the size of foundational LLMs decreases without sacrificing performance, it is anticipated that FedLLM will gain significant traction in healthcare applications, particularly in ADR research.

Author Contributions

Conceptualization: KRC (lead), DG (equal)

Data curation: DG

Formal analysis: DG (lead), KRC (supporting)

Investigation: DG
Methodology: DG
Project administration: KRC (lead), DG (equal)
Resources: DG
Supervision: KRC
Validation: KRC
Visualization: DG (lead), KRC (supporting)
Writing – original draft: DG (lead), KRC (supporting)
Writing – review & editing: DG (lead), KRC (supporting)

Conflicts of Interest

None declared.

Reference

- [1] Institute of Medicine (US) Committee on Quality of Health Care in America, *To Err is Human: Building a Safer Health System*. Washington (DC): National Academies Press (US), 2000. Accessed: Oct. 31, 2024. [Online]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK225182/>
- [2] Institute of Medicine (US) Committee on Quality of Health Care in America, *To Err is Human: Building a Safer Health System*. Washington (DC): National Academies Press (US), 2000. Accessed: Jul. 15, 2024. [Online]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK225182/>
- [3] “WISQARS,” Centers for Disease Control and Prevention. Accessed: Jul. 15, 2024. [Online]. Available: <https://wisqars.cdc.gov/about/nonfatal-injury-data>
- [4] “FDA Adverse Event Reporting System (FAERS): Latest Quarterly Data Files | FDA.” Accessed: Jun. 09, 2024. [Online]. Available: <https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-latest-quarterly-data-files>
- [5] S. Li *et al.*, “Federated and distributed learning applications for electronic health records and structured medical data: A scoping review,” Apr. 2023, doi: 10.48550/arXiv.2304.07310.
- [6] K. R. Madhavi, V. K. Suri, V. Mahalakshmi, R. O. Reddy, and C. S. kumar Reddy, “Federated Learning and Adaptive Privacy Preserving in Healthcare,” in *Innovations in Bio-Inspired Computing and Applications*, A. Abraham, A. Bajaj, N. Gandhi, A. M. Madureira, and C. Kahraman, Eds., in Lecture Notes in Networks and Systems. Cham: Springer Nature Switzerland, 2023, pp. 543–551. doi: 10.1007/978-3-031-27499-2_51.
- [7] M. Liu, S. Ho, M. Wang, L. Gao, Y. Jin, and H. Zhang, “Federated Learning Meets Natural Language Processing: A Survey,” Jul. 2021, doi: 10.48550/arXiv.2107.12603.
- [8] O. Choudhury *et al.*, “Anonymizing Data for Privacy-Preserving Federated Learning,” Feb. 2020, doi: 10.48550/arXiv.2002.09096.
- [9] L. Yi, J. Zhang, R. Zhang, J. Shi, G. Wang, and X. Liu, “SU-Net: An Efficient Encoder-Decoder Model of Federated Learning for Brain Tumor Segmentation,” in *Artificial Neural Networks and Machine Learning – ICANN 2020*, I. Farkas, P. Masulli, and S. Wermter, Eds., Cham: Springer International Publishing, 2020, pp. 761–773. doi: 10.1007/978-3-030-61609-0_60.
- [10] M. F. Almufareh, N. Tariq, M. Humayun, and B. Almas, “A Federated Learning Approach to Breast Cancer Prediction in a Collaborative Learning Framework,” *Healthcare*, vol. 11, no. 24, p. 3185, Dec. 2023, doi: 10.3390/healthcare11243185.
- [11] Z. Zhang *et al.*, “Improving Semi-supervised Federated Learning by Reducing the Gradient Diversity of Models,” in *2021 IEEE International Conference on Big Data (Big Data)*, Dec. 2021, pp. 1214–1225. doi: 10.1109/BigData52589.2021.9671693.
- [12] S. Hussain, H. Afzal, R. Saeed, N. Iltaf, and M. Y. Umair, “Pharmacovigilance with Transformers: A Framework to Detect Adverse Drug Reactions Using BERT Fine-Tuned with FARM,” *Comput. Math. Methods Med.*, vol. 2021, p. 5589829, Aug. 2021, doi: 10.1155/2021/5589829.
- [13] T. Che *et al.*, “Federated Learning of Large Language Models with Parameter-Efficient Prompt Tuning and Adaptive Optimization,” Oct. 2023, doi: 10.48550/arXiv.2310.15080.
- [14] Z. Zhang *et al.*, “Improving Semi-supervised Federated Learning by Reducing the Gradient Diversity of Models,” in *2021 IEEE International Conference on Big Data (Big Data)*, Dec. 2021, pp. 1214–1225. doi: 10.1109/BigData52589.2021.9671693.
- [15] J. Zhang, S. Guo, X. Ma, H. Wang, W. Xu, and F. Wu, “Parameterized Knowledge Transfer for Personalized Federated Learning,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 10092–10104. Accessed: Apr. 03, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/hash/5383c7318a3158b9bc261d0b6996f7c2-Abstract.html
- [16] Y. Tian, Y. Wan, L. Lyu, D. Yao, H. Jin, and L. Sun, “FedBERT: When Federated Learning

- Meets Pre-training,” *ACM Trans Intell Syst Technol*, vol. 13, no. 4, p. 66:1-66:26, Aug. 2022, doi: 10.1145/3510033.
- [17] R. Ye *et al.*, “FedLLM-Bench: Realistic Benchmarks for Federated Learning of Large Language Models,” *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 111106–111130, Dec. 2024.
- [18] R. Ye *et al.*, “OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning,” in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, in KDD ’24. New York, NY, USA: Association for Computing Machinery, Aug. 2024, pp. 6137–6147. doi: 10.1145/3637528.3671582.
- [19] “PRISMA for Scoping Reviews (PRISMA-ScR),” PRISMA statement. Accessed: Jun. 09, 2024. [Online]. Available: <https://www.prisma-statement.org/scoping>
- [20] “[1909.00384] DeepHealth: Review and challenges of artificial intelligence in health informatics.” Accessed: Jun. 09, 2024. [Online]. Available: <https://arxiv.org/abs/1909.00384>
- [21] S. A. Dubey and A. A. Pandit, “A Comprehensive Review and Application of Interpretable Deep Learning Model for ADR Prediction,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, 2022, doi: 10.14569/IJACSA.2022.0130924.
- [22] L. Ntantiso, A. Bagula, O. Ajayi, and F. Kahenga-Ngongo, “A Review of Federated Learning: Algorithms, Frameworks and Applications,” in *Towards new e-Infrastructure and e-Services for Developing Countries*, R. A. Saeed, A. D. Bakari, and Y. H. Sheikh, Eds., in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Cham: Springer Nature Switzerland, 2023, pp. 341–357. doi: 10.1007/978-3-031-34896-9_20.
- [23] R. M. Murphy *et al.*, “Adverse drug event detection using natural language processing: A scoping review of supervised learning methods,” *PLOS ONE*, vol. 18, no. 1, p. e0279842, Jan. 2023, doi: 10.1371/journal.pone.0279842.
- [24] F. Wang and A. Preininger, “AI in Health: State of the Art, Challenges, and Future Directions,” *Yearb. Med. Inform.*, vol. 28, no. 1, pp. 16–26, Aug. 2019, doi: 10.1055/s-0039-1677908.
- [25] W. Ali, W. Zuo, R. Ali, X. Zuo, and G. Rahman, “Causality Mining in Natural Languages Using Machine and Deep Learning Techniques: A Survey,” *Appl. Sci.*, vol. 11, no. 21, p. 10064, Jan. 2021, doi: 10.3390/app112110064.
- [26] M. Bampa, “Handling temporality of clinical events with application to Adverse Drug Event detection in Electronic Health Records: A scoping review,” *ArXiv*, Apr. 2019, Accessed: Jun. 11, 2024. [Online]. Available: <https://www.semanticscholar.org/paper/f4cc573a929f023c5e3da06c43cad09b827f502>
- [27] X. Chen, G. Cheng, F. L. Wang, X. Tao, H. Xie, and L. Xu, “Machine and cognitive intelligence for human health: systematic review,” *Brain Inform.*, vol. 9, no. 1, p. 5, Feb. 2022, doi: 10.1186/s40708-022-00153-9.
- [28] J. Fu *et al.*, “Differentially Private Federated Learning: A Systematic Review,” May 19, 2024, *arXiv*: arXiv:2405.08299. doi: 10.48550/arXiv.2405.08299.
- [29] C. Preiksaitis *et al.*, “The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review,” *JMIR Med. Inform.*, vol. 12, no. 1, p. e53787, May 2024, doi: 10.2196/53787.
- [30] D. Li *et al.*, “Assessing the performance of large language models in literature screening for pharmacovigilance: a comparative study,” *Front. Drug Saf. Regul.*, vol. 4, Jun. 2024, doi: 10.3389/fdsfr.2024.1379260.
- [31] M. Salas *et al.*, “The Use of Artificial Intelligence in Pharmacovigilance: A Systematic Review of the Literature,” *Pharm. Med.*, vol. 36, no. 5, pp. 295–306, Oct. 2022, doi: 10.1007/s40290-022-00441-z.
- [32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 24, 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.
- [33] C. McMaster *et al.*, “Developing a deep learning natural language processing algorithm for

- automated reporting of adverse drug reactions,” *J. Biomed. Inform.*, vol. 137, p. 104265, Jan. 2023, doi: 10.1016/j.jbi.2022.104265.
- [34] E. Hassan, T. Abd El-Hafeez, and M. Y. Shams, “Optimizing classification of diseases through language model analysis of symptoms,” *Sci. Rep.*, vol. 14, no. 1, p. 1507, Jan. 2024, doi: 10.1038/s41598-024-51615-5.
- [35] OpenAI *et al.*, “GPT-4 Technical Report,” Mar. 04, 2024, *arXiv*: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.
- [36] J. R. Nebeker, P. Barach, and M. H. Samore, “Clarifying adverse drug events: a clinician’s guide to terminology, documentation, and reporting,” *Ann. Intern. Med.*, vol. 140, no. 10, pp. 795–801, May 2004, doi: 10.7326/0003-4819-140-10-200405180-00009.
- [37] “MSSO Updates | MedDRA.” Accessed: Jun. 09, 2024. [Online]. Available: <https://www.meddra.org/>
- [38] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang, “Cadec: A corpus of adverse drug event annotations,” *J. Biomed. Inform.*, vol. 55, pp. 73–81, Jun. 2015, doi: 10.1016/j.jbi.2015.03.010.
- [39] “European Medicines Agency (EMA).” Accessed: Nov. 01, 2024. [Online]. Available: <https://www.ema.europa.eu/en/homepage>
- [40] C. for D. E. and Research, “FDA Adverse Event Reporting System (FAERS) Public Dashboard,” *FDA*, Dec. 2023, Accessed: Nov. 01, 2024. [Online]. Available: <https://www.fda.gov/drugs/fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard>
- [41] H. Canada, “Adverse Reaction Database.” Accessed: Nov. 01, 2024. [Online]. Available: <https://www.canada.ca/en/health-canada/services/drugs-health-products/medeffect-canada/adverse-reaction-database.html>
- [42] T. G. Administration (TGA), “Database of Adverse Event Notifications (DAEN) - medicines | Therapeutic Goods Administration (TGA).” Accessed: Nov. 01, 2024. [Online]. Available: <https://www.tga.gov.au/safety/safety/database-adverse-event-notifications-daen-medicines>
- [43] D. Xu *et al.*, “Overview of the 9th Social Media Mining for Health Applications (#SMM4H) Shared Tasks at ACL 2024 – Large Language Models and Generalizability for Social Media NLP,” in *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, D. Xu and G. Gonzalez-Hernandez, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 183–195. Accessed: Nov. 01, 2024. [Online]. Available: <https://aclanthology.org/2024.smm4h-1.40>
- [44] “National NLP Clinical Challenges (n2c2).” Accessed: Nov. 01, 2024. [Online]. Available: <https://n2c2.dbmi.hms.harvard.edu/home>
- [45] “ade-benchmark-corpus/ade_corpus_v2 · Datasets at Hugging Face.” Accessed: Nov. 01, 2024. [Online]. Available: https://huggingface.co/datasets/ade-benchmark-corpus/ade_corpus_v2
- [46] “Common Terminology Criteria for Adverse Events (CTCAE) | Protocol Development | CTEP.” Accessed: Nov. 01, 2024. [Online]. Available: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm
- [47] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, “G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment,” May 23, 2023, *arXiv*: arXiv:2303.16634. doi: 10.48550/arXiv.2303.16634.
- [48] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a Method for Automatic Evaluation of Machine Translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds., Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. doi: 10.3115/1073083.1073135.
- [49] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. Accessed: Jun. 29, 2024. [Online]. Available:

<https://aclanthology.org/W04-1013>

- [50] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” Feb. 24, 2020, *arXiv*: arXiv:1904.09675. doi: 10.48550/arXiv.1904.09675.
- [51] A. Aynedinov and A. Akbik, “SemScore: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity,” *arXiv.org*. Accessed: Feb. 04, 2024. [Online]. Available: <https://arxiv.org/abs/2401.17072v1>
- [52] D. Peer, P. Schöpf, V. Nebendahl, A. Rietzler, and S. Stabinger, “ANLS* -- A Universal Document Processing Metric for Generative Large Language Models,” May 25, 2024, *arXiv*: arXiv:2402.03848. doi: 10.48550/arXiv.2402.03848.
- [53] O. Choudhury, Y. Park, T. Salonidis, A. Gkoulalas-Divanis, I. Sylla, and A. k. Das, “Predicting Adverse Drug Reactions on Distributed Health Data using Federated Learning,” *AMIA. Annu. Symp. Proc.*, vol. 2019, pp. 313–322, Mar. 2020.
- [54] A. Sakhovskiy and E. Tutubalina, “Multimodal model with text and drug embeddings for adverse drug reaction classification,” *J. Biomed. Inform.*, vol. 135, p. 104182, Nov. 2022, doi: 10.1016/j.jbi.2022.104182.
- [55] Q. Xie, Z. Luo, B. Wang, and S. Ananiadou, “A Survey for Biomedical Text Summarization: From Pre-trained to Large Language Models,” Jul. 13, 2023, *arXiv*: arXiv:2304.08763. doi: 10.48550/arXiv.2304.08763.
- [56] T. Ishikawa, T. Yakoh, and H. Urushihara, “An NLP-Inspired Data Augmentation Method for Adverse Event Prediction Using an Imbalanced Healthcare Dataset,” *IEEE Access*, vol. 10, pp. 81166–81176, 2022, doi: 10.1109/ACCESS.2022.3195212.
- [57] F. Zhu, Y. Zhang, L. Chen, B. Qin, and R. Xu, “Learning to Describe for Predicting Zero-shot Drug-Drug Interactions,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 14855–14870. doi: 10.18653/v1/2023.emnlp-main.918.
- [58] T. Fan *et al.*, “FATE-LLM: A Industrial Grade Federated Learning Framework for Large Language Models,” Oct. 16, 2023, *arXiv*: arXiv:2310.10049. doi: 10.48550/arXiv.2310.10049.
- [59] W. Kuang *et al.*, “FederatedScope-LLM: A Comprehensive Package for Fine-tuning Large Language Models in Federated Learning,” in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, in KDD ’24. New York, NY, USA: Association for Computing Machinery, Aug. 2024, pp. 5260–5271. doi: 10.1145/3637528.3671573.
- [60] J. Zhu *et al.*, “Promoting Data and Model Privacy in Federated Learning through Quantized LoRA,” Jun. 16, 2024, *arXiv*: arXiv:2406.10976. doi: 10.48550/arXiv.2406.10976.
- [61] E. Diao, J. Ding, and V. Tarokh, “HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients,” Dec. 13, 2021, *arXiv*: arXiv:2010.01264. doi: 10.48550/arXiv.2010.01264.
- [62] A. Bera, R. Das, S. Ghosh, R. Chakraborty, I. Mitra, and P. Nandy, “Harnessing Transformers for Detecting Adverse Drug Reaction and Customized Causality Explanation using Generative AI,” in *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, Aug. 2023, pp. 1–6. doi: 10.1109/ICCUBEA58933.2023.10392249.
- [63] Y. Gu *et al.*, “Distilling Large Language Models for Biomedical Knowledge Extraction: A Case Study on Adverse Drug Events,” Jul. 12, 2023, *arXiv*: arXiv:2307.06439. doi: 10.48550/arXiv.2307.06439.
- [64] “Hugging Face – The AI community building the future.” Accessed: Jun. 09, 2024. [Online]. Available: <https://huggingface.co/>
- [65] E. J. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models,” Oct. 16, 2021, *arXiv*: arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685.
- [66] S.-Y. Liu *et al.*, “DoRA: Weight-Decomposed Low-Rank Adaptation,” Jun. 03, 2024, *arXiv*:

- arXiv:2402.09353. doi: 10.48550/arXiv.2402.09353.
- [67] W. J. Yun, S. Kim, and J. Kim, "Multi-Site Clinical Federated Learning Using Recursive and Attentive Models and NVFlare," in *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*, Jul. 2023, pp. 1033–1036. doi: 10.1109/ICDCS57875.2023.00115.
- [68] M. Wortsman *et al.*, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," Jul. 01, 2022, *arXiv*: arXiv:2203.05482. doi: 10.48550/arXiv.2203.05482.
- [69] T. White, "Sampling Generative Networks," Dec. 06, 2016, *arXiv*: arXiv:1609.04468. doi: 10.48550/arXiv.1609.04468.
- [70] L. Yu, B. Yu, H. Yu, F. Huang, and Y. Li, "Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch," Jun. 13, 2024, *arXiv*: arXiv:2311.03099. doi: 10.48550/arXiv.2311.03099.
- [71] D.-H. Jang, S. Yun, and D. Han, "Model Stock: All we need is just a few fine-tuned models," Mar. 28, 2024, *arXiv*: arXiv:2403.19522. doi: 10.48550/arXiv.2403.19522.
- [72] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," Oct. 16, 2013, *arXiv*: arXiv:1310.4546. doi: 10.48550/arXiv.1310.4546.
- [73] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.
- [74] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs," Aug. 14, 2018, *arXiv*: arXiv:1603.09320. Accessed: Oct. 31, 2024. [Online]. Available: <http://arxiv.org/abs/1603.09320>
- [75] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Oct. 18, 2019, *arXiv*: arXiv:1901.08746. doi: 10.48550/arXiv.1901.08746.
- [76] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," Nov. 29, 2020, *arXiv*: arXiv:1904.05342. doi: 10.48550/arXiv.1904.05342.
- [77] H. Y. Kwan, "User-Focused Telehealth Powered by LLMs: Bridging the Gap Between Technology and Human-Centric Care Delivery," in *2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI)*, May 2024, pp. 187–191. doi: 10.1109/CCAI61966.2024.10603150.
- [78] "Streamlit • A faster way to build and share data apps." Accessed: Jan. 05, 2025. [Online]. Available: <https://streamlit.io/>
- [79] "LlamaIndex - Build Knowledge Assistants over your Enterprise Data." Accessed: Jan. 05, 2025. [Online]. Available: <https://www.llamaindex.ai/>
- [80] "LangChain." Accessed: Jan. 05, 2025. [Online]. Available: <https://www.langchain.com/>
- [81] "Docker: Accelerated Container Application Development." Accessed: Jan. 05, 2025. [Online]. Available: <https://www.docker.com/>
- [82] "Podman." Accessed: Jan. 05, 2025. [Online]. Available: <https://podman.io/>
- [83] C. Chen *et al.*, "Interactive crowdsourcing to spontaneous reporting of Adverse Drug Reactions," in *2014 IEEE International Conference on Communications (ICC)*, Jun. 2014, pp. 4275–4280. doi: 10.1109/ICC.2014.6883992.
- [84] "A survey on federated learning," *Knowl.-Based Syst.*, vol. 216, p. 106775, Mar. 2021, doi: 10.1016/j.knosys.2021.106775.
- [85] S. Almutairi and A. Barnawi, "Federated learning vulnerabilities, threats and defenses: A systematic review and future directions," *Internet Things*, vol. 24, p. 100947, Dec. 2023, doi: 10.1016/j.iot.2023.100947.

- [86] S. Zhou, T. Zhu, D. Ye, X. Yu, and W. Zhou, "Boosting Model Inversion Attacks with Adversarial Examples," Jun. 24, 2023, *arXiv*: arXiv:2306.13965. doi: 10.48550/arXiv.2306.13965.
- [87] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019, doi: 10.1126/science.aaw4399.
- [88] "OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning." Accessed: Oct. 21, 2024. [Online]. Available: <https://arxiv.org/html/2402.06954v1>
- [89] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "Multi-Head Attention: Collaborate Instead of Concatenate," May 20, 2021, *arXiv*: arXiv:2006.16362. doi: 10.48550/arXiv.2006.16362.
- [90] J. Vig, "Visualizing Attention in Transformer-Based Language Representation Models," Apr. 11, 2019, *arXiv*: arXiv:1904.02679. doi: 10.48550/arXiv.1904.02679.
- [91] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 01, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.
- [92] "The AI Act Explorer | EU Artificial Intelligence Act." Accessed: Apr. 03, 2025. [Online]. Available: <https://artificialintelligenceact.eu/ai-act-explorer/>
- [93] S. Samsi *et al.*, "From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference," in *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2023, pp. 1–9. doi: 10.1109/HPEC58863.2023.10363447.
- [94] S. Ge, F. Wu, C. Wu, T. Qi, Y. Huang, and X. Xie, "FedNER: Privacy-preserving Medical Named Entity Recognition with Federated Learning," Mar. 25, 2020, *arXiv*: arXiv:2003.09288. doi: 10.48550/arXiv.2003.09288.
- [95] J. Wei, T. Hu, J. Dai, Z. Wang, P. Han, and W. Huang, "Research on named entity recognition of adverse drug reactions based on NLP and deep learning," *Front. Pharmacol.*, vol. 14, p. 1121796, Jun. 2023, doi: 10.3389/fphar.2023.1121796.
- [96] G. H. Kwak and P. Hui, "DeepHealth: Review and challenges of artificial intelligence in health informatics," Aug. 08, 2020, *arXiv*: arXiv:1909.00384. doi: 10.48550/arXiv.1909.00384.
- [97] Y. Zhou, J. Wu, H. Wang, and J. He, "Adversarial Robustness through Bias Variance Decomposition: A New Perspective for Federated Learning," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, in CIKM '22. New York, NY, USA: Association for Computing Machinery, Oct. 2022, pp. 2753–2762. doi: 10.1145/3511808.3557232.
- [98] R. Tang, X. Han, X. Jiang, and X. Hu, "Does Synthetic Data Generation of LLMs Help Clinical Text Mining?," Apr. 10, 2023, *arXiv*: arXiv:2303.04360. doi: 10.48550/arXiv.2303.04360.
- [99] L. R. Soenksen *et al.*, "Integrated multimodal artificial intelligence framework for healthcare applications," *Npj Digit. Med.*, vol. 5, no. 1, pp. 1–10, Sep. 2022, doi: 10.1038/s41746-022-00689-4.

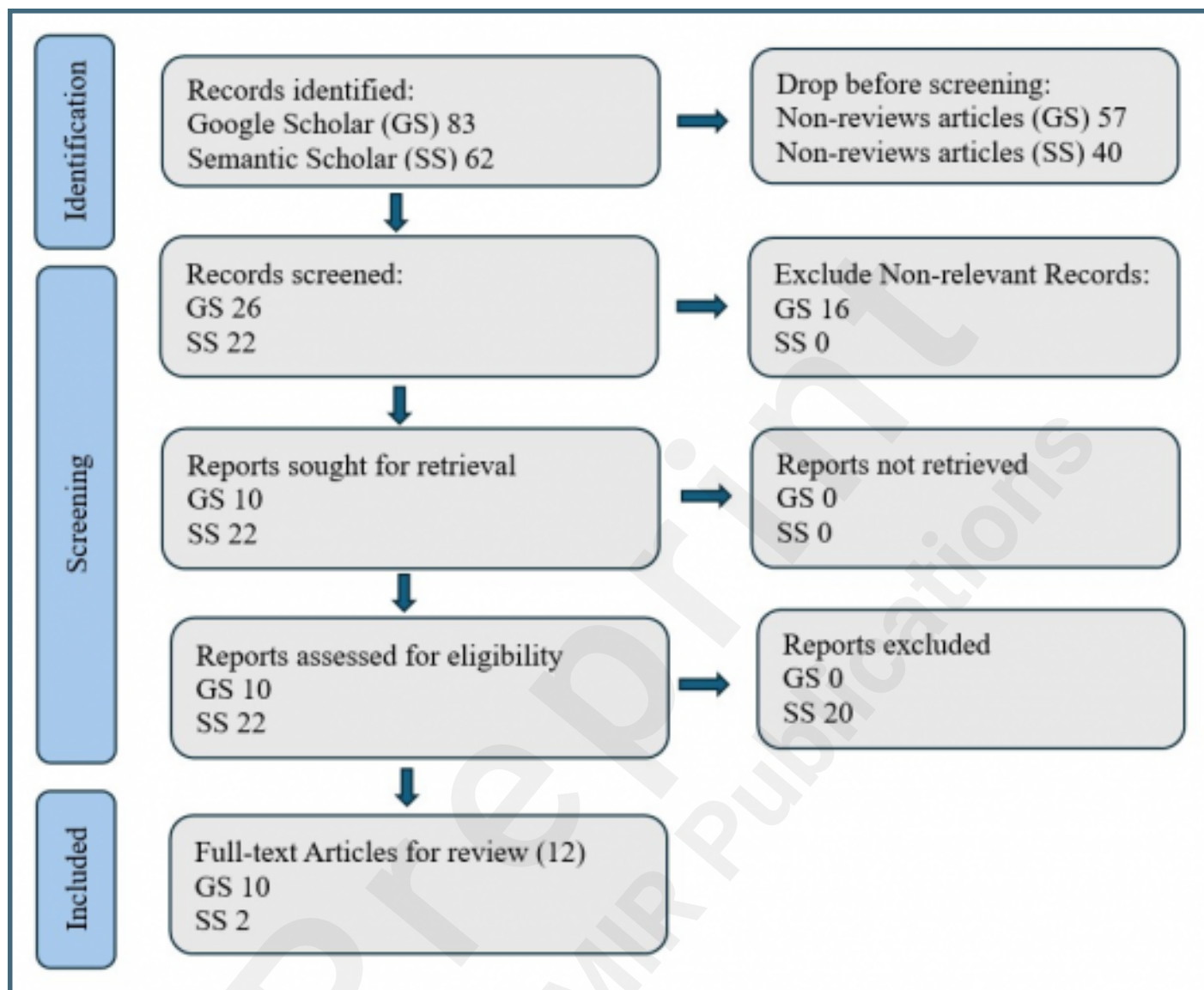
Supplementary Files

Untitled.

URL: <http://asset.jmir.pub/assets/0406cb2be389fdac1919cd1a0cb3ebef.docx>

Figures

Literature Search Strategy for FedLLM in ADR Prediction.



Multimedia Appendixes

PRISMA-ScR checklist.

URL: <http://asset.jmir.pub/assets/1cff7f6bd4bd93c7d8279f80c5c20550.docx>

