Leveraging temporality, dose effect, and co-medication to improve drug safety surveillance

YiFan Wu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Trevor A. Cohen, Chair

Ian H. De Boer

Meliha Yetişgen

Program Authorized to Offer Degree:
Biomedical Informatics and Medical Education

University of Washington

Abstract

Leveraging temporality, dose effect, and co-medication to improve drug safety surveillance

YiFan Wu

Chair of the Supervisory Committee:
Trevor A. Cohen
Department of Biomedical Informatics and Medical Education

Adverse drug reactions (ADRs) rank among the top causes of morbidity and mortality worldwide, yet current post-market drug surveillance systems often relying on spontaneous reporting. They suffer from under-reporting of ADRs and limited capture of clinical context. This dissertation addresses these gaps by leveraging electronic health record (EHR) data and transformer-based models to detect ADRs and drug–drug interactions (DDIs) more effectively.

First, we develop and evaluate a generative transformer architecture (GPT-2) trained from scratch on longitudinal EHR data from two distinct repositories (MIMIC-IV and a large university health system, UW). Unlike traditional disproportionality metrics that focus on cross-sectional drug-event co-occurrences, the proposed model captures temporal relationships and contextual dependencies among medications, diagnoses, and outcomes. Second, we introduce a "value-aware" embedding approach to incorporate continuous numeric data, such as drug dosages and lab measurements. Experimental results show that these value-aware embeddings further improve model performance, outperforming baseline transformer architectures that did not have numeric data. Third, we extend the model's scope to evaluate DDIs under polypharmacy conditions, demonstrating that a transformer exceeded the predictive accuracy of simpler machine learning baselines.

Across all EHR datasets tested, the transformer-based methods consistently surpass existing standard approaches in ADR detection, measured by improved area under the receiver operating characteristic curve (AUROC). By capturing time-dependent patterns, integrating numeric variables, and accounting for interacting drug exposures, this work broadens the capabilities of

pharmacovigilance beyond conventional "signal detection" practices. Despite practical limitations such as limited generalizability to other health systems and the need for rigorous validation—these findings demonstrated the promise of generative transformers as a scalable, data-driven framework for enhancing patient safety for pharmacovigilance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

SRS: Spontaneous reporting systems. each particular usage.

EHR: Electronic health records.

ROR: Reporting Odds Ratio

PRR: Proportional Reporting Ratio

LR: Logistic Regression

FDA: Food and Drug Administration

FAERS: FDA Adverse Event Reporting System

SVM: Support Vector Machine

DDI: Drug Drug Interaction

GPT: Generative Pre-trained Transformer

ADR: Adverse Drug Reaction

AUROC: Area Under the receiver operating curve

RWE: Real-world evidence

EMA: European Medicines Agency

SNOMED-CT: Systematized Nomenclature of Medicine - Clinical Terms

RXNORM: RxNorm is US-specific controlled terminology in medicine that contains all medications

UMLS: Unified Medical Language System

NLM: National Library of Medicine

IMDS: Institute of Medical Data Science

RWS: Real World Evidence

ER: Emergency Room

AKI: Acute Kidney Injury

NLP: Natural Language Processing

LSTM: Long Short-Term Memory

CRF: Conditional random fields

RNN: Recurrent Neural Network

BERT: Bidirectional Encoder Representations from Transformers

BHERT: BERT for EHR

WHO: World Health Organization

HVAT: Hybrid Value Aware Transformer

VA: Veterans Affairs

ADRD: Alzheimer's disease and related dementia

MIMIC: Medical Information Mart for Intensive Care

ROPE: Rotary positional embeddings

BCPNN: Bayesian Confidence Propagation Neural Network

GPS: Gamma Poisson Shrinkage

ML: Machine learning

CNN: Conventional neural network

DNN: Deep neural network

GNN: Graph neural network

PCA: Principle component analysis

NMF: Non-negative matrix factorization

ITHS: Institute of Translational Health Sciences

EUADR: Exploring and Understanding Adverse Drug Reactions

OMOP: Observational Medical Outcomes Partnership

EARP: Embeddings Augmented by Random Permutation

WTE: Word token embeddings

PE: Positional embeddings

AZCERT: Arizona Center for Education and Research on Therapeutics

UMN: An ADR-drug reference set curated by Minnesota University

AKI: Acute kidney injury

GH: Acute gastrointestinal hemorrhage

CUI: Concept unique identifier

HVAT: Hybrid value-aware transformer

MLM: Masked language models

VAE: Value-aware embeddings

VSA: Vector symbolic architectures

HRR: Holographic reduced representation

VTB: Vector derived binding

MBAT: Matrix-based binding

RHRR: Fourier holographic reduced representation

MSE: Mean squared error

DDE: Drug drug event

CRESCENDDI: Clinical-relevant Reference set Centered around Drug Drug Interactions

ACKNOWLEDGMENTS

I would first like to express my deepest gratitude to my advisor, Dr. Trevor Cohen. Throughout my PhD journey, I felt continuously supported and encouraged by his guidance and the time he invested in my growth. I am also truly grateful to my reading committee members, Ian and Meliha, as well as Dr. Donna Berry for serving as the GSR.

I extend my appreciation to the CIPCT program, in particular Dr. Donna Berry and Dr. Andrea Hartzler, for their invaluable support. My thanks also go to Dr. Pat Reid Ponte and Dr. Tom Payne for providing me with opportunities to learn and teach, moments of which I will always cherish.

I am especially thankful to the faculty at Institute for Health metrics and Evaluation, including Dr. Theo Vos, Dr. Sarah Hanson, and Dr. Damian Santomauro, for their mentorship and support. My time at IHME was both eye-opening and rewarding, and I am grateful for the chance to work with WHO officers during that period.

I would like to recognize those who mentored and supported me early in my academic career, particularly Dr. Cijiang He, Dr. Qi Qian, Dr. Markus Bitzer, Dr. Wenjun Ju and Dr. Adam Wilcox. Their guidance was indispensable in shaping my path forward (undergraduate and early graduate).

My gratitude extends to the BIME department for their continual support. Special thanks to Peter TH, John G, and the staff members—Heidi, Heather, Moiya, Marni, who provided assistance and encouragement throughout this process.

Above all, I owe everything to my parents, Yi Wu and Nan Chen. They are my role models, both in their moral example and their dedication to scholarly excellence. Their support and belief in my "crazy ideas" have shaped me into who I am today. As my mother always says: "Petite sa petite, l'oiseau fait son nid" (step by step, the bird builds its nest). This reminds me that even the smallest progress is meaningful. I also want to thank my friends, especially George Xu and Qifei

Dong, as well as my accountability partners from around the world, especially Xiantian Lin, Ayuko K., and Nigu L. etc for spending countless hours studying and working alongside me.

During the COVID-19 pandemic, I experienced the loss of a dear friend at just 31 years old due to colorectal cancer. This tragedy prompted me to reflect on mental and physical health, and it reinforced the importance of self-care and gratitude. Despite having moments of doubt, I am grateful I persevered. I strive to leave behind any notions of defining myself by limitations. The past two years, especially after COVID, have led me to appreciate a healthier mindset and overall well-being.

Finally, I acknowledge the support throughout my PhD: the National Library of Medicine (NLM) predoctoral fellowship, the CIPCT program, Dr. Trevor Cohen's startup and teaching funds, IHME RAship, and the Institute of Medical Data Science(IMDS) Pilot Award.

Thank you to everyone who has been part of this journey. I am forever grateful for your encouragement, mentorship, friendship, and love.

DEDICATION

to science

Chapter 1

INTRODUCTION AND OVERVIEW

1.1 Introduction

Adverse drug reactions (ADR) constitute a significant burden to healthcare systems globally and rank among the top causes of mortality. Financial costs are estimated at $30.1 billion (US) and $79 billion (EU), respectively [67, 105, 9]. Unlike phase I–III clinical trials, which are tightly controlled, post-market drug surveillance relies heavily on observational data [119]. The field of pharmacovigilance concerns the early detection, reporting, and evaluation of ADRs, often through spontaneous reporting systems (SRSs) such as the FDA Adverse Event Reporting System (FAERS). However, SRSs heavily rely on provider reporting and suffer from an under-reporting rate of over 90% [87].

Electronic health record (EHR) data presents an alternative opportunity to enhance ADR signal detection because they capture a longitudinal and provide a more comprehensive view of a patient's healthcare journey. EHRs include diagnoses, prescriptions, dosages, laboratory values, and clinical notes. Yet, integrating EHR data into pharmacovigilance methods introduces challenges such as defining consistent outcome phenotypes, modeling time-dependencies, and dealing with missing information at scale. Recent advances in deep learning—especially transformer-based architectures—offer a potential solution to these challenges[125], because transformers can naturally handle sequence data and learn relationships among medications, doses, and clinical outcomes.

1.2 Hypotheses

I hypothesize that:

- **(H1)** Transformer based models will outperform conventional disproportionality metrics to improve ADR signal detection.

- **(H2)** Representing numeric dose and lab data in transformer models will further enhance

their predictive utility on this task

- **(H3)** Modeling co-medications (polypharmacy) will allow us to capture clinically significant drug-drug interactions, improving overall risk predictions.

1.3 Motivations and Aims

Pharmacovigilance methods frequently rely on cross-sectional snapshots of drug-event co-occurrence, limiting their ability to exploit the longitudinal structure of EHR data. The application of generative pretrained transformer in pharmacovigilance remains unexplored. Lastly, polypharmacy introduces drug-drug interactions (DDIs) that may exacerbate or mitigate adverse events.

To address these limitations, this dissertation proposes a novel framework that leverages generative transformer models (e.g., GPT-2) and value-aware embeddings for numeric dose data. Specifically, I propose three Specific Aims:

- **Aim 1:** Develop and evaluate generative transformer models for ADR signal detection in EHR datasets.

- **Aim 2:** Integrate value-aware embeddings to represent numeric clinical data (e.g., lab values and dosages) and improve ADR detection accuracy.

- **Aim 3:** Leverage the resulting transformer-based models to investigate drug-drug interactions (DDIs) in observational EHR data.

1.4 Relevance to Biomedicine and Health

Addressing these aims has direct implications for patient safety and public health. The underreporting and incomplete capture of ADRs hinder regulatory decision-making and contribute to preventable harms. EHR-based pharmacovigilance techniques that utilize machine learning, particularly transformers, can identify signals earlier, potentially reducing morbidity, mortality, and healthcare costs. These innovations also align with ongoing FDA initiatives such as the Sentinel Initiative, which emphasizes use of real-world evidence from longitudinal datasets [15].

1.5 Overview of Dissertation

The dissertation is organized as follows:

- **Chapter 2** presents a selective literature review of pharmacovigilance methods, highlighting key gaps in leveraging EHR data for the ADR detection and polypharmacy analysis.

- **Chapter 3** corresponds to Aim 1, detailing the development of a GPT-2-based model to incorporate temporal sequences from the EHR. The model is evaluated against 4 ADR-drug reference sets by AUROC.

- **Chapter 4** addresses Aim 2, introducing novel strategies for representing numeric data (dose, lab values) within a transformer architecture. The model is evaluated against 4 ADR-drug reference sets by AUROC.

- **Chapter 5** extends the work from Chapter 4, exploring more complex embedding strategies and whether such approaches boost ADR detection further.

- **Chapter 6** focuses on Aim 3, applying the best-performing models to identify harmful drug-drug interactions in EHR data. The performance is evaluated against a drug-drug interaction reference set by AUROC.

- **Chapter 7** synthesizes findings, discusses limitations, and suggests directions for future work.

1.6 Roadmap

In the chapters ahead, I begin with a thorough background (Chapter 2) to understand the challenges of ADR detection in observational data and to review existing methodological approaches. We then introduce our proposed model architectures (Chapters 3 and 4), demonstrate the embedding approaches for numeric data (Chapters 4 and 5), and conclude with an exploration of how these methods can be applied to detect critical drug-drug interactions (Chapter 6). Finally, I present the overall conclusions and future directions in Chapter 7 (Figure 1.1).

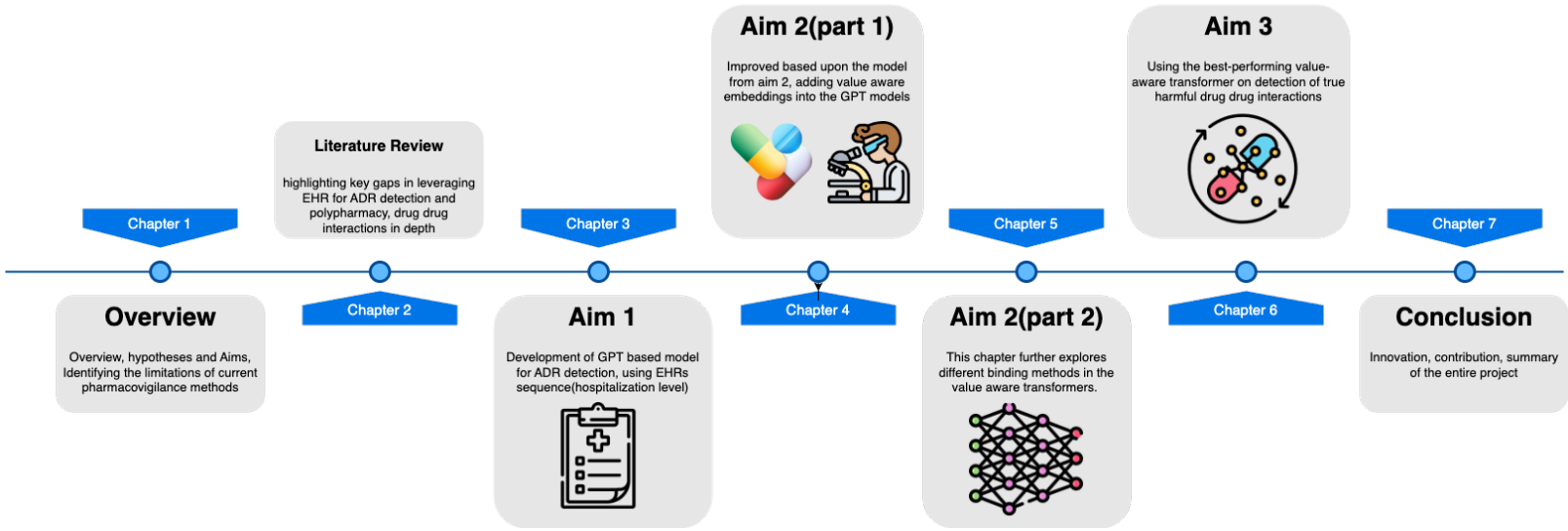

Figure 1.1: Roadmap for this dissertation.

Chapter 2

LITERATURE AND GAPS IN PRIOR WORK

2.1 Introduction to Pharmacovigilance and Need for Improved ADR Detection

Adverse drug reactions (ADRs) pose a substantial threat to public health, ranking among the top causes of mortality, with a huge financial burden of approximately $30.1 billion in the United States and $79 billion in European countries [67, 105, 9]. In contrast to the highly controlled environment of phase I–III clinical trials, post-marketing surveillance is based more on observational data [119]. The field of pharmacovigilance focuses on detecting, reporting, and evaluating these ADRs. Historically, this effort has relied on Spontaneous reporting systems (SRSs), such as the FDA Adverse Event Reporting System (FAERS) in the United States and the Vigilance system in Europe. These systems effectively capture post-market drug safety issues but depend on voluntary reporting by healthcare professionals, leading to severe under-reporting (estimated over 90% of ADRs are missed) [87]. Such under-reporting causes downstream issues in analytics, and results in inadequate prioritization of patient risks.

To overcome the limitations of SRS-based monitoring, EHR data have gained prominence as a complementary source. Unlike SRSs, EHRs document a patient's medical history—diagnoses, prescriptions, laboratory tests, and additional clinical notes. They depict a more comprehensive landscape for real-world drug usage and potential ADRs. However, traditional approaches to ADR detection (e.g., disproportionality metrics such as proportional reporting ratio and reporting odds ratio) still dominate, especially in the European countries. Although these metrics work well for SRS data, they do not fully leverage the longitudinal nature of EHRs. Particularly, the conventional disproportionality metrics result in relatively poor performance when applied to EHR data as shown by Li et al. as an example that when using EHR alone for ADR detection achieved 0.53 area under the roc curve [71, 78].

Given the surge in real-world evidence (RWE) initiatives, there is a demand for novel methods that leverage the richness of EHRs for post-marketing surveillance. Consequently, the development

of robust EHR-based pharmacovigilance strategies becomes increasingly urgent for ensuring high-quality to support drug safety monitoring.

2.2 EHR-Based Pharmacovigilance

2.2.1 EHR data and Major Challenges

EHRs contain rich longitudinal data spanning diagnoses, treatments, prescriptions, dosages, laboratory values, and more [1, 40]. Additionally, EHR systems frequently use standardized terminologies (e.g., SNOMED-CT, RxNorm, UMLS), which can facilitate the extraction of interoperable ADR signals.

Despite these advantages, several challenges persist when using EHRs for pharmacovigilance:

- Temporal Alignment: A scoping review reported that about 25% of EHR-based ADR studies didn't enforce a temporal relationship ensuring that the drug is administered before the alleged side effect, thereby undermining causal inference [24].

- Missing and Irregular Data: Patient records can vary widely in length, and timing of clinical events may be sporadic; these gaps complicate time-series modeling.

Given these complexities, it is imperative to incorporate the sequence of clinical events into post market drug surveillance. In EHR timestamps, drug administration must precede the appearance of a suspected side effect. (Figure 2.1) offers a schematic depiction of how medications, tests, and diagnoses can occur in sequence as a patient within a hospitalization moves from initial diagnosis to further treatment. The following is a practical example (Figure. 2.1):

In this hypothetical case, a patient is admitted to the ER, and an initial diagnosis is assigned (Figure 2.1). The clinicians would order tests for further examination. The orders of medical tests, medication usage and the diagnosis are depicted as a sequence of clinical actions.

2.2.2 Examples of EHR-based ADR detection attempts

Studying temporality in the EHR for ADR detection is challenging due to the data's varying length, and missingness of data across time and length. Translating longitudinal EHR data into a feature

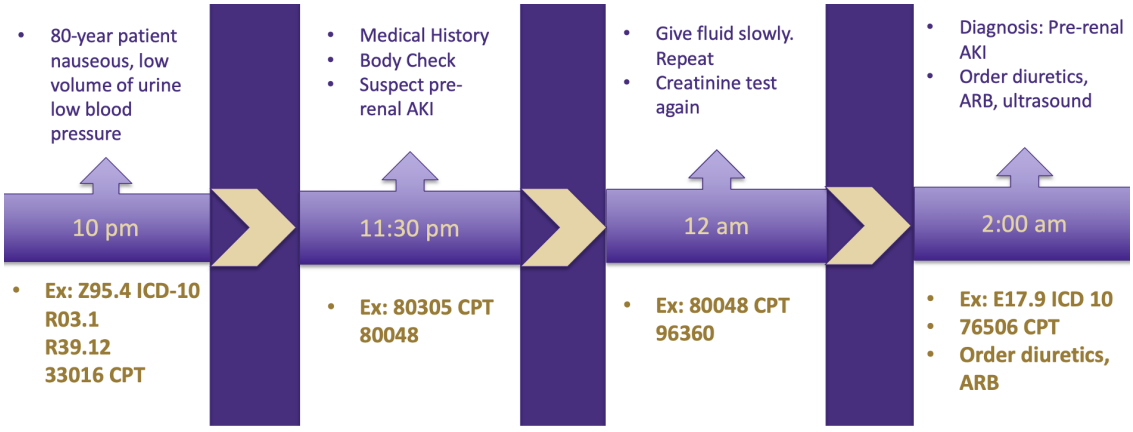

Figure 2.1: Schematic depiction of EHR data sequence.

representation for machine learning is also challenging.

There have been a few attempts to work with temporal data in ADR detection. Between 2014 and 2017, Zhao et al. published a series of papers using the temporality of clinical events for drug safety surveillance. In this work, the authors attempted to simplify time series data by transforming them into symbolic time intervals using a symbolic aggregate approximation (SAX) based method[144, 145, 146]. They transformed the time-series data by reducing dimensionality(transform the original time-series into piecewise aggregate approximate, similar to binning) and converting it into strings. The similarities of the strings are measured by extending Euclidean and piecewise aggregate approximate(PAA) distances. The method showed good performance but did not handle data with high sparsity. The string comparison method used in the studies, Levenshtein distance, is also highly dependent on the length of the string because the additional elements of longer strings are viewed as insertions, which exact a penalty that increases distance. In 2019, Bagattini et al. improved the previous techniques by alleviating the sparsity issue and measured the distance of event strings using less length-dependent methods modified based upon edit distance, which is an improvement[7]. Nevertheless, the technique Bagattini et al. proposed was a stand-alone method that has not been used for other tasks or other datasets, and the generalizability of the model need to be improved.

In 2020, Estri et al. proposed Transitive Sequential Pattern Mining (TSPM) to mine longitudinal discrete EHR data. Essentially, the method applies dimensionality reduction by minimizing sparsity (removing features with large numbers of missing values) and maximizing relevance (computing the entropy of the empirical probability distribution) (MSMR). The researchers used logistic regression with L1 regularization to predict five diseases(with no involvement of ADR): congenital heart failure, COPD, type I and II diabetes, rheumatoid arthritis, and ulcerative colitis. However, the method has some critical limitations. It uses a brute force method to find all permutations of drug-event pairs and only keeps the drug-event pair with the highest probability for a single drug (i.e. if Pr(ibuprofen|ADR1) > Pr(ibuprofen|ADR2); only the former relationship is kept in the analysis) [27, 28]. An additional limitation of the choice of data is that the researchers only used medication and diagnosis data in the analysis[27, 28].

Levine et al. developed a lagged regression model for detecting physiologic drug effects using EHR data [68]. The models were built upon an autoregressive model, known for its robustness on medical data. An autoregressive model uses a linear combination of its past values as input to predict future values of the time series. The first observation profoundly affects the following observations in the forecast window. The researchers derived two datasets with 7 drugs and 4 labs with 64 unique drug lab permutations. The procedure starts with patient normalization by using a Z transformation within each physiologic drug effect to improve the quality of time-series data[21]. The mean of the autoregressive models in this study achieved an AUROC of 0.633, and the results have high interpretability. However, one issue is that many ADRs are acute events. The developed models used 30 lags (a statistical term referring to the number of past observations in a time series that are used as input to predict the current value of the series). A first clinical event, for example, the time a drug was given, might not affect a last clinical event, if this occurs 30 days later, then the forecast window is too broad.

2.2.3 FDA Sentinel Initiative and TreeScan

In 2007, the U.S. Food and Drug Administration (FDA) launched the Sentinel Initiative to establish a national system for monitoring the safety of regulated medical products such as drugs, vaccines, biologics, and medical devices. Officially operational in 2008 and expanded in 2019, Sentinel integrates partnerships across epidemiology, pharmacy, and health informatics, with a notable emphasis on leveraging EHRs for real-world evidence[15].

Among the analytical tools employed within Sentinel, the FDA adopted TreeScan as an advancement over traditional disproportionality metrics. Originally introduced by Kulldorff and colleagues[66], TreeScan was designed to identify clusters of adverse events in spontaneous reporting systems (like FAERS)[1]. The software can simultaneously evaluate thousands of potential adverse drug reaction (ADR) signals, while adjusting for multiple comparisons. In addition, TreeScan accommodates various data sources—including EHRs and insurance claims—and can incorporate temporal information via purely temporal scan statistics or a Bernoulli model that flags higher-than-expected event rates within defined time windows.

2.2.4 Summary

Overall, EHR data represent a rich longitudinal resource that can capture nuanced safety signals overlooked by traditional spontaneous reporting systems. Existing methods—such as symbolic aggregate approaches, Transitive Sequential Pattern Mining, lagged regression, and TreeScan—offer partial solutions for analyzing this high-dimensional, irregular clinical data. However, many of these methods still treat drug-event relationships in a relatively isolated manner, under-utilizing the temporal ordering, numeric dose information, or multifaceted interactions that characterize real-world patient care.

The next section discusses, recent advances in NLP and transformer-based architectures that have opened up new possibilities. These techniques can potentially address the limitations of earlier approaches by handling sequence data more effectively, incorporating numeric representations, and modeling complex relationships among multiple drugs, diagnoses, and clinical outcomes in a unified framework.

---

[1] https://www.TreeScan.org

2.3 Natural Language Processing and Transformer-Based Approaches in Pharmacovigilance

Artificial intelligence (AI) has seen rapid growth in pharmacovigilance research over the last two decades. For instance, in early 2003, Google Scholar yielded only 140 publications when searching for "pharmacovigilance" and "machine learning," whereas by October 6, 2024, this number had increased to nearly 12,600. As of February 9, 2025, the count for "natural language processing" (NLP) and "pharmacovigilance" surpassed 5,700. This explosion of interest highlights the expanding role of AI techniques, especially NLP, in analyzing clinical text and other observational data sources for ADR detection, trend analysis, regulatory reporting, and more [87, 24].

2.3.1 Early AI/NLP in ADE detection

Most initial NLP efforts for pharmacovigilance have centered on extracting key entities (e.g., drug names, adverse events) and relations (e.g., drug-ADR pairs) from unstructured text, such as clinical notes, SRS narratives, and social media posts [86]. Models commonly used in these NLP studies include Long Short-Term Memory (LSTM) networks and Conditional Random Fields (CRFs), which proved effective for sequence labeling tasks and entity extraction [87]. Yet, when it came to signal detection—i.e., identifying potentially novel ADRs—only a handful of studies attempted to leverage these NLP outputs in a more integrated, large-scale manner[86].

2.3.2 The evolution of vector representations

Traditionally, pharmacovigilance uses disproportionality analysis based on a simple 2×2 contingency table in systems like FAERS. In this table, we count how often a specific drug and adverse event (ADR) appear together, then calculate metrics such as the Proportional Reporting Ratio (PRR). Although this count-based method is straightforward and has proven effective, it treats each drug and ADR as separate "labels" without recognizing that similar drugs may share similar side effects (Figure 2.2).

To address this limitation, our group introduced a distributed representation approach called aer2vec, inspired by the word2vec model, where drugs and ADRs are mapped into a shared vector space [99]. In this space, similar drugs or side effects naturally cluster together. As a result,

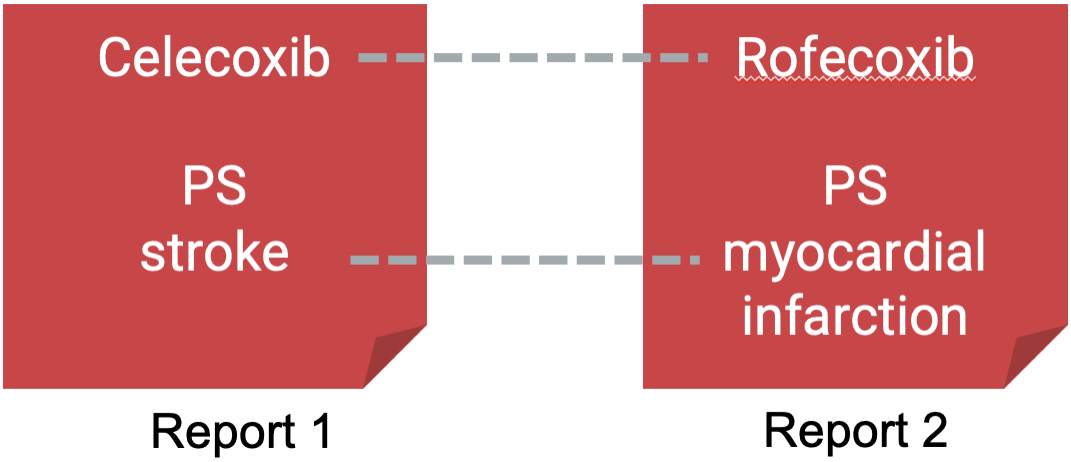

Figure 2.2: FAERS reports showing drugs with similar therapeutic effect might have similar side effect.

if a new drug's vector closely resembles that of a known drug, we can infer potential side-effect profiles by analogy (Figure 2.3). We have successfully used these embeddings for other tasks as well, such as predicting blood–brain barrier penetration from adverse event data [135]. However, aer2vec embeddings are "static" and do not change based on different contexts, such as varying doses, patient demographics, or combination therapies—all of which can alter a drug's side-effect profile. To capture these relationships, transformer-based contextual embeddings have emerged, allowing each instance of a drug–ADR pairing to be represented according to the specific clinical or demographic context in which it appears. This shift from static to contextual embeddings helps create more flexible and accurate models for understanding drug safety data.

2.3.3 Transformers and why they suit EHR data

The paper *Attention Is All You Need* in 2017 introduced the revolutionary Transformer architecture, gradually surpassing earlier recurrent architectures like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) [124]. Transformers inherently capture sequential relationships using positional encoding to indicate token order [54]. This design also excels in parallelizing computations, making it more scalable to large datasets compared with recurrent neural networks.

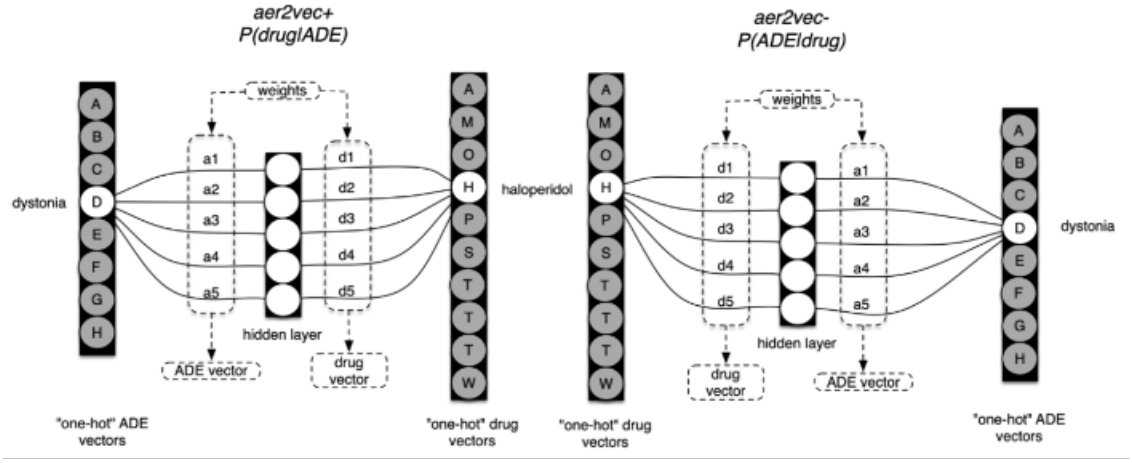

Figure 2.3: Distributed representation of aer2vec [99]. Figure was obtained from the original paper.

In the context of EHR data, transformers are advantageous because they handle long-range dependencies and heterogeneous data types (e.g., medications, diagnoses, lab results). By applying self-attention mechanisms, transformers can focus on the most relevant parts of a patient's longitudinal record—regardless of time gaps or complexity—thereby modeling intricate interactions among symptoms, treatments, and outcomes. Their parallel processing further facilitates efficient training on large healthcare datasets, offering potential improvements in tasks such as predictive modeling, disease progression analysis, and ADR detection [70].

BERT (Bidirectional Encoder Representations from Transformers) expanded on the original Transformer architecture by encoding both left and right contexts simultaneously [25]. This bidirectionality enhances hidden state representations of text, a feature that can also be useful for structured EHR content where multiple clinical elements interact in non-sequential ways.

2.3.4 Specific transformer-based EHR models

Inspired by BERT, Li et al. created BHERT (BERT for EHR) in 2020, where diagnoses are treated as words, clinical visits as sentences, and entire patient histories as documents [70]. BHERT handles temporal irregularities and heterogeneous data types, both continuous and categorical, and showed promising performance in predicting future diagnoses in a specific time frame. Age is the only

continuous variable, and it is not encoded as a raw continuous number. It is converted into a discrete category and is assigned to a trainable embedding vector[70]. Despite these successes, challenges such as effectively encoding long-term dependencies across multiple visits persist.

Another architecture, Med-BERT, was developed using data from 28 million patients in the Cerner Health Facts database [102]. Evaluated on disease prediction tasks with Truven claims data, Med-BERT primarily relied on disease code embeddings and did not enforce a strict order of concepts within visits (relative positional information). The results were promising but the model was expensive to train and failed to outperform simpler baselines on smaller datasets, suggesting that further optimization or domain-specific tuning may be necessary for EHR-based use cases.

Unlike BERT-based systems, the Generative Pretrained Transformer (GPT) employs a unidirectional approach, predicting tokens from left to right [124]. Although GPT has found success in various text-generation tasks, its application to structured EHR data remains limited. One study explored using GPT for EHR timeline forecasting by merging structured and unstructured data, reporting high agreement with expert reviews on 34 synthetic patient timelines [63]. This implies that GPT-based methods could also be adapted for dynamic clinical modeling, yet additional research is still needed in this area.

Among GPT's many versions, GPT-2 stands out for its public availability, and relatively modest computational requirements compared to larger models like GPT4 and open-source models such as LLaMA 2,3 [118, 38]. Researchers with limited resources may still find GPT-2 accessible for fine-tuning and experimentation, especially through the Hugging Face ecosystem, which provides pretrained models and streamlined interfaces for customization.

2.3.5 Limitations

While transformers offer promising results in the medical area, they still face notable barriers. Most models in this domain treat medication and lab information as discrete tokens[70, 102], overlooking the importance of numeric data such as drug dosage or lab results. Training these architectures on large-scale healthcare datasets also demands intensive computational resources, posing challenges for institutions with limited hardware capacity. Moreover, healthcare data can vary considerably across sites due to differences in coding practices, local terminologies, and patient demographics,

raising questions about how well models trained in one institution will adapt to another.

Despite these obstacles, transformer-based approaches remain an exciting approach in EHR data analytics, and ongoing research continues to refine them for clinical settings. The next section addresses one of the most important gaps: the lack of numeric representation.

2.4 Lack of Numeric representation in Transformer

Although transformer architectures have excellent performance in a range of natural language processing (NLP) tasks, they struggle to effectively represent numeric data [129, 34]. According to the World Health Organization (WHO), an adverse drug reaction (ADR) is harm "caused by the use of a drug at regular doses," emphasizing the clinical importance of dose information. Taking 650 mg of acetaminophen twice daily is markedly different from administering 4 g daily, especially in patients with compromised renal function [129]. Yet, most existing pharmacovigilance models or otherwise—treat medication use as a binary variable and disregard the exact dosage. Past research occasionally captured the number of doses per patient but rarely incorporated granular dose or frequency data into machine learning models for ADR detection, especially in elderly patients [79, 45, 46].

Recent innovations indicate growing awareness of this limitation. In 2023, researchers proposed a Hybrid Value-Aware Transformer (HVAT) that modifies a BERT architecture by scaling embedded medical concepts with normalized numeric values [108]. HVAT was tested on Veterans Affairs (VA) data to examine the relationship between ADRD (Alzheimer's disease and related dementia) and physical fitness, and it showed superior performance compared to support vector machines (SVM). However, reproducing these findings is challenging because the VA dataset is not publicly available, and the method was demonstrated on a single use case.

Another study used a similar value-scaling approach to predict numbers in an astronomical dataset [36], while the Labrador model embedded continuous lab values from the Medical Information Mart for Intensive Care (MIMIC)-IV dataset using an empirical cumulative distribution function (eCDF) [13]. The developers of Labrador found that masked language models pretrained on numeric EHR data did not outperform simpler models like random forests or XGBoost for downstream tasks, suggesting that numeric embeddings alone may not be enough to ensure generalization—particularly if the model is not specifically optimized for the target prediction.

Beyond HVAT and Labrador, other techniques aim to better integrate numeric data with transformers. Graded semantic vectors create embeddings by interpolating between orthogonal vectors representing minimal and maximal values, a strategy that has shown promise in encoding numeric information [132, 17]. Rotary positional embeddings (RoPE), was first proposed in 2021, was introduced in open-source models such as Llama II [118, 111], use continuous rotation matrices to handle varying sequence lengths. RoPE could potentially be adapted to represent drug dosages or lab values by transforming them into sine and cosine components. Such an approach would allow transformers to incorporate numeric values more naturally into their attention mechanisms and could enhance the modeling of dose-response relationships in patient data.

Despite these promising developments, fundamental barriers remain. Reproducibility is one of the most pressing issues. Many advanced methods, such as HVAT, are demonstrated on proprietary datasets (e.g., VA data) that are not publicly accessible[108]. Generalizability also poses a significant challenge. Numeric data in EHRs can vary substantially across healthcare institutions due to differences in coding systems or institutions, lab reference ranges, or patient demographics[14]. As a result, a model trained on one dataset may fail to perform well on another, limiting its broader applicability and undermining the potential benefits of numeric-aware transformer solutions.

2.5 Polypharmacy and Drug-Drug Interactions (DDIs)

Historically, DDI detection relied heavily on spontaneous reporting systems (SRSs), such as the FDA Adverse Event Reporting System (FAERS). Researchers often used disproportionality metrics such as the proportional reporting ratio (PRR) and reporting odds ratio (ROR), which compare observed and expected co-occurrences of particular drug–event pairs to flag potential interactions [88]. Other early approaches included association rules, which identify frequent item sets, and logistic regression (LR), which sometimes incorporated covariates like patient age, sex, and interaction terms for drug pairs [122, 123, 41, 88].

As data grew more complex, Bayesian methods offered additional robustness. One such model, the Bayesian Confidence Propagation Neural Network (BCPNN), introduced by Bate et al. in 1998, computes an Information Component (IC) to quantify deviations from expected frequencies under independence assumptions [11]. BCPNN is notably used by the World Health Organization–Uppsala Monitoring Center (WHO-UMC) for early DDI signal detection. Another Bayesian

approach, Gamma Poisson Shrinkage (GPS), uses a Gamma prior with a Poisson likelihood to shrink estimates in cases where raw frequencies might be unreliable [29, 113]. Unlike PRR and ROR, GPS inherently adjusts for uncertainty and often performs better in sparse or noisy datasets. Extensions include a multi-item gamma Poisson shrinker adapted to zero-inflated data, addressing the frequent occurrence of zero cells in adverse reporting databases [42].

Although these methods have proven valuable, many rely on cross-sectional snapshots and under-reporting, especially for less severe events. LR, while simple and interpretable, does not address collinearity across multiple predictors [41, 88]. Bayesian logistic regression can partially mitigate high-dimensional data constraints [2], but the overall complexity of polypharmacy often exceeds what these models can readily capture.

2.5.1 Machine Learning Approaches for DDI Detection

Recent developments of artificial intelligence have greatly expanded the methodologies for DDI research [41]. As the number of approved drugs increases, combinatorial complexity pushes researchers to adopt more advanced, scalable algorithms. One review divides machine learning methods for DDI detection into three groups: traditional ML (e.g., similarity-based methods, Bayesian networks, and SVMs), non-traditional ML, and deep learning [143].

Before the emergence of transformer architectures, deep learning approaches to DDI typically included convolutional neural networks (CNNs), deep neural networks (DNNs), graph embeddings, and graph neural networks (GNNs). Researchers also explored multimodal models such as dual-pathway DNNs (for example, TP-DDI) and factorization autoencoders to handle varied data sources [143]. Although attention mechanisms was developed before the transformer architecture [8], the transformer architecture popularized self-attention and inspired a new wave of attention-based DDI models. A selected list of attention based models include AttentionDDI—a Siamese self-attention multimodal network for analyzing multiple drug similarity measures [105]—and MDF-SA-DDI, which relies on multi-source drug fusion and self-attention-based feature blocks [73, 140, 130]. AMDE, proposed by Pang et al. in 2022, encodes chemical structures (e.g., SMILES) using autoencoders and attention mechanisms [89], while others incorporate knowledge graph paths to further refine similarity measures [112].Not all of these rely on the full transformer framework;

however, they all use the attention modules to capture complex relationships among drug features.

Though these attention-based models often achieve high accuracy, all focus on pharmacokinetic or structural drug data rather than leveraging EHRs. The use of transformer in DDI detection will be discussed in section 2.6.

2.5.2 Why EHR Data Is Rarely Used for DDI Detection

Despite the inherent value of EHRs—due to their longitudinal, patient-centric viewpoint—DDI detection that relies solely on EHR data remains relatively uncommon. Challenges include data incompleteness, heterogeneous coding, and interoperability issues, where records from different institutions may lack consistency or omit crucial events [100]. EHR data can also introduce substantial computational burden, given the volume of time-stamped entries spread across diagnoses, prescriptions, lab results, and clinical notes.

Consequently, many researchers still utilize more standardized sources, such as SRSs or pharmacokinetic databases, when screening for DDIs. Although EHRs contain richer contextual information (e.g., comorbidities, lab values, and treatment timelines), integrating these elements for large-scale analyses requires advanced frameworks that can manage complex, incomplete, and irregularly sampled clinical data.

2.6 Selected previous EHR-Based DDI Studies

A limited but notable set of projects illustrates the feasibility of EHR-based approaches to DDI detection. One example is Vanderbilt University's Drug–Drug Interaction Wide Association Study (DDIWAS), which analyzed 616 drugs in EHR allergy lists to detect potential interactions [134]. This method identified numerous known and novel DDIs, but its reliance on allergy list data restricts general use, and the approach requires extensive expert review. Earlier work by Tatonetti et al. used FAERS data to identify 171 previously unreported DDIs, validated through EHR analysis and multivariate models [115]. Another study by Iyer et al. involved large-scale text mining of over 50 million clinical notes across two healthcare systems(using EHR only), applying adjusted disproportionality ratios to highlight significant drug–event associations; the method achieved AUROCs exceeding 0.8 in both datasets, when evaluated against DrugBank, Medi-Span Drug Therapy Monitoring System and Drugs.com as references[49].

Further exploring QT prolongation, Lorberbaum et al. combined EHR-based analyses with laboratory experiments, revealing how observational data and in vitro testing can reinforce one another [75].

2.6.1 Transformer in DDI detection

Transformer architectures offer a unified way to incorporate structured, unstructured, and temporal data from EHRs. Their self-attention mechanism can capture multifaceted interactions between drugs, clinical events, and patient traits, making them suitable for modeling polypharmacy scenarios. Transformers also adapt more readily to new data through transfer learning, a feature that is useful in continuously evolving healthcare environments.

By learning from large observational datasets, attention-based models may uncover both synergistic therapeutic effects and harmful ADRs. Additionally, their capacity to handle concurrent medication use with varying dosages and timestamps addresses a limitation of older DDI detection methods. Nonetheless, open questions remain about effectively integrating numeric dosage information, ensuring interpretability, and generalizing across disparate health systems. Despite these challenges, transformer-based methods are capable of offering more detailed and adaptive detection of DDIs.

Transformer architectures offer a unified way to incorporate structured, unstructured, and temporal data from EHRs. Their self-attention mechanism can capture multifaceted interactions between drugs, clinical events, and patient traits, making them suitable for modeling polypharmacy scenarios. Transformers also adapt more readily to new data through transfer learning, a feature that is useful in continuously evolving healthcare environments. By learning from large observational datasets, attention-based models may uncover both synergistic therapeutic effects and harmful ADRs. Additionally, their capacity to handle concurrent medication use with varying dosages and timestamps addresses a limitation of older DDI detection methods. Nonetheless, open questions remain about effectively integrating numeric dosage information, ensuring interpretability, and generalizing across disparate health systems. Despite these challenges, transformer-based methods are capable of offering more detailed and adaptive detection of DDIs.

Recently, knowledge-graph-enhanced transformer approaches have shown promise in address-

ing these interpretability and integration challenges. One example is "DDI-GPT: Explainable Prediction of Drug-Drug Interactions using Large Language Models enhanced with Knowledge Graphs," which uses a GPT-based backbone and a comprehensive drug-centered knowledge graph to produce robust DDI predictions. In this framework, segments of the knowledge graph—covering drug targets, pathways, transporters, and other pharmacological details—are serialized into textual "sentence trees" for each pair of drugs. By preserving key biomedical relationships and using self-attention over this enriched text, the model not only outperforms prior graph-based methods on large-scale benchmarks but also exhibits strong zero-shot performance on newly updated FDA data. Moreover, DDI-GPT incorporates an interpretability layer that assigns importance scores to gene and pathway tokens, providing clinicians and researchers with mechanistic insights into why a particular DDI is predicted. This explainability is particularly valuable in real-world settings, where trust in a model's reasoning is crucial and where relevant drug information evolves continuously.

2.7 Summary of Gaps in Prior Work

Despite rapid advancements in both pharmacovigilance and machine learning, several critical gaps persist:

- **Lack of robust temporal modeling:** Many current EHR-based methods overlook the longitudinal sequence of events, focusing on cross-sectional co-occurrences.

- **Insufficient representation of numeric data:** Many transformer-based methods do not incorporate dosage and lab information, even though dose-response relationship, is recognized as a core causality criteria for ADR risk assessment[43].

- **Limited polypharmacy characterization:** While advanced AI models exist for drug-drug interaction detection work on EHR-based DDI detection remains sparse.

2.8 Conclusion

In summary, opportunities exist for improving post-market drug surveillance by leveraging EHR data. However, achieving this goal requires advanced modeling of temporal sequences, integration of numeric dose data, and explicit consideration of polypharmacy. The subsequent chapters of this

dissertation build on these insights, proposing a transformer-based approach that addresses each of these key limitations.

Chapter 3

GENERATIVE TRANSFORMER MODELS FOR ADR SIGNAL DETECTION IN EHR DATASETS

3.1 Abstract

Adverse drug reactions (ADRs) present substantial challenges to patient safety and healthcare systems, often leading to hospitalizations and economic burdens. To mitigate these burdens, post-market surveillance is used to monitor ADRs that did not occur in clinical trials. The commonly employ methods include disproportionality metrics. However, these methods are limited in their ability to capture temporal relationships inherent in ADR occurrences, fail to fully leverage comprehensive data from electronic health records (EHRs), and lack robust mechanisms for signal detection. In this study, we propose novel methods using generative pre-trained transformers for enhanced ADR signal detection in EHR data. Evaluations on two EHRs demonstrate that our models outperform conventional disproportionality metrics and previous machine learning approaches. This study highlights the potential of transformer-based models in advancing pharmacovigilance by integrating comprehensive clinical data

3.2 Introduction

Adverse drug reactions (ADRs) pose a significant challenge to patient safety and healthcare systems, often resulting in hospitalizations, prolonged treatments, and substantial economic burdens [6]. It has been reported that ADRs and other preventable adverse events cost $3.5 billion annually to taxpayers in the United States [7]. However, many ADRs remain undetected or insufficiently characterized during clinical trials, particularly those that emerge after a drug's release to the market. Pharmacovigilance, the science dedicated to detecting, assessing, understanding, and preventing ADRs, thus faces an ongoing struggle to identify these issues in a timely and comprehensive manner.

One obstacle involves data sources and their inherent limitations. The U.S. Food and Drug

Administration's (FDA) Adverse Event Reporting System (FAERS) and the European Vigibase system rely on spontaneous reporting, which often suffers from 1) underreporting [3], 2) bias influenced by news and social media [147], 3) inconsistency, and incomplete information [126]. For example, previous research suggests that approximately 50% to 96% of ADR incidences remain unreported [3]. To address these shortcomings, researchers have turned to electronic health records (EHRs) as a more comprehensive data source. EHRs capture both structured and unstructured clinical information, providing richer data that can enhance ADR signal detection [23, 20].In addition, EHRs are not contingent on reporters.

Despite the promise of EHRs, current analytical methods pose another challenge. Traditional disproportionality metrics focus on simple co-occurrences of drugs and ADRs, failing to account for nuanced relationships and temporal ordering. Such methods cannot easily generalize across similar drugs or related side effects, and they struggle to establish necessary temporal sequences where drug administration precedes an observed adverse event. Moreover, commonly used metrics such as Proportional Reporting Ratio (PRR) and Reporting Odds Ratio (ROR) often exhibit low AUROC when relying solely on EHR data [71, 78]. These shortcomings highlight the need for more sophisticated analytical frameworks that capture the complexity and temporal nature of clinical data.

In response to these issues, large-scale initiatives have attempted to apply more robust approaches. The FDA's 2007 Sentinel initiative integrates claims data and EHR information to improve

pharmacovigilance [15]. The Sentinel program employs TreeScan statistics, which enhance traditional disproportionality measures by performing a temporal likelihood ratio test on predefined adverse event groups [64, 65, 90]. This approach has demonstrated certain level of robustness according to the authors across multiple epidemiological studies [131]. However, the development team behind Sentinel acknowledges the ongoing need for advanced analytical methods, including deep learning and natural language processing (NLP), to further refine ADR signal detection [15].

Recent methodological advancements in machine learning have shown promise in addressing these gaps. Techniques such as logistic regression, gradient-boosted trees, random forests (RF), and support vector machines (SVM) have outperformed conventional disproportionality methods(PRR, Multiple-item gamma Poisson Shrinker) [22]. Despite this progress, disproportionality metrics

remain the most commonly used approaches (48.6%), followed by machine learning (37%) and regression (31%) for drug safety signal identification from EHR data[23]. This indicates that while there is momentum toward more sophisticated models, traditional statistical methods still dominate, underlining the importance of continued methodological innovation.

Deep learning, particularly the application of language models, has shown remarkable promise in health informatics [138]. Models inspired by language processing are especially relevant because clinical events over time resemble sequences of words, enabling the application of NLP techniques to healthcare data. Before transformer architectures became widespread, embedding-based methods such as skip-gram models showed promise in surpassing traditional metrics and aiding in drug repurposing[99]. Embeddings trained on observational data have improved pharmacovigilance signal detection from EHR notes by capturing positional and directional dependencies within clinical texts[85]. These findings suggest that more advanced architectures could further improve the detection of ADR signals by leveraging temporal context and complex semantic relationships.

However, incorporating the full complexity of patient data into these models remains a significant challenge. Patients often have multiple visits over several years, receive numerous prescriptions, and produce extensive diagnostic and laboratory results. Early methods represented patients using high-dimensional "one-hot" vectors, later applying dimensionality reduction techniques such as principle component analysis (PCA) or non-negative matrix factorization(NMF). More recently, deep learning methods, specifically de-noising auto-encoders, have produced richer, low-dimensional patient representations, outperforming conventional techniques[84]. As the volume and complexity of clinical data grow, so does the need for models that can integrate these heterogeneous data sources effectively.

The rise of transformer-based architectures provides a natural next step. Transformers have advanced NLP by efficiently handling long sequences and capturing complex dependencies without relying on recurrent structures[125] and it addresses the problem of lacking contextual approach. Instead of assigning one fixed vector per word/drug, the representation changes depending on the surrounding text or data. Generative pre-trained transformers (GPT), developed by OpenAI, represent one of the state-of-the-art methodologies for language modeling and are inherently well-suited for sequence data. Unlike traditional neural networks, transformers can better manage extensive training data and use attention mechanisms to provide context across entire sequences.

Their ability to model order and context makes transformers highly applicable to the temporal and heterogeneous nature of EHR data.

Several transformer-based models for EHRs have already emerged. BEHRT, developed in 2020, extends BERT to represent diagnoses as words, visits as sentences, and a patient's medical history as a document, employing positional embeddings to capture temporal structures[70]. BHERT addresses challenges such as managing long-term dependencies, heterogeneous data types, and irregular recording intervals. Similarly, Med-BERT trained on large-scale EHR datasets has shown promise in predicting future diagnoses[102]. However, these bidirectional models(model processes text by simultaneously looking at preceding and succeeding context) may not align perfectly with the inherently causal and sequential nature of healthcare data, where future events should not inform the interpretation of past observations.

In contrast, unidirectional GPT architectures mirror the forward progression of time in healthcare data, potentially offering a more suitable modeling paradigm. GPT-2, for example, uses token and positional embeddings to interpret sequences in a left-to-right manner, which may better respect the temporal ordering critical for causal inference. While limited research has explored GPT architectures directly for EHR timeline prediction[63], the method was developed for a different application. Additionally, although GPT-like models have been adapted to extract clinical entities from text[69], their direct application to ADR signal detection is virtually unexplored.

In this chapter, I propose novel methods that integrate GPT-2-based architectures with EHR data to improve ADR signal detection. GPT-2 was chosen for its accessibility, transparency, and computational efficiency compared to more resource-intensive models like Llama models and Deepseek[101]. Our central hypothesis is that a transformer-based generative model, which naturally incorporates temporal ordering, will outperform conventional disproportionality metrics and non-transformer-based machine learning approaches on detecting ADR. Additionally, I aim to understand the role of temporal information in ADR signal detection, providing insights that can inform future model development and data integration strategies.

3.3 Methods

3.3.1 Datasets

MIMIC IV Hospitalization

I used the Medical Information Mart for Intensive Care (MIMIC-IV) dataset as the training set because of the granularity of its data. MIMIC-IV is a publicly available database sourced from the inpatient units at Beth Israel Deaconess Medical Center that includes patient lab, measures, orders, diagnosis, procedures, treatments, and deidentified free-text clinical notes from visits 2008 and 2019. One advantage of MIMIC-IV is that it contains a medicine administration record, which is essential for our pharmacovigilance work. The dataset contains hospital admissions and ICU admission records for 180,733 and 50,920 patients, respectively[52]. Since the data are publicly available, they can be used to compare different models in the field to increase reliability. The dataset contains all the hospitalization data from MIMIC IV. Diagnosis, admission, lab, and medication data were used. Using the lab data, I created a binary flag for the targeted lab measures and ternary flag is created for the rest of the lab measures as low, normal and high. All the lab and drug description records were sorted by hospitalization and charting time of an individual record. The admission records were sorted by hospitalization identifier and the discharge time.

UW EHR

This study uses a dataset created by the Institute of Translational Health Sciences(ITHS) in 2020 from the University of Washington Enterprise Data Warehouse(EDW). The dataset was originally extracted based on the criteria derived from Ryan et al.'s ODHSI ADR reference set and the Exploring and Understanding Adverse Drug Reactions (EUADR) reference set in order to capture the information that would be required to assess the strength of the association between drugs and events in this set[104]. According to the patient selection method, the medical history of patients with at least one of the ICD diagnosis codes from the reference standards in the EHR was selected, including anaphylactic shock, aplastic anemia, erythema multiforme, mitral valve disease, neutropenia, rhabdomyolysis, acute renal failure, acute liver failure, acute myocardial infarction and gastrointestinal hemorrhage for patients with data recorded between 2000 to 2020. The information

from six months before the diagnosis and 12 months after the diagnosis was extracted. Patients younger than 18 years old who had no events within 18 months from the initial timestamp were excluded. All eligible encounters, comorbidities, medications, procedures, labs and clinical notes were included. Unfortunately, this dataset only contains crude timestamps(by date) that do not have the exact time of procedures and diagnoses. The notes were processed using MetaMap to extract UMLS concepts from clinical notes[4](extracted about 50K medical concepts). Data was processed in the same way as the MIMIC IV hospitalization data.

3.3.2 Representation of hospitalization from clinical data

The input data for this model was derived from EHRs and encoded to represent hospitalizations through a wide range of clinical data: lab test results, medication records, diagnoses, and medical observations. Each hospitalization was represented as a chronological sequence of medical concept tokens, such as laboratory test, medication and diagnosis codes. Figure 3.2 shows an example of the EHR sequence. To enhance the model's ability to focus on recent clinical information or to predict the next event of interest, these sequences were processed in reverse order, with the most recent tokens appearing first. The processed clinical sequence was then converted into a structured dataset format using Hugging Face's Dataset library[1]. To maintain consistency and handle variations in sequence length, each record was limited to 512 tokens, truncating longer sequences and padding shorter ones with zeroes. This uniform representation allowed the transformer-based model to process hospitalizations effectively, capturing temporal patterns in clinical events.

3.3.3 Tokenizer

A whitespace tokenizer from the Hugging Face's tokenizers library[2], was used. The tokenizer was configured with a vocabulary generated from the clinical data, and padding was enabled to ensure uniform sequence length. The vocabulary size of the tokenizer was calculated and incorporated into the model configuration to align the model's embedding layer with the tokenization process. During tokenization, sequences exceeding the 512-token limit(for computational efficiency) were truncated

[1]https://huggingface.co/docs/datasets/en/index

[2]https://huggingface.co/docs/tokenizers/en/index

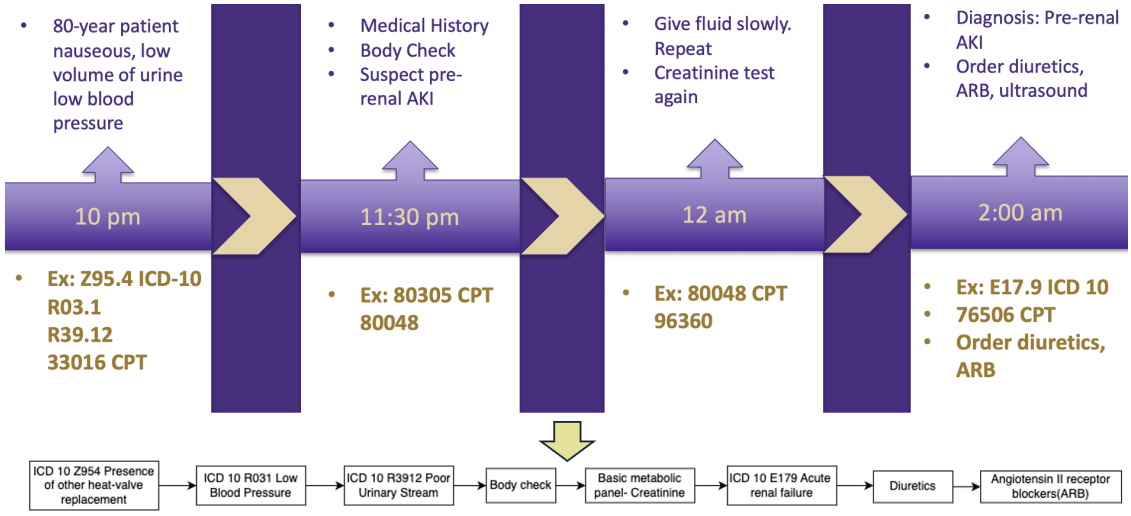

Figure 3.1: EHR sequence example

to retain only the most recent tokens, while shorter sequences were padded with zeroes.

3.3.4 Baseline Models

I used the disproportionality analysis and the TreeScan [64, 65] as the baselines. I calculated the PRR for each drug-ADR pair from the reference sets described above[29]. PRR is a metric used to detect potential drug safety signals by comparing the proportion of reports for a specific adverse event associated with a particular drug to the proportion of the same adverse event reports for all other drugs in the database. For the second baseline, the TreeScan method, which is the current signal detection method using in the FDA sentinel initiative that accounts for the temporality effects. I utilized the "Purely Temporal Scan Statistic" option in the TreeScan software[3] with uniform probability model in the software, which accounts for the number of ADR cases and the time of developing ADRs. The data was prepared following the example in the manual of TreeScan software. The time period was calculated between the initial use of drug to the first ADR event. For each hospitalization, the maximum drug-ADR count is 1, if the drug-ADR pair occurs multiple times in one hospitalization, I would only count it as 1 for disproportionality based methods. The

[3]https://www.TreeScan.org

same procedure was applied to all hospitalizations and the (total number of drug-ADR, time period in days) tuple was recorded to feed into the TreeScan software. The software can handle multiple signal detection tasks simultaneously and address the multiple testing. I applied the method on both MIMIC and UW-EHR datasets. I take all the per-interval p-values by TreeScan for a given (drug, condition) pair and transform them via $-log_{10}(p)$. By summing them, I obtain a continuous log-sum score that reflects how strongly the pair is associated with the outcome across all intervals.

For comparison purpose, I also applied the proximity-based method described previously in Mower et al[85]. The study uses locality sensitive neural embeddings dervied from EHR clinical notes to improve ADR detection. The sequences derived from MIMIC and UW-EHR went through Semantic Vector Java package[4] and the Embeddings Augmented by Random Permutation(EARP) embeddings were created[85]. EARP directional consider a window of several terms around a given observed concept, but directional information was encoded as well. I took the average between the left and right context vectors of the observed term within the window for evaluation(within 5 tokens).

Because TreeScan's purely temporal scan looks only for time intervals with more events than expected, it yields only 'positive signals' (rather than negatives). Typically, one must set a minimum threshold(n>2) for the number of ADR events or exposures so that it's statistically meaningful to detect an excess. We used a threshold of three observed ADR occurrences per drug–ADR pair; any pair below that threshold was excluded. At the end, 506 drug-ADE pairs were included in the TreeScan analysis for the MIMIC hospitalization data(Figure 3.3). For UW data, 447 drug-ADE pairs were included(Figure 3.4). The reverse model uses the ADR in the reference sets as prompts and predicts the probability of the previous token being the drug.

3.3.5 GPT2 model

I trained a GPT-2 model from scratch(no pretrained weights) using the OpenAI GPT-2 architecture provided by the Hugging Face library. Hugging Face's open-source implementation of GPT-2, published in 2019 [101], enabled the construction of the model architecture, with configurations reaching up to 1.5 billion parameters. I trained a few different configurations of the GPT2 models

[4]https://github.com/semanticvectors/semanticvectors

with small variations in each one on the MIMIC IV and UW EHRs. The Transformer architecture uses positional embeddings(PE) to incorporate order information and word token embeddings (WTE) to represent the semantic meaning of each token. These embeddings are typically stored in lookup tables (embedding matrices), making them easily retrievable[101].

I replaced the WTE with the semantic embeddings that were developed using the Semvecpy package[5] to learn the semantic relationship among the tokens in the EHR sequences for both MIMIC IV and UW EDW data. The original work on semvecpy and its variants are from a previous publication[16]. The development of the package was inspired by the *word2vec* model[83].

- GPT2 model with positional embeddings
- GPT2 + pretrained embeddings
- GPT2 without positional embeddings

For the pretrained embeddings, in every case, models were trained with 5 negative samples, a window radius of 5 from the token, 5 training epochs and real valued vectors of 768 dimensions. The same method was repeated twice for both the MIMIC and UW datasets. See the parameters in Table A.1 in the appendix(Appendix Table A.1). The sequence was trained in a reverse order meaning the most recent tokens were fed in first and the prediction would be retrospective. For each sequence and to improve efficiency, the last 512 tokens for each sequence were used. By turning the positional embeddings on and off, I will be able to investigate the impact of temporality.

The following hyperparameters were used during training: an effective batch size of 16, achieved via gradient accumulation over 16 steps; a learning rate of $6 * 10^{-6}$ with a weight decay of 0.1; warm-up steps set to 2000; and a total of 8 training epochs. Early stopping was implemented with a patience of two evaluation steps to prevent overfitting and be more efficient. The model was optimized using AdamW[76], and mixed-precision training (FP16) was applied to enhance computational efficiency. The custom tokenizer used for this study incorporated padding tokens and ensured consistency across varying input lengths, with all sequences truncated or padded to a maximum length of 512 tokens(Table A.1).

[5]https://github.com/semanticvectors/semvecpy

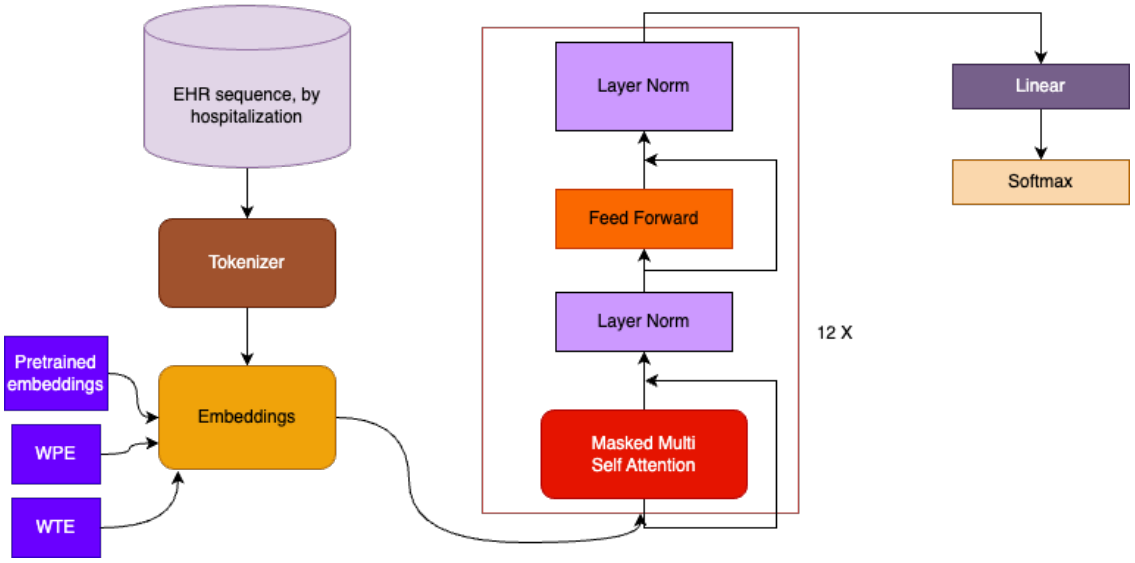

Figure 3.2: GPT2 model illustration, WPE: positional embeddings, WTE: token embeddings

Reference sets

Four evaluation reference sets with ADR-drug pairs were used. The OMOP ADR reference set and EUADR are well-validated in previous studies[104, 120]. All reference sets contain drug-ADR pairs and a binary label indicating if the drug is associated with the ADR. Due to the limited numbers of positive and negative ADR-drug pairs, I combined the EU+OMOP together in the inference stage. The OMOP reference set contains four ADRs: acute renal failure, acute liver injury, acute myocardial infarction and gastrointestinal bleed. The EUADR reference set contains these four along with six additional ADRs: anaphylactic shock, erythema multifore, mitral valve disease, neutropenia, aplastic anemia, and rhabdomyolysis[120]. I utilized two more reference sets, Arizona Center for Education and Research on Therapeutics (AZCERT) and a set curated by a Minnesota group (UMN). The AZCERT list was developed by combining ADR cases reported on CredibleMeds.org, medical literature and new FDA labels. All the ADR-QT Prolongation and Torsade De Pointes pairs underwent a rigorous evaluation procedure and casual relationships have been confirmed[133, 128]. Lastly, a fourth set was curated by Dauner et al.(from UMN) which contains nine conditions: dysglycaemia, hepatic decompensation and hepatic failure, angioedema, ascites,

encephalopathy, hepatic encephalopathy, hyperglycaemia, hypoglycaemia, jaundice, oesophageal varices haemorrhage and varices oesophageal with 155 positive ADR-drug pairs and 45 negative pairs[22]. Each pair was well-validated in other research and has a binary label of whether the drug causes the ADR (Table. 3.1).

Table 3.1: The Adverse Drug Reactions and the number of binary labels

Reference Set	**Label +/-**	**#Adverse Drug Reaction**
OMOP	164+, 235-	Acute renal failure, acute liver injury, acute myocardial infarction, gastrointestinal bleeding
EUADR	44+, 50-	Acute renal failure, acute liver injury, acute myocardial infarction, gastrointestinal bleeding, anaphylactic shock, erythema multiforme, mitral valve disease, neutropenia, aplastic anemia, rhabdomyolysis
AZCERT	55+, 865-	QT Prolongation and Torsade De Pointes
Minnesota	110+, 45-	Dysglycaemia, hepatic decompensation and hepatic failure, angioedema, ascites, encephalopathy, hepatic encephalopathy, hyperglycaemia, hypoglycaemia, jaundice, oesophageal varices haemorrhage, varices oesophageal

3.3.6 Evaluation

For model training, I calculated the loss for each epoch and monitored its decrease over time. For model inference, the ability of each method to detect adverse drug reaction (ADR) signals was evaluated by ranking positively labeled drug-ADR pairs before negatively labeled pairs using the Area Under the Receiver Operating Characteristic Curve(AUROC). The evaluation utilized all four reference sets, with ADRs mapped to their respective International Classification of Diseases (ICD-9, ICD-10) codes. For ADRs associated with multiple ICD codes (e.g., acute renal failure mapped

to eight codes across both versions), I calculated a weighted average AUROC across all ICD codes within each ADR. Reference set-specific AUROCs were computed to enable comparisons across methods. The TreeScan method was used as a baseline, and I ensured that all drug-ADR pairs met the constraints for inclusion across all baseline models, including PRR, TreeScan, and permutation-based approaches. For the overall AUROC, all drug-ADR pairs across the four reference sets were aggregated for a single calculation.

I focused on evaluating the likelihood that, given an ADR from the reference set, the model predicts the drug of interest as the proceeding token. This can be conceptualized as treating the ADR as a prompt within the GPT-2 model and examining the likelihood of the drug being the predicted proceeding token. ADRs and drugs with a frequency below 10 were excluded from the analysis. Too few samples result in unbalance in evaluation. Both sigmoid and softmax probabilities were calculated. The softmax function provides the probability of the drug token relative to all possible tokens in the dataset, while the sigmoid function offers a non-mutually exclusive probability for each drug independently. To account for ICD-code-specific contributions, the weighted probabilities were computed as the summation of the product of the probability for each ICD code and its associated weight.

For instance, there are 8 ICD codes associated with acute renal failure, and I calculated each ICD code's weight based on its count. Mathematically, this is expressed as:

$$\text{Weighted Sum } P(\text{drug} \mid \text{ADR})_{j,c} = \sum_{i \in \mathcal{I}_c} \Big[P\big(\text{drug}_j \mid \text{ADR}_c\big)_i \times \text{ICDWeight}_i \Big].$$

where i represents the ICD codes associated with the ADR, j represents the drug, and c represents

the condition.

$\mathcal{I}_c$: Set of ICD codes for the adverse drug reaction c,

j : Index for the drug,

c : Index for the ADR (or condition),

i : Index over ICD codes in $\mathcal{I}_c$,

$P(\text{drug}_j \mid \text{ADR}_c)_i$: Probability for drug j given ADR c, specific to code i,

ICDWeight_i : The weight associated with ICD code i.

Both approaches were applied to calculate softmax and sigmoid probabilities, producing metrics such as Weighted Softmax/Sigmoid. For the comparison of the models with positional and without positional embeddings, the better AUROC was calculated and reported by ADR, by reference set, and overall.

3.4 Results

Table 3.2 compares our baseline methods (PRR, TreeScan, EARP_dir) and the GPT-2–based models for both MIMIC and UW data. For the MIMIC dataset, GPT-2 (without pretrained embeddings) generally performed comparably to the PRR and TreeScan baselines, achieving an AUROC of around 0.50–0.53 when aggregating across all reference sets. EARP_dir slightly surpassed GPT-2 alone on some subsets (0.56 vs. 0.53 overall). However, GPT-2 combined with pretrained embeddings consistently surpassed all baseline methods, reaching 0.58 across all sets.

For the UW dataset, the GPT-2 model alone outperformed the baselines (0.64–0.65 vs. 0.50–0.58), and GPT-2 plus pretrained embeddings further improved performance, up to 0.66 overall. Although TreeScan excelled in certain subsets (e.g., the UMN references, AUROC 0.65) and EARP_dir was best for certain AZCERT intervals (0.63), neither approach matched the top scores from the transformer-based methods when averaging across all reference sets.

To illustrate performance on individual conditions, we examined acute renal failure (AKI) and acute gastrointestinal hemorrhage (GH). Using MIMIC data, the GPT-2+embeddings model yielded an AUROC above 0.60 for AKI (10,714 AKI+ vs. 75,457 AKI–) and 0.74 for GH (1,802 GH+ vs. 84,369 GH–). With UW data, GPT-2 achieved 0.68 for AKI (1,829 AKI+ vs. 15,625

AKI–) and 0.88 for GH (380 GH+ vs. 17,074 GH–), suggesting that the richer note-derived concepts in UW may bolster detection for some conditions.

Finally, Table 3.3 presents results comparing GPT-2+embeddings models with (Pos) and without (woPos) positional embeddings. Consistent with the broader findings, we observed no uniform advantage of explicit positional embeddings; on some subsets (e.g., MIMIC AZCERT), positional embeddings increased the AUROC, while in others (e.g., MIMIC OMOP+EU), disabling them slightly improved results.

Overall, these findings indicate that GPT-2 alone can match or exceed classical baselines, particularly in the UW data, and that GPT-2+embeddings offers further gains across reference sets—highlighting the potential of transformer-based methods for ADR signal detection in EHR data.

Table 3.2: Comparison of non-value aware models on MIMIC and UW inpatient data

MIMIC inpatient summary from non-value aware models

	PRR	TreeScan	EARP_dir	GPT2 Weighted Average Softmax	GPT2 Weighted Average Sigmoid	GPT2+embeddings Weighted Average Softmax	GPT2+embeddings Weighted Average Sigmoid
OMOP + EU (82+, 101-)	0.407	0.384	0.490	0.506	**0.521**	0.466	0.491
UMN (58+, 16-)	0.483	0.475	0.507	0.497	0.530	**0.533**	0.514
AZCERT (33+, 215-)	0.459	**0.630**	0.582	0.460	0.461	0.623	0.616
Across all sets (173+, 332-)	0.499	0.451	0.560	0.501	0.530	0.544	**0.583**

UW inpatient summary from non-value aware models

3.5 Error Analysis by Conditions

Table 3.4 shows the performance of the reversed GPT-2–based model for four key conditions, comparing the results to the AUROCs reported by Li et al. using FAERS data[71]. They adapted an approach, adjusted AUROC, by first estimating confounding-adjusted associations for each

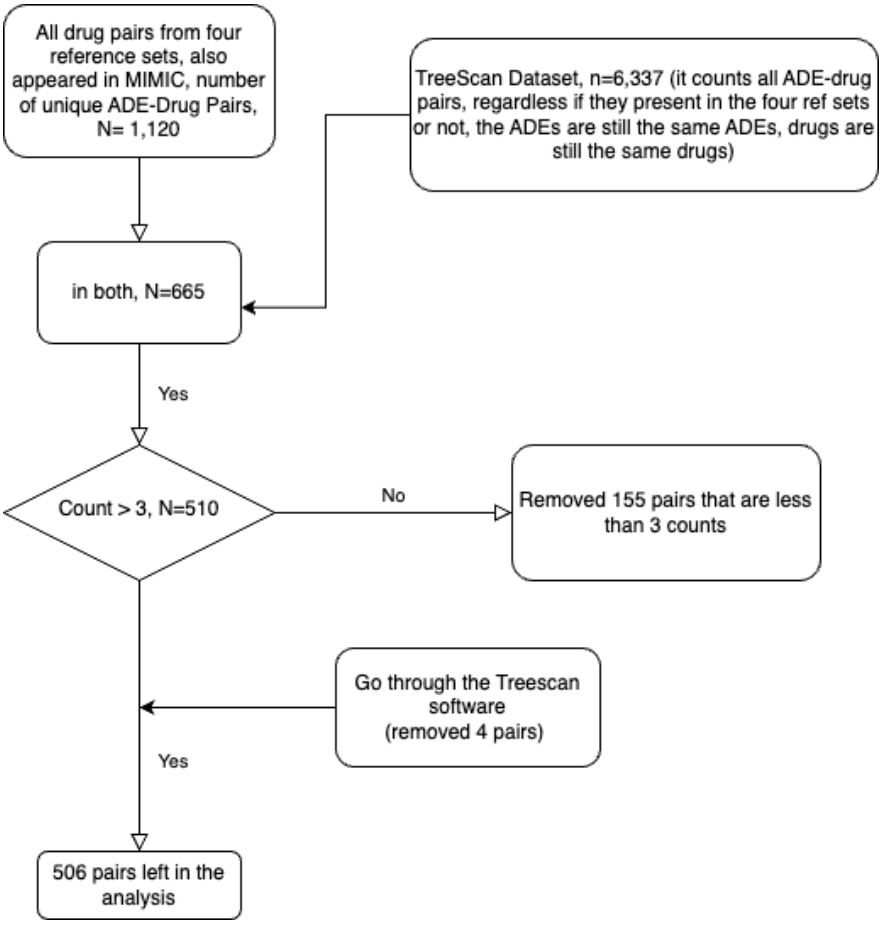

Figure 3.3: MIMIC drug inclusion flowchart

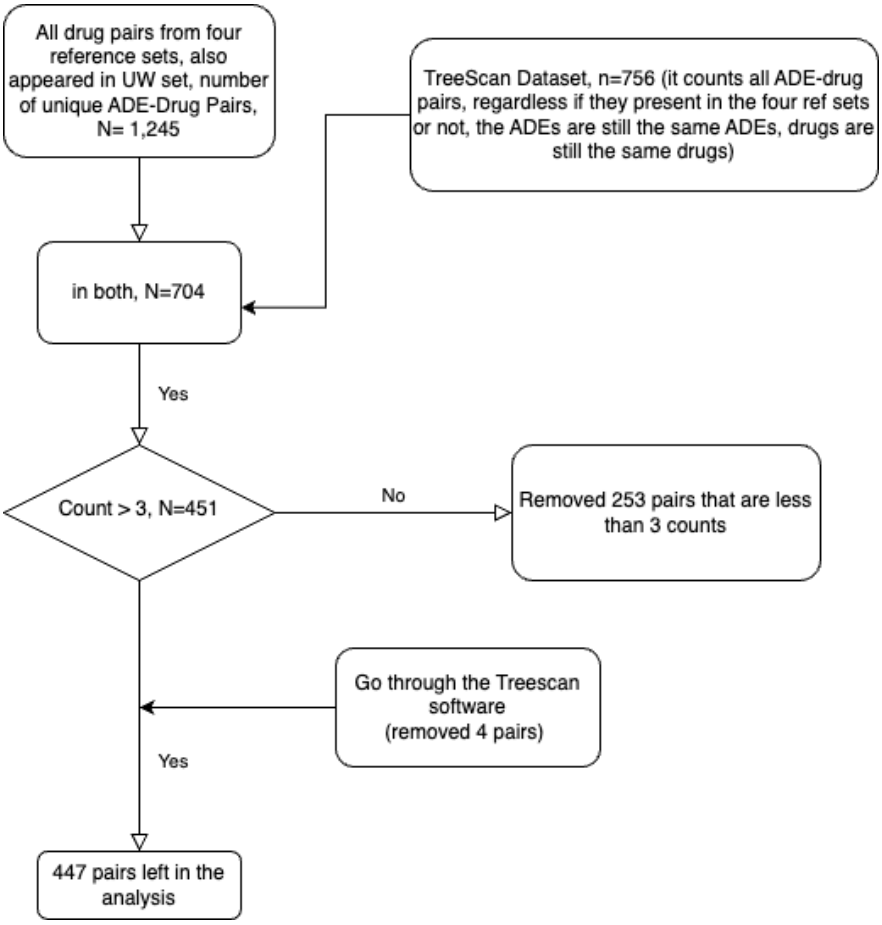

Figure 3.4: UW drug inclusion flowchart

Table 3.3: GPT2+ pretrained Embeddings with vs without positional embeddings. woPos:the configurations without positional embeddings. Pos: the configurations with positional embeddings.

	MIMIC-woPos	MIMIC-Pos	UW-woPos	UW-Pos
OMOP + EU	0.593	0.491	0.581	0.640
UMN	0.582	0.514	0.613	0.632
AZCERT	0.520	0.616	0.624	0.590
Across all sets	0.548	0.583	0.640	0.655

drug–ADR pair, then measuring how well these associations separated known positives from known negatives in the OMOP reference standard via a two-step lasso model[71].

For each condition, we note the number of positive and negative cases in MIMIC and UW, as well as the model's weighted softmax and weighted sigmoid outcomes. In general, UW data yields higher AUROCs than MIMIC for acute liver failure, myocardial infarction, and gastrointestinal hemorrhage, even though MIMIC typically provides larger sample sizes. This pattern suggests that having clinical note-based concepts (available in UW) can enrich the context for detecting ADR signals, whereas MIMIC's more precise temporal records do not necessarily translate to better predictive power. One possible explanation is that transformers, in our current configuration, might not fully exploit day-to-day or hour-to-hour timestamps, particularly if the ICD diagnosis codes are entered at the end of hospitalization.

Examining the specific conditions, gastric bleeding stands out with AUROCs around 0.73–0.74 in MIMIC and an even higher 0.87–0.88 in UW, each near or above FAERS's 0.83 benchmark. AKI also shows moderate success (0.60 in MIMIC vs. 0.68 in UW), though still below FAERS's 0.89—indicating that observational EHR data can sometimes dilute the observed drug–renal outcome relationship. Meanwhile, acute liver failure and myocardial infarction remain challenging in MIMIC (AUROCs around 0.33–0.37), but moderate in UW (0.56–0.59), suggesting that diagnosing or coding these events in UW's clinical notes may better capture the signal. Nevertheless, neither EHR consistently matches FAERS levels, which aligns with previous findings in the literature [71, 78].

Other conditions in the EUADR set, such as aplastic anemia, neutropenia, and rhabdomyolysis, have substantially fewer positives and often under a few hundred instances in MIMIC or UW. For these rarer events, AUROCs typically hover near 0.50 or show missing values, implying that the sample size is too small to reliably detect a drug-safety association. Beyond these four conditions, the AZCERT set comprises only one condition, while the UMN reference sets include additional ADRs (e.g., ascites) exhibiting variable performance (AUROCs 0.4–0.8) across the two datasets. In general, such infrequently coded or confounded conditions yield only sparse signals for our GPT-2–based approach to exploit, which further underscores the importance of robust clinical documentation and additional data sources for enhancing ADR detection.

3.6 Discussion

In this chapter, I introduced a novel transformer-based approach for ADR signal detection, built on a GPT-2 model trained using EHR data from two large healthcare systems. Across multiple experiments—comparing pretrained vs. non-pretrained embeddings, and enabling vs. disabling positional embeddings—the GPT-2–based methods consistently outperformed traditional disproportionality metrics (PRR), the temporal TreeScan method, and our previously developed permutation-based approach (EARP_dir). This aligns with broader observations that machine learning and deep learning methods generally surpass classical metrics when working with rich EHR data.

Despite TreeScan's capacity to incorporate temporality, it did not always surpass PRR, echoing past findings that TreeScan may detect fewer signals but is advantageous for controlling Type I error rates. Meanwhile, EARP_dir uses local permutations for more nuanced sequence modeling than PRR or TreeScan, yet it still lagged behind the best GPT-2 configurations—suggesting that the greater complexity of a transformer can pay off in pharmacovigilance. Notably, GPT-2 alone, without pretrained embeddings, often matched or exceeded baseline approaches in the UW dataset (AUROCs 0.64–0.65 vs. 0.50–0.58 for PRR/TreeScan), though in MIMIC it reached 0.50–0.53 overall—slightly below EARP_dir (0.56). Nevertheless, this underlines that even a relatively straightforward generative model can compete with established methods in EHR-based signal detection.

The GPT-2 model with pretrained embeddings yielded further gains. For MIMIC, GPT-2+embeddings improved overall AUROCs from 0.53 to 0.58, and in UW it rose to 0.66 outperformed

other baselines. The new condition-level error analysis(referring to the previous section) revealed that GPT-2+embeddings commonly handled AKI and GI bleed better than earlier methods, matching or surpassing FAERS benchmarks in GI bleed. However, acute liver failure (ALF) and myocardial infarction (AMI) remained difficult to detect, particularly in MIMIC, where AUROCs hovered around 0.33–0.37. The UW dataset showed moderate success for ALF and AMI (0.56–0.59), suggesting that clinical note-based data in UW may enrich ADR context. Yet neither EHR consistently matched FAERS performance levels, in line with previous literature findings[71, 78].

For including positional embeddings or not, we found no consistent advantage in including them. While adding positional encodings helped in some subsets (e.g., MIMIC AZCERT), it was mildly detrimental in others (e.g., MIMIC OMOP+EU). This finding aligns with prior time-series language modeling research[31, 114], indicating that explicit positional signals do not always improve EHR-based modeling—especially if ICD codes are typically assigned late for billing, rather than capturing real-time progression.

Beyond these core experiments, I also tested forward vs. reversed GPT-2 training. Although forward modeling is more intuitive for predicting future events, the reversed approach—having the model "see" the ADR and predict preceding treatments—proved more effective. This reversed ordering aligns with the clinical logic that medications generally precede side-effect diagnoses, producing more coherent training examples for the model.

Data quality and institutional coding practices also shape these outcomes. MIMIC, representing a single hospital system, has more precise timestamps but fewer note-derived concepts, whereas UW spans a longer timeframe (2000–2020) and has day-level timestamps offset by a richer, note-based vocabulary. These differences help explain why GPT-2 alone trails EARP_dir in some MIMIC subsets, yet surpasses classical baselines more robustly in UW.

In the error analysis, we observed an AUROC gap between Li et al.'s approach and ours, and it is important to note several key methodological differences[71]. First, Li et al. leverage both FAERS and EHR data, whereas our method relies only on EHRs. Second, unlike Li et al.'s confounder-adjusted, p-value–calibrated framework, we do not perform confounder screening or negative-control calibration; our transformer model directly outputs predictions. Third, we exploit self-attention mechanisms for temporal sequencing rather than a lasso regression.

Moreover, the results from the error analysis(Table 3.4) further demonstrated that GI bleed is

comparatively easier to detect in both EHRs, while AKI shows moderate performance. By contrast, ALF and AMI could have been confounded by other clinical factors or by delayed coding. Moreover, some rarer EUADR conditions (e.g., aplastic anemia, neutropenia, rhabdomyolysis) having too few positive examples—often below a few hundred—yielding near-random AUROCs or missing values.

There are a few limitations. First, the UW dataset was extracted under narrower selection criteria (232 drugs or 10 ADRs in EU+OMOP), which could bias condition coverage. Second, event timestamps in both MIMIC and UW may not fully reflect true real-time clinical progression, complicating causal modeling. Third, while performance improves over older methods, certain ADRs—particularly ALF and AMI—remain difficult to capture. Finally, this work involved only two major healthcare systems, so broader generalizability needs to be further validated.

For future directions, I will attempt to address the limitations by acquiring datasets from different sources or include the clinical notes from MIMIC as well. In addition, I would like to verify the method using bigger models with more parameters and better trained. The next-generation generative models could further improve performance. Another direction I considered is to include multimodal data. For example, using EHRs alone can provide a comprehensive view of individual patients, but incorporating spontaneous reporting system data, like the FDA Adverse Event Reporting System (FAERS), has the potential to further improve performance.

3.7 Conclusion

This chapter builds a foundation for the next two chapters by implementing a GPT-2 model trained from scratch with some variations by replacing the WTE with pretrained embeddings, which demonstrates another use case of transformer models in pharmacovigilance.

Table 3.4: Error analysis for the four conditions in our EHR-based approach, compared with the method described by Li et al.[71], who employed a two-stage LASSO regression framework on FAERS (spontaneous reporting) data.

		MIMIC IV Inpatient Reverse			**UW Inpatient Reverse**		
	Compare to FAERS (Li et al., 2015)	**N(+cases, -controls)**	**AUROC Weighted Softmax**	**AUROC Weighted Sigmoid**	**N(+cases, -controls)**	**AUROC Weighted Softmax**	**AUROC Weighted Sigmoid**
Acute kidney failure	0.89	10048+, 39362-	0.60	0.58	1830+, 15624-	0.68	0.67
Acute liver failure	0.70	3435+, 45975-	0.34	0.37	1115+, 16339-	0.56	0.57
Acute myocardial infarction	0.65	3357+, 46053-	0.33	0.35	347+, 17107-	0.59	0.57
Acute gastrointestinal hemorrhage	0.83	1486+, 47924-	0.74	0.73	376+, 17078-	0.88	0.87

Chapter 4

ENHANCE THE GENERATIVE TRANSFORMER MODEL FOR SIGNAL DETECTION BY INTEGRATING VALUE AWARE EMBEDDINGS

In this chapter, I build upon the GPT2 model from Aim 1 and introduce value-aware transformers with various value-aware embeddings. The goal is to further improve the transformer based ADR detection methods. The introduction shares some similarity with Aim 1, data processing and evolution procedures are as described in Chapter 3.

4.1 Abstract

Building on the insights from the previous chapter, this work addresses the need for incorporating numeric data—such as lab results and medication doses—into transformer-based models for pharmacovigilance. I introduce a value-aware GPT-2 architecture that encodes numeric information into token embeddings, extending the traditional transformer framework. Specifically, I investigate both graded and rotary embeddings (in universal, concept-specific, and trainable variants) and integrate them alongside GPT-2's standard token and positional embeddings. Experiments on data from two EHRs demonstrate notable performance gains over non-value-aware baselines. These findings emphasize the importance of numeric representation in improving ADR detection with generative transformer models.

4.2 Introduction

Despite the transformer model's success in improving ADR signal detection, it has fundamental limitations. A notable one is transformer's inability to comprehend numeric data effectively[129, 34]. Before the advent of GPT-o1, transformer models had limited capability to recognize that 500 mg is greater than 5 mg of a drug. Similarly, the latest Deepseek R1 model reasoning model, still at the time of this writing could not reason that 7.110 is less than 1000. This is because transformer as a language model that the token repsentations of these values do not explicitly encode their magnitude. Even with improved reasoning and arithmetic skills, the underlying mechanism

by which GPT-o1 achieves this understanding remains unclear. This limitation is critical when monitoring drug safety. According to the WHO's definition, an ADR is: "harm directly caused by the use of a drug at regular dose"[18], emphasizing the importance of drug dose information. For example, taking 650 mg of acetaminophen twice daily is substantially different from taking 4 g daily for a patient with deteriorating renal function. Capturing dose information could enhance machine learning models for ADR detection by reducing noise in these models[21].

Several recent attempts have been made to integrate numeric values into transformers in the medical field, highlighting the growing attention to this issue. In 2023, a study introduced the Hybrid Value-Aware Transformer (HVAT), which jointly models longitudinal and non-longitudinal EHR data. HVAT features two branches: the main branch, similar to conventional transformer models, processes longitudinal data, while the other branch uses a feed-forward layer to handle non-longitudinal data such as sex and race/ethnicity. HVAT modifies the bidirectional encoder representations from transformers (BERT) by scaling the value embeddings of medical concepts with normalized numeric values. The researchers evaluated this architecture using Veterans Affairs (VA) data, specifically examining the relationship between ADRD and physical fitness. The performance was promising, outperforming other machine learning methods such as support vector machine (SVM) [108]. However, the VA data is not publicly available, and the lack of implementation details makes it difficult to reproduce these results.

Another recent study used a similar approach to scale embedding vectors based on numeric values and showed promising results for numeric tokenization strategies on metrics[36] on a large astronomical dataset. This method shares many similarities with HVAT. Furthermore, a recent publication introduced another approach, Labrador, which encodes continuous lab values from the Medical Information Mart for Intensive Care (MIMIC)-IV dataset using an empirical cumulative distribution function (eCDF) for numeric data transformation. First the eCDF is computed for each lab test and form a distribution. Then each lab value is mappe to a probability between 0 and 1. For example, If a patient's creatinine level is at the 70th percentile of the training data, it is mapped to 0.7. The authors trained various configurations of masked language models (MLMs) and applied them to four clinical cases. However, during inference, the MLM models did not outperform

random forest and XGBoost models. The authors concluded that MLM pretraining on MIMIC-IV does not transfer well to downstream inference tasks[13].

Beyond the HVAT, xVal and Labrador methods, some potential modifications to existing approaches could improve numeric data integration in transformers. One promising method involves using graded semantic vectors to generate value embeddings[132, 17]. Graded semantic vectors are constructed by interpolating between a pair of orthogonal demarcator vectors. The maximal ω and minimal α values of the demarcator vectors define the boundaries, while the distance between these nearly orthogonal vectors is projected into the high dimensional vector space using proximity metrics. A vector between α and ω represents a feature proportionate to its distance from the minimum to the maximum in vector space. This algorithm has demonstrated reasonable effectiveness on enconding immunoglobulin sequences of west Nile virus data[132, 17] and shows potential for preserving numeric information.

A similar method called rotary positional embeddings (RoPE) was used in the recent open-source generative transformer model, Llama II[118]. Unlike traditional fixed positional encodings, RoPE introduces a continuous rotation matrix that transforms token embeddings based on their relative positions[111]. This method efficiently handles varying sequence lengths while encoding the relative positional relationships between tokens. As an extension, RoPE could be adapted to encode numerical values, such as drug doses and lab values, within a transformer model. By converting these values into sine and cosine functions, followed by applying a dot product with a weight matrix learned during training, rotary embeddings can capture the continuous nature of numeric data. This could enhances the model's ability to process sequences with embedded quantitative information.

In this paper, I propose new methods to incorporate value-aware embeddings representing lab and drug dose data into generative pre-trained transformers, focusing primarily on GPT-2 and using EHR data exclusively. While later models like GPT-3.5 or GPT-4 offer superior performance, they require significantly more computational resources and are not publicly available. The GPT-2 model, on the other hand, provides a robust and efficient solution that aligns with our research

goals, enabling effective and flexible implementation. The goals of this paper are twofold: 1) to encode numeric clinical information into embeddings for downstream ADR prediction tasks and 2) to compare different types of value-aware embeddings. Our main hypothesis is that value-aware algorithms, combined with generative pre-trained models, can outperform non-value-aware methods, including both temporal and non-temporal disproportionality metrics.

4.3 Method

4.3.1 MIMIC-IV Hospitalization and UW EHR

The details of the MIMIC-IV hospitalization and UW EHR datasets are described in Chapter 3. While the MIMIC-IV dataset does not include unstructured clinical notes, the UW EHR dataset does. For the MIMIC-IV dataset, I utilized diagnostic codes, laboratory test results, and medication administration records. Laboratory results were preprocessed to retain relevant numeric values, which were categorized as "high", "low," or "normal" based on standard reference ranges in the two EHR datasets. The values were then normalized using Min-Max scaling to ensure consistency across different measurement units and I did categorized and numeric value for creatinine. Similarly, medication records were filtered to include administered doses, with dose values normalized to a range between 0 and 1.

I represented each hospitalization as a chronological sequence of events, where each event corresponds to a diagnosis, lab test, medication, or other clinical entry. Events were sorted by their occurrence time and assigned an event type (e.g., diagnosis, lab or drug for concepts extracted from unstructured text). Each event was then mapped to a discrete token in our vocabulary.

When an event included a numeric measurement- for example, a lab result (e.g., creatinine = 2.1 mg/dL) or a medication dose (e.g., furosemide 40 mg), I stored that value in a corresponding numeric field. For tokens without numeric values, the lab and dose fields are (0,0). Thus, each token can optionally have an additional numeric value for labs or medication doses. These numeric fields were preserved in our serialized PyTorch datasets to be used by the model alongside the tokens themselves (Figure 4.1).

For the UW EHR sequences, the data structure was similar to MIMIC-IV, except that UW data

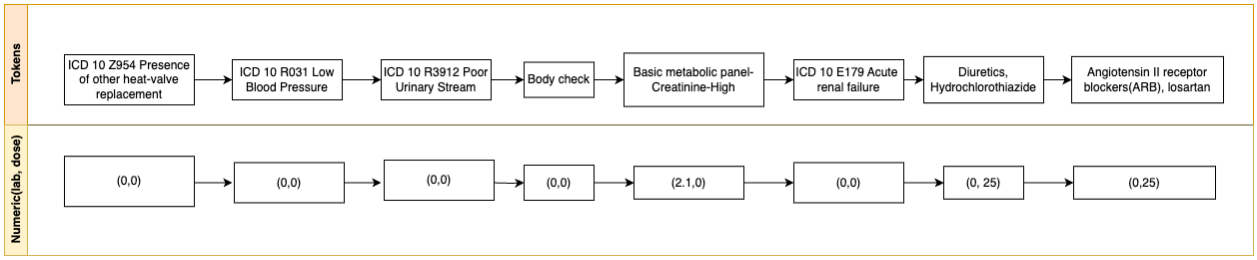

Figure 4.1: An example EHR sequence with numeric data, the values showed on the plots are values before min max scaling.

also included tokens derived from clinical text using MetaMap [4].

4.3.2 Representations of value aware embeddings

Value-aware algorithm

Inspired by the HVAT and xVAL algorithms[108, 36], I developed a customized GPT-2 model to improve clinical concept representations by incorporating numeric values into embeddings. In the first configuration, I extended the longitudinal data processing concept from HVAT by introducing lab and dose values associated with medical concepts and scaling these values using value embeddings. Numeric embeddings (per clinical concept with associated numeric values) were obtained from separate embedding layers and scaled via element-wise multiplication with corresponding value embeddings. This enriched the embeddings with both conceptual and numeric information.

Unlike HVAT, which uses a hybrid structure with separate branches for longitudinal and discrete data, our method focuses on the longitudinal aspect. I integrated these value-aware embeddings directly into the GPT-2 framework, combining(addition) them with word token embeddings matrix(WTE) and positional embeddings (PE). The Transformer architecture uses PE to incorporate order information and token embedding matrix to represent the semantic meaning of each token. These embeddings are typically stored in lookup tables (embedding matrices), making them easily retrievable[101]. Unlike HVAT's sinusoidal PE, our PE are learned during training.

In the second configuration, we replaced the token embedding matrix with pretrained embeddings derived in the previous chapter while preserving the integration of numeric values.

In general, the combined embedding for each medical concept C_i at position i is calculated as:

$$\mathbf{E}_{\text{combined}} = \mathbf{E}_{\text{token}}(C_i) + \mathbf{E}_{\text{position}}(i) + \text{value} \times \mathbf{E}_{\text{value}}(C_i)$$

where $\mathbf{E}$ represents embeddings, C_i is the medical concept, and i is the position of the token, value is the numeric value. The $\mathbf{E}_{\text{token}}(C_i)$ is either the randomly initialized token embedding matrix or the pretrained embeddings derived using the Semantic Vectors(semvecpy) package described previously.

Graded Embeddings

Universal Graded Embeddings: I used the graded semantic vectors, also known as demarcator vectors to encode the numeric data[132, 17]. I generate a pair of normalized demarcator vectors $D(\alpha)$ and $D(\omega)$ that serve as universal "anchors" for all concepts. These vectors are generated once, and their interpolation is concept-dependent through min-max normalization of the concept's range. The alpha and omega vectors are universal across all medical concepts. I used randomly generated, normalized vectors $D(\alpha)$ and $D(\omega)$, which in high-dimensional spaces tend to be nearly orthogonal. These graded vectors are then directly added to the token embedding matrix and PE(bundling) of our customized GPT-2 model.

For a given value v, the interpolated vector is calculated as: I encode each numeric value v (e.g., a lab result) between a global minimum and maximum (min and max) by linearly interpolating between two near-orthogonal boundary vectors, D_α and D_ω. Specifically,

$$\mathbf{E}_{\text{graded}}(C_i, L_i, D_i) = \left(\frac{\max - v}{\max - \min}\right) D_\alpha + \left(\frac{v - \min}{\max - \min}\right) D_\omega, \tag{4.1}$$

where min and max represent the numeric bounds for the concept (e.g., 0.5 and 5.0 mg/dL for creatinine). I define two interpolation weights:

$$x = \frac{\max - v}{\max - \min}, \quad y = \frac{v - \min}{\max - \min},$$

so that

$$\mathbf{E}_{\text{graded}}(C_i, L_i, D_i) = x\, D_\alpha \;+\; y\, D_\omega.$$

Hence, if v is close to min, the resulting vector is near D_α; if v is close to max, the vector is near D_ω. Figure 4.2 illustrates how a creatinine value is mapped into this graded vector space.

The combined embedding for a concept C_i at position i with L_i (lab value) and D_i (dose value) is given by:

$$\mathbf{E}_{\text{combined}} = \mathbf{E}_{\text{token}}(C_i) + \mathbf{E}_{\text{position}}(i) + \mathbf{E}_{\text{graded}}(C_i, L_i, D_i)$$

where $\mathbf{E}_{\text{graded}}(C_i, L_i, D_i)$ is the graded embedding for concept C_i with L_i or D_i values.

Concept-specific Graded Embeddings: I also explored the concept-specific demarcator vectors $D(\alpha)^{(c)}$ and $D(\omega)^{(c)}$, both are randomly initialize at the beginning. For each medical concept C, I determine the minimum and maximum values observed in the data and then compute the graded vector by linearly interpolating between $D(\alpha)^{(c)}$ and $D(\omega)^{(c)}$. In this setup, besides the concept specific min max normalization, the alpha and omega are not universal across all medical concepts, but concept-specific. The resulting concept-specific graded vectors are combined with the token and positional embeddings in the same manner by addition, thereby providing a more tailored representation of numeric values.

Trainable Graded Embeddings: In this variation, rather than loading fixed graded embeddings, I make the demarcator vectors trainable parameters that can be updated during model training. For each medical concept C, I initialized $D(\alpha)^{(c)}$ and $D(\omega)^{(c)}$ as trainable parameters. In this setup, $D(\alpha)^{(c)}$ and $D(\omega)^{(c)}$ are no longer fixed but are optimized jointly with the language model's parameters through back-propagation. Theoretically, the resulting embeddings evolve to better capture the semantics of numeric values. Then I incorporate it into the models' input representation by bundling it to the WTE(word token embedding matrix) and PE.

Rotary Embeddings

Rotary Position Embeddings (RoPE) can be adapted to encode numeric values by replacing the usual position index with numeric inputs. Instead of merely encoding a token's position in the se-

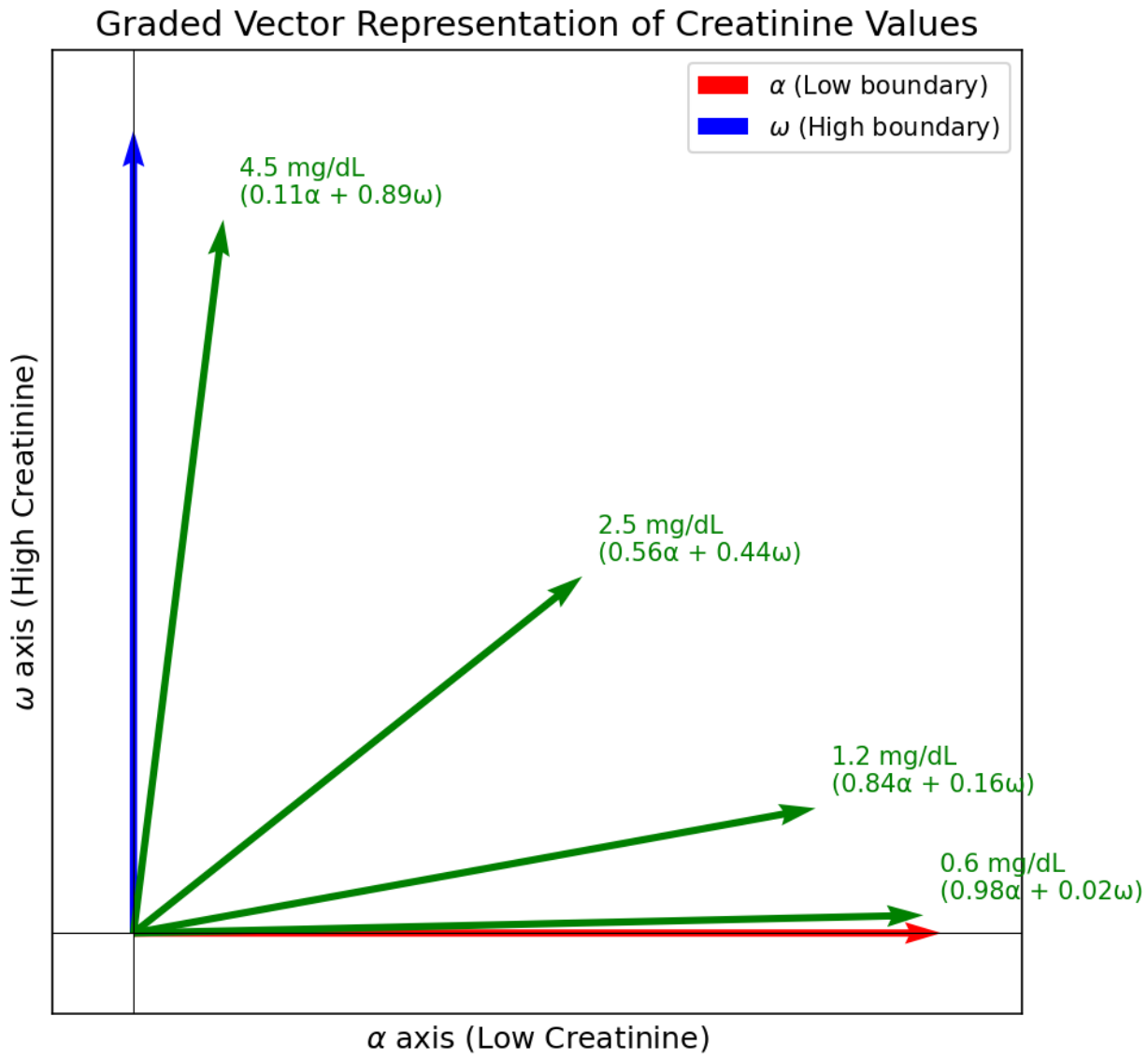

Figure 4.2: Graded vector representation of a creatinine value, with min = 0.5 mg/dL and max = 5.0 mg/dL.

quence, I treat the lab and/or dose value as the "position" for rotation, thereby injecting continuous numeric information into the embedding space(Figure. 4.3).

Angle Computation Let d be the embedding dimension. As usual with RoPE, I split the d-dimensional vector into $\frac{d}{2}$ coordinate pairs $\big(2i,\, 2i+1\big)$ for $i = 0, \ldots, \frac{d}{2} - 1$. Given a numeric value v, I compute angles

$$\theta_i \;=\; \frac{v}{10000^{\,\frac{i}{(d/2)}}} \;\times\; \text{scale},$$

where scale is a learnable or fixed factor that controls how sharply the numeric value (v) influences the rotation. I then define a rotation operator $R\big(\theta(v)\big)$ that acts on each coordinate pair:

$$\begin{pmatrix} x'_{2i} \\ x'_{2i+1} \end{pmatrix} = \begin{pmatrix} \cos\big(\theta_i\big) & -\sin\big(\theta_i\big) \\ \sin\big(\theta_i\big) & \cos\big(\theta_i\big) \end{pmatrix} \begin{pmatrix} x_{2i} \\ x_{2i+1} \end{pmatrix},$$

thus "injecting" the numeric value v into the embedding by rotating pairs of coordinates in proportion to v.

Complex-Plane View A helpful way to see this is through complex numbers. For each index i, consider the pair $\big(x_{2i},\, x_{2i+1}\big)$ as the real and imaginary parts of a complex number

$$z_i \;=\; x_{2i} \;+\; i\, x_{2i+1}.$$

Rotating $\big(x_{2i},\, x_{2i+1}\big)$ by θ_i is then equivalent to multiplying z_i by $e^{\,i\,\theta_i}$, i.e.,

$$z'_i \;=\; e^{\,i\,\theta_i}\, z_i \;=\; \big(x_{2i}\cos\theta_i - x_{2i+1}\sin\theta_i\big) \;+\; i\,\big(x_{2i}\sin\theta_i + x_{2i+1}\cos\theta_i\big).$$

This complex-plane perspective is mathematically identical to the 2D rotation in real coordinates and clarifies how RoPE encodes continuous values via rotation.

Universal Variant In the universal case, the model learns a small set of trainable base vectors that are shared across all concepts. Specifically, I initialize two trainable vectors, $\mathbf{b}^{(\text{lab})}$ and $\mathbf{b}^{(\text{dose})}$, to represent lab and dose concepts, respectively. Let v_L and v_D be the numeric values for lab and dose; often only one is nonzero (e.g., $(2.1, 0)$ for a lab, or $(0, 25)$ for a medication dose, Figure 4.1), but I allow both to be nonzero in theory. I compute

$$R\big(\theta(v_L)\big)\,\mathbf{b}^{(\text{lab})} \quad \text{and} \quad R\big(\theta(v_D)\big)\,\mathbf{b}^{(\text{dose})},$$

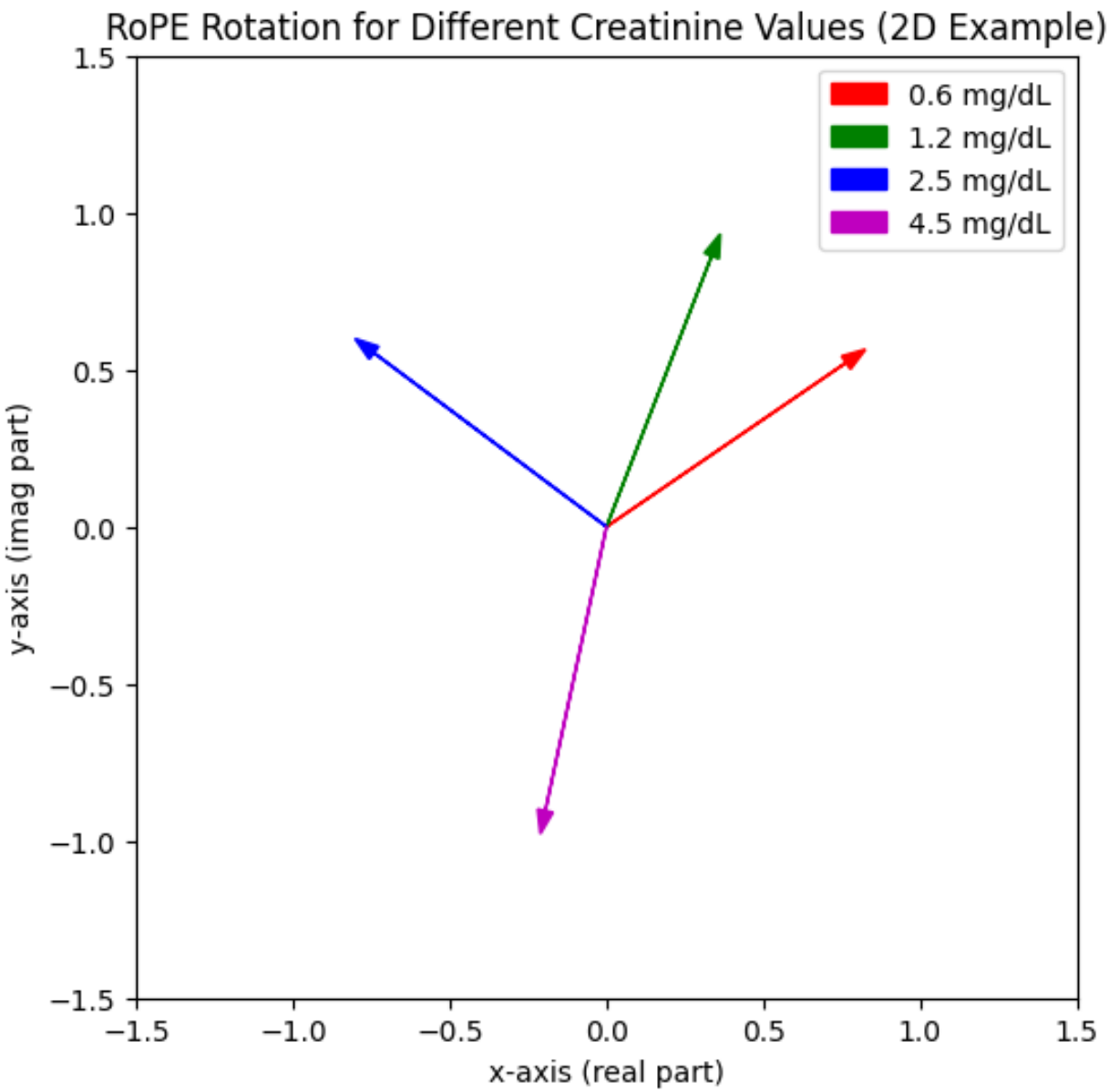

Figure 4.3: Rotary representation of a creatinine value in 2D plane.

then add them:

$$R\big(\theta(v_L)\big)\,\mathbf{b}^{(\text{lab})} \;+\; R\big(\theta(v_D)\big)\,\mathbf{b}^{(\text{dose})}.$$

Finally, I apply a linear transformation $\mathbf{W}$:

$$\mathbf{E}_{\text{univ}}(v_L, v_D) = \mathbf{W}\left(R\big(\theta(v_L)\big)\,\mathbf{b}^{(\text{lab})} \;+\; R\big(\theta(v_D)\big)\,\mathbf{b}^{(\text{dose})}\right).$$

I implement $\mathbf{W}$ as an nn.Linear(d, d) layer (or equivalently, a matrix in $R^{d\times d}$). It is trained jointly with the model so that after rotation, the resulting vector can be further adapted to the downstream task. If omitted (i.e., $\mathbf{W}$ is the identity), I simply have the sum of the rotated base vectors. This single $\mathbf{W}$ is shared for all lab/dose events and does not vary by concept.

Because lab and dose don't co-occur in the same medical concept(how the data was prepared), either $v_L \neq 0$ or $v_D \neq 0$. I retain the additive form to keep the formulation general.

Concept-Specific Variant. In contrast, the *concept-specific* variant allocates separate base parameters for *each* concept c. That is, I created two learnable parameters $\mathbf{b}^{(c,\text{lab})}$ and $\mathbf{b}^{(c,\text{dose})}$, allowing every concept to have its own baseline representation. Given v_L and v_D, I apply the same RoPE rotation to produce

$$R\big(\theta(v_L)\big)\,\mathbf{b}^{(c,\text{lab})} \quad \text{and} \quad R\big(\theta(v_D)\big)\,\mathbf{b}^{(c,\text{dose})}$$

which are then optionally combined and transformed:

$$\mathbf{E}_{\text{cs}}(c, v_L, v_D) = \mathbf{W}\left(R\big(\theta(v_L)\big)\,\mathbf{b}^{(c,\text{lab})} \;+\; R\big(\theta(v_D)\big)\,\mathbf{b}^{(c,\text{dose})}\right).$$

Concept-specific parameters enable each concept to capture nuanced numeric relationships, at the cost of additional memory.

Integration with GPT-2. Regardless of the universal or concept-specific variant, I add the numeric rotary embedding to GPT-2's usual token embedding $\mathbf{E}_{\text{token}}$ and positional embedding $\mathbf{E}_{\text{position}}$. Let c_i be the i-th token's *concept* (e.g., a lab test or medication), and let $v_i^{(L)}$ and $v_i^{(D)}$ denote its associated lab or dose values. In this setup, these two numeric fields are mutually exclusive for a single token, meaning one of them is always zero. I retain both parameters, however,

so that the model can uniformly process all tokens regardless of whether they represent labs or drugs. Formally, the combined embedding is:

$$\mathbf{E}_{\text{combined}}(c_i, v_i^{(L)}, v_i^{(D)}, i) \;=\; \mathbf{E}_{\text{token}}(c_i) \;+\; \mathbf{E}_{\text{position}}(i) \;+\; \mathbf{E}_{\text{rotary}}(c_i, v_i^{(L)}, v_i^{(D)}),$$

where

$$\mathbf{E}_{\text{rotary}}(c_i, v_i^{(L)}, v_i^{(D)}) = \begin{cases} \mathbf{E}_{\text{univ}}(v_i^{(L)}, v_i^{(D)}) & \text{(universal variant)}, \\ \mathbf{E}_{\text{cs}}(c_i, v_i^{(L)}, v_i^{(D)}) & \text{(concept-specific variant)}. \end{cases}$$

If c_i corresponds to a lab concept, then $\big(v_i^{(L)} \neq 0,\, v_i^{(D)} = 0\big)$; if c_i is a drug concept, I have $\big(v_i^{(L)} = 0,\, v_i^{(D)} \neq 0\big)$. The $\mathbf{E}_{\text{combined}}$ representation is then passed into the self-attention layers, enabling GPT-2 tto process continuous numeric information (lab or dose values) alongside discrete tokens in a unified embedding space.

4.3.3 Value-aware GPT2 small

Four different configurations and variations within each model were examined:

- WTE + PE + Value-Aware Embeddings(VAE): Learned word token embeddings combined with positional and value-aware embeddings.
- Pretrained Semantic Embeddings + PE + VAE: Pretrained embeddings replaced WTE, integrated with positional and value-aware embeddings.
- WTE + PE + Graded Embeddings: Graded embeddings in universal variant
- WTE + PE + Rotary Embeddings: Rotary embeddings in universal variant

In the second part of the analysis, I compared specifically compared different configurations of rotary and graded embeddings.

- WTE + PE + Graded Embeddings: in universal variant
- WTE + PE + Graded Embeddings, concept specific, fixed

- WTE + PE + Graded Embeddings: concept specific with all trainable parameters
- WTE + PE + Rotary Embeddings: in universal variant
- WTE + PE + Rotary Embeddings: concept specific with all trainable parameters

Training used a learning rate of 1e-6, with early stopping applied if validation loss did not decrease by more than 10% across two consecutive epochs. Optimization was performed using the Adam optimizer with a weight decay of 0.01, and the maximum number of epochs was set to 20.

4.3.4 Evaluation

During model training, I tracked perplexity across epochs to monitor their progressive reduction over time for the validation and the training sets. For model inference, I evaluated the performance of each approach in detecting ADR signals by assessing its ability to prioritize positively labeled drug-ADR pairs over negatively labeled ones, using the Area Under the Curve of the Receiver Operating Characteristic (AUROC) across four reference sets. ADRs were mapped to corresponding International Classification of Diseases (ICD-9, 10) codes, and a weighted average(based on the frequency) was computed for all ICD codes associated with each ADR. For example, acute renal failure is linked to eight distinct ICD codes across both ICD versions. AUROC values were calculated for each reference set. For the overall AUROC, all drug-ADR pairs were included in the calculation rather than averaging across the reference sets. The reference sets were described explicitly in the chapter 3.

The evaluation is mostly the same as chapter 3(also added the max probability). To reiterate, I assessed how likely the model is to predict a specific drug token given an ADR from the reference set, effectively using the ADR as a prompt within GPT-2(trained on the sequence in reverse order) and checking the probability that the previous token is the relevant drug. I excluded any ADR or drug with fewer than 10 occurrences to maintain a balanced evaluation. Both sigmoid and softmax probabilities were calculated: the softmax indicates the drug token's likelihood relative to all other tokens, whereas the sigmoid provides an independent probability for each potential drug.

To incorporate ICD codes, I applied two methods. In the Weighted Probabilities approach, I multiplied each code's probability by its corresponding weight and summed the results. For instance,

if an ADR is mapped to multiple ICD codes, each code's probability is combined proportionally to its frequency-based weight. Formally:

$$\text{Weighted Sum } P(\text{drug} \mid \text{ADR})_{j,c} = \sum_{i \in \mathcal{I}_c} \Big[P\big(\text{drug}_j \mid \text{ADR}_c\big)_i \times \text{ICDWeight}_i \Big].$$

where i represents the ICD codes associated with the ADR, j represents the drug, and c represents the condition.

Alternatively, in the Max Probabilities method, I selected the ICD code that yielded the highest probability:

$$\text{MaxProb}_{j,c} = \max_{i \in \mathcal{I}_c} \Big[P\big(\text{drug}_j \mid \text{ADR}_c\big)_i \Big].$$

Both techniques were evaluated under softmax and sigmoid regimes (Weighted Softmax/Sigmoid and Max Softmax/Sigmoid), and I report the highest area under the ROC curve (AUROC) among these four metrics. The AUROC was calculated by ADR, by reference set, and in aggregate.

4.4 Results

For the value-aware models, I used the best-performing model from the non-value-aware stage as the baseline(Table. 3.2. All models incorporating value-aware embeddings outperformed the baseline model (Table 4.1), except for the UW-UMN reference set. The addition of lab values and drug dose information improved the model's ability to distinguish between ADR+ and ADR- labels. When the word token embedding matrix was replaced with pretrained embeddings, performance improved even further. The best-performing models combined pretrained embeddings, with value-aware embeddings to represent numeric values for the MIMIC data, and for the UW data, graded embeddings had the best performance.

A detailed analysis showed that the method had strong signals on detecting acute renal failure and gastrointestinal hemorrhage related ADR across all configurations. In models using graded embeddings, the AUROC reached 0.69 for acute renal failure and 0.8 for gastrointestinal bleeding for the MIMIC set. For the UW set, the AUROC reached 0.74 for acute renal failure and gastrointestinal bleeding, it was 0.84. Both value-aware and non-value-aware methods demonstrated consistent results across the two EHR datasets.

Table 4.1: Comparison of models' AUROC on MIMIC and UW inpatient data, freezed, with "+" binding, trained for 5 epochs due to the earlystopping rule.

MIMIC inpatient summary

	best non-value aware	VAT	embeddings + VAT	Graded Universal	Rotary Universal
OMOP + EU (82+, 101-)	0.521	0.600	0.615	0.622	**0.655**
UMN (58+, 16-)	0.533	0.528	0.563	**0.651**	0.639
AZCERT (33+, 215-)	0.630	0.643	**0.662**	0.655	0.581
Across all sets (173+, 332-)	0.583	0.676	**0.692**	0.688	0.666

UW inpatient summary

	best non-value aware	VAT	embeddings + VAT	Graded Universal	Rotary Universal
OMOP + EU (74+, 86-)	0.640	0.640	0.639	**0.650**	0.591
UMN (49+, 13-)	**0.652**	0.557	0.610	0.630	0.561
AZCERT (32+, 211-)	0.658	0.657	**0.671**	0.640	0.613
Across all sets (155+, 340-)	0.655	0.695	0.697	**0.710**	0.659

I further compared the performance of graded and rotary embedding configurations(Table 4.2). For both the MIMIC and UW inpatient datasets, the best-performing models were those utilizing graded embeddings with universal parameters. Models with universal parameters consistently outperformed their concept-specific counterparts.

4.5 Discussion

In this study, I proposed a novel method for ADR signal detection by customizing the GPT-2 architecture for EHR data from two large healthcare institutes. I examined if the value-aware transformers can outperform the best performing models from Aim 1. I hypothesized that models with value-aware embeddings would outperform the non-value-aware models. In our experiments, all models utilizing the GPT-2 architecture with value embeddings outperformed the models with non-value aware embeddings(except for the UW-UMN reference set, TreeScan still performed the

Table 4.2: Comparison of value-aware models with graded or rotary embeddings on MIMIC and UW inpatient data, train at least 5 epochs, with 10% earlystop validation loss

MIMIC inpatient summary

	Graded Universal, fixed	Graded Concept Specific, fixed	Graded Concept Specific, trainable	Rotary Universal, Trainable	Rotary Concept Specific, Trainable
OMOP + EU (82+, 101-)	0.622	0.634	0.613	0.655	0.580
UMN (58+, 16-)	0.651	0.587	0.601	0.639	0.509
AZCERT (33+, 215-)	0.655	0.649	0.574	0.581	0.625
Across all sets (173+, 332-)	**0.688**	0.684	0.636	0.666	0.653

UW inpatient summary

	Graded Universal, fixed	Graded Concept Specific, fixed	Graded Concept Specific, trainable	Rotary Universal, trainable	Rotary Concept Specific, trainable
OMOP + EU (74+, 86-)	0.650	0.650	0.665	0.607	0.659
UMN (49+, 13-)	0.630	0.589	0.583	0.536	0.615
AZCERT (32+, 211-)	0.640	0.647	0.647	0.613	0.634
Across all sets (155+, 340-)	**0.710**	0.702	0.702	0.657	0.702

best). Simply adding scaled value embeddings alongside token embedding matrix and PE from the GPT-2 model enhanced the model's performance in distinguishing adverse drug events. Additionally, numeric values were integrated using various embeddings, including graded embeddings, rotary embeddings, and other configurations.

GPT models are known for their versatility and wide application across many fields. To our knowledge, this is the first attempt to use a GPT model specifically for enhancing ADR signal detection using EHR data alone. While larger language models, such as Llama, exist, GPT-2 remains accessible, hands-on, and less computationally intensive. By integrating value-aware methods, the performance improved significantly.

To date, very few studies have attempted to represent numeric data within transformers. HVAT, xVAL, and Labrador are three relevant studies[108, 13, 36]. The xVAL approach utilizes a method

similar to HVAT involving scaling embeddings. I adapted the HVAT framework because it aligned with the disease-related content of our study. However, our approach differs in several ways. HVAT employed sinusoidal positional embeddings from the BERT model and had customized methods for categorical data such as gender and ethnicity. They reported an AUROC of 0.96 on ADRD and physical fitness data from the VA[108]. Unfortunately, I could not reproduce their results due to data accessibility limitations. In our analysis, I used trainable positional embeddings based on sequences, where each unit represents a hospitalization. Additionally, I prioritized overall model performance rather than focusing on specific associations. Lastly, our method excluded gender and age, which are important risk factors. Interestingly, in the HVAT study, excluding gender and age only slightly reduced performance, suggesting that the value-aware components had the greatest influence[108].

Our best-performing models achieved an AUROC between 0.7 and 0.8 overall, with specific ADRs such as gastric bleeding related prediction exceeding 0.8, compared to the non-value aware setting, 0.77. These results were consistent across the two EHR datasets. The only discrepancy I observed was that the best non-value aware method outperformed the value aware transformer(VAT) on the UMN reference set (Table 4.1). The UMN set is much smaller compared to the other two and always has fluctuations. Further analysis, for example, on different seeds, might introduce higher stability into the analysis.

I utilized graded and rotary embeddings to represent numeric values, which is innovative given transformers' well-known limitations with arithmetic tasks[129]. Graded embeddings have been shown to increase the similarity between immunoglobulin receptor sequences produced in response to West Nile Virus[132, 17]. Rotary embeddings, which use a rotation matrix to encode absolute distances while capturing relative positional dependencies, performed poorer to graded embeddings. Universal variants of these embeddings performed better than concept-specific trainable variants. The added complexity of concept-specific embeddings appeared to not adding additional values when handling highly variable clinical data. Theoretically, trainable embeddings should outperform fixed versions due to their flexibility, but this was not observed in our experiments. Further analysis of this discrepancy is presented in the next chapter.

I also applied the Labrador method[13], transforming numeric values using the representation of eCDF algorithm. This nonparametric approach maps numeric values to their percentile ranks, offering advantages such as handling outliers and standardizing distributions. However, the results did not outperform non-value-aware models, consistent with findings from the original Labrador study, where transformer models performed poorly on downstream classification tasks[13]. I also experimented with Z-score normalization but found it less effective than Min-Max scaling, which was used consistently across all experiments.

In addition to the reported configurations, I explored other setups. For example, in our previous work on the aer2vec method[99], I tested both $P(\text{drug} \mid \text{ADR})$ and $P(\text{ADR} \mid \text{drug})$. In this paper, I focused on $P(\text{drug} \mid \text{ADR})$ which had more consistent and reliable performance than the forward model, which is $P(\text{ADR} \mid \text{drug})$. This observation aligns with the previous aer2vec work[99].

There are a few limitations of the study. First of all, EHR data inherently has drawbacks, for instance, missingness and the ICD codes are originally used for billing purposes, which might not reflect the true diagnoses[44]. The UW EHR data were prepared originally to perform a different ADR-drug analysis. The initial event had a selection criterion of one of the 232 drugs or one of the 10 ADRs in the EU+Ryan's reference set. The UW EHR data has a wide time span and I mapped UMLS CUI from the clinical notes, which may introduce a large controlled vocabulary. I compared the distribution of the pretrained embeddings of the norm length between the two sets. The distributions were quite similar to each other, but in the UW set, there are a small amount of tokens that are more influential(large norms) than the others (Appendix Figure A.1), which I did not observe in the MIMIC set. Both UW data and MIMIC data have timestamps on events, but they might not reflect the time series accurately. The UW data uses crude timestamps for the events, which do not provide the exact time of occurrence. Secondly, these data constraints might effect generalizability. I only tested the method on data from two EHRs from large healthcare system. Further analysis is needed to confirm the generalizability of the methods. Thirdly, the overall performance is promising, but the breakdown on individual ADR leaves room for further improvement. Mostly, the performance is consistent within the reference set and across different

EHRs. But there are still circumstances, for example, on individual condition such as anaphylactic shock, in which the predictability on binary ADR-drug event is random.

For future directions, I will attempt to address these limitations by acquiring more datasets, such as adding MIMIC clinical notes into the analysis. The next generation of generative models could further improve the performance. For instance, new numeric representations have been introduced lately, for instance, McLeish et al. developed a new type of positional embeddings to encode the significance of digits. It showed promising ability for addition, multiplication, and sorting tasks[80]. One potential use is to incorporate it into the transformer model I developed and it has the potential to further improve the performance. Another direction I considered is to include multimodal data. For example, using EHRs solely can capture comprehensive view for individual patients, but having spontaneous reporting system data, like FDA Adverse Event Reporting System (FAERS) has the potential to further improve performance. With rapid change in the field, more advanced numeric representation will likely be developed and could be used to further improve the performance of the methods(similar to chapter 3).

4.6 Conclusion

In this chapter, I further developed the GPT-2 based pharmacovigilance method by integrating value-aware embeddings. All models with value-aware embeddings outperformed the non-value aware methods(except for UW-UMN). The same observation was consistent across the two EHRs. The best performing model in the value aware stage is the pretrained embeddings + value aware embeddings for the MIMIC data, while for the UW, the graded universal variant had the best performance. Further work is needed to examine these generalizability of the methods. In the next chapter, I will analyze the roles of graded and rotary embeddings, and will explore different bindings.

Chapter 5

BUNDLING AND BINDING

5.1 Abstract

In this chapter, I investigate whether explicit variable-value binding (via the Hadamard product) provides a clear advantage over simple addition-based "bundling" for encoding numeric information in a GPT-2–based architecture. I also compare universal and concept-specific embeddings, including configurations that store and train alpha–omega vectors for direct numeric recovery. Using MIMIC and UW inpatient datasets, I evaluate each approach on perplexity, AUROC, and—where applicable—mean squared error (MSE) for evaluation. Overall, these findings suggest that a straightforward additive (bundling) strategy with universal embeddings is sufficient for numeric encoding in high-dimensional transformer models, indicating that explicit binding operations may not be strictly necessary.

5.2 Introduction

In Chapter 4, I developed GPT-2–based value aware models for encoding numeric information (e.g., lab results, drug dosages) alongside token and positional embeddings. My approach was to add the numeric embeddings with the word token embeddings (WTE) and positional embeddings (PE). In the terminology of *Vector Symbolic Architectures* (VSAs), this summation is known as bundling or superposition, that is, combining multiple vectors by addition to form a single composite vector [96, 94, 56].

In contrast, binding refers to merging vectors using an invertible operation, such as the Hadamard (element-wise) product, circular convolution, pairwise exclusive or XOR[37]. The Hadamard product performs element-wise multiplication, preserving dimensionality, and is often used in VSAs as a dimension-preserving alternative to the outer product. Circular convolution provides a computationally convenient alternative to Smolensky's tensor product, a foundational operation in tensor-based symbolic models. To develop graded vectors, the authors applied circular convolution

to encode continuous values, and bind this information to variables of interest. For example, the major events during presidency were bound to the name of the president [132, 17]. Other binding methods also exist, including some with self-inverse operations [106, 37]. Binding typically preserves the ability to "unbind" or decode original vectors from a combined representation [94, 56].

Although both bundling and binding can integrate information in high-dimensional spaces, they differ in their algebraic properties and computational requirements. A recent study reviewed 11 VSA implementations, highlighting their varying strengths in dealing with fuzziness and ambiguity [106]. In general, bundling (addition) is simpler and cheaper but partially loses invertibility, whereas binding (element-wise multiplication, circular convolution, etc.) preserves the potential to separate components later [56].

In this review[106], it also summarizes several high-dimensional computing or vector symbolic architecture methods support the encoding of continuous variables by projecting them into real or complex vector spaces (Table 5.1). For instance, MAP-C initializes and operates on real-valued hypervectors in R^D with element-wise multiplication as the binding operation [33]. HRR (Holographic Reduced Representation) employs circular convolution on real-valued hypervectors, allowing scalar information to be superimposed and retrieved through correlation [98]. Similarly, vector-derived binding (VTB)[37] and matrix-based binding (MBAT) use transformations (permutation or matrix multiplication) on real vectors to represent continuous features [32]. In contrast, FHRR (Fourier Holographic Reduced Representation) represents continuous data through complex vectors, binding them via element-wise complex multiplication that corresponds to phase addition and subtraction [95]. These real or complex vector spaces allow for smooth similarity metrics, such as cosine or angular distance, and permit natural encoding of graded information.

Some researchers have attempted to understand how language models store and track the evolving states of entities[31, 55, 82]. The major findings are that transfomer models implicitly use a binding mechanism to associate entities with attributes [31]. One study pointed out that specific neurons in the feedforward neural network(MLP layer) play more important roles than the attention mechanism for storing and retrieving factual knowledge, and the researchers were able to directly edit the associations between entities and attributes via a causal intervention technique[82]. However, there are still many unknowns. Kim and Schuster proposed that the ability to track entity states emerges in the larger, more advanced transformer models, like GPT-3.5. Older models

like GPT-3 and vanilla models(Language models without fine-tuning) failed to track entity state changes effectively [55].

In this study, I explore whether using an explicit binding operator (such as the Hadamard product) offers a clear advantage over simple addition for representing numeric data in a customized GPT-2 model. Our goal is to understand how binding operations affect performance, and to investigate why concept-specific variants may not necessarily outperform universal configurations.

5.3 Method

The method consists of two main parts. First, I extend the approach from Chapter 4 by comparing addition-based bundling to an element-wise (Hadamard) product for binding numeric embeddings. Second, I examine why concept-specific variants do not necessarily outperform universal variants(was our findings in the previous chapter), focusing on how well numeric values are preserved in each transformer layer. In other words, I ask: is explicit binding truly necessary for value-aware GPT-2 models?

5.3.1 Part I: Bundling vs. Hadamard Binding

Previously, I formed combined embeddings by *adding* the word token embedding (WTE), positional embedding (PE), and a numeric embedding. Specifically,

$$\mathbf{E}_{\text{combined}}(C_i, L_i, D_i) = \big(\mathbf{E}_{\text{token}}(C_i) + \mathbf{E}_{\text{position}}(i)\big) \; + \; \mathbf{E}_{\text{rotary/graded}}(C_i, L_i, D_i).$$

While this addition effectively "bundles" the components, it does not explicitly maintain distinct roles for each. In this study, I compare the additive approach to a *Hadamard product* variant, where

$$\mathbf{E}_{\text{combined}}(C_i, L_i, D_i) = \big(\mathbf{E}_{\text{token}}(C_i) + \mathbf{E}_{\text{position}}(i)\big) \; * \; \mathbf{E}_{\text{rotary/graded}}(C_i, L_i, D_i),$$

and $*$ denotes element-wise multiplication. The goal is to see whether this more explicit form of binding (rather than simple bundling) better encodes numeric information in the transformer.

5.3.2 Part II: Universal vs. Concept-Specific Embeddings

Next, I examine why concept-specific numerical embeddings and models with all trainable parameters (including $\mathbf{D}_\alpha{}^{(c)}$ and $\mathbf{D}_\omega{}^{(c)}$) might not surpass universal numerical embeddings. Following

Table 5.1: Methods Supporting Continuous Encoding in High-dimensional Vector Spaces. Table was partly adapted from Schlegel et al., 2022 [106].

Method	**Vector Space (Elements)**	**Binding Operation**	**Bundling Operation**	**Continuous Encoding**
MAP-C (Multiply-Add-Permute - Continuous)	Real (R^D)	Element-wise multiplication	Element-wise addition with cutting	Yes, real valued hypervectors
HRR (Holographic Reduced Representation)	Real (R^D)	Circular convolution	Element-wise addition with normalization	Yes, real valued vectors
VTB (Vector-Derived Binding)	Real (R^D)	Learned/permuting transform	Element-wise addition with normalization	Yes, real valued vectors
MBAT (Matrix Binding of Additive Terms)	Real (R^D)	Matrix multiplication	Element-wise addition with normalization	Yes, real valued vectors
FHRR (Fourier Holographic Reduced Rep.)	Complex (C^D)	Element-wise complex multiplication	Superposition via element-wise phase averaging	Yes, continuous phase values

Chapter 3, I define two high-dimensional "demarcator" vectors, $\mathbf{D}_\alpha$ and $\mathbf{D}_\omega$, that map numeric values to high-dimensional vectors. In the *universal* variant, one shared $\mathbf{D}_\alpha$ and $\mathbf{D}_\omega$ handle all numeric concepts; in the *concept-specific* variant, each numeric feature has its own pair of demarcator vectors. In the fully trainable variant, every parameter (including ${\mathbf{D}_\alpha}^{(c)}$ and ${\mathbf{D}_\omega}^{(c)}$ for each concept) can be updated by back propagation.

To construct a numeric embedding for a concept c with value $f(c)$, let $f_{\min}$ and $f_{\max}$ be the lowest and highest values for that concept. I interpolate:

$$\mathbf{D}\big(f(c)\big) = w_\alpha \, \mathbf{D}_\alpha + w_\omega \, \mathbf{D}_\omega,$$

where

$$w_\alpha = \frac{f_{\max} - f(c)}{f_{\max} - f_{\min}}, \quad w_\omega = \frac{f(c) - f_{\min}}{f_{\max} - f_{\min}}, \quad w_\alpha + w_\omega = 1.$$

To understand how well our approach preserves numeric information, I *recover* $f(c)$ at each layer. Let $\mathbf{h}_\ell(f(c))$ denote the hidden state (at layer ℓ) corresponding to our numeric token. I measure its dot products with $\mathbf{D}_\alpha$ and $\mathbf{D}_\omega$:

$$\mathrm{dot}_\alpha = \mathbf{h}_\ell\big(f(c)\big) \cdot \mathbf{D}_\alpha, \quad \mathrm{dot}_\omega = \mathbf{h}_\ell\big(f(c)\big) \cdot \mathbf{D}_\omega.$$

I define the normalized weight

$$w_\omega = \frac{\mathrm{dot}_\omega}{\mathrm{dot}_\alpha + \mathrm{dot}_\omega},$$

which approximates $\frac{f(c)-f_{\min}}{f_{\max}-f_{\min}}$. From here, I recover

$$f(c) = w_\omega \, (f_{\max} - f_{\min}) + f_{\min},$$

with $w_\alpha = 1 - w_\omega$.

Although this alpha–omega formulation lets us reconstruct $f(c)$ by dot products in principle, not all configurations store $\mathbf{D}_\alpha$ and $\mathbf{D}_\omega$ at inference time. For instance, in the universal variant, I reuse the same $\mathbf{D}_\alpha$ and $\mathbf{D}_\omega$ across all numeric concepts to generate a single composite embedding for each numeric value. This approach does not preserve concept-specific demarcator vectors at inference time, making direct do-based numeric recovery more challenging. In practice, I still pass tokens through the same sequence of attention layers, feed-forward blocks, and the final linear projection to produce logits. I retrieved standard language-model metrics such as perplexity to see whether numeric detail correlates with overall modeling quality.

When a configuration does maintain $\mathbf{D}_\alpha{}^{(c)}$ and $\mathbf{D}_\omega{}^{(c)}$) (e.g., in a fully trainable concept-specific model), I can measure numeric reconstruction quality via the mean squared error (MSE) between the recovered $f(c)$ and the true $f(c)$. To analyze the impact of deeper layers on $\mathbf{D}_\alpha$ and $\mathbf{D}_\omega$, I feed these vectors into the transformer as input tokens. At each layer ℓ, I record the hidden-state vectors $\mathbf{h}_\ell(\mathbf{D}_\alpha)$ and $\mathbf{h}_\ell(\mathbf{D}_\omega)$. By measuring their norms at each layer, I quantify how the network transforms or amplifies numeric information. Larger norms may indicate the transformer amplifies numeric distinctions in deeper layers, whereas more gradual changes could suggest stable numeric representations.

Finally, I compare each configuration (addition vs. Hadamard, universal vs. concept-specific) on perplexity and, where possible, alpha-omega numeric recovery.

5.4 Results

First, I examined how different binding methods affect performance. I found that Hadamard-based binding does not necessarily improve upon simple addition (Table.5.2). Despite in the MIMIC data, the graded-concept-specific-addition had the best performance. The graded-universal-addition was the second best performing model. For the UW data, Graded-universal-addition was the best performing model across all configurations. This suggests that bundling (addition) may be enough or perhaps that the Hadmard product is numerically unstable. In our perplexity analysis over training epochs, the Hadamard product variant initially shows a sharp decrease in perplexity for about three epochs, then only modest improvements afterward(Table.5.2).

Second, I investigated how MSE changes across layers for models that do allow alpha–omega recovery(only the trainable models)(Tables 5.3, 5.4). Overall, MSE often grows monotonically from lower to higher layers, suggesting that attention and feed-forward mechanisms progressively reshape the numeric embeddings(Appendix Figures A.1 and A.2). Notably, the Hadamard-product configurations tend to show a steeper rise for the norms at deeper layers. The MSE for the addition bundling is slightly higher than the Hadamard binding (Tables 5.3, 5.4), which indicates a worse preservation of the original numeric value.

Finally, perplexity results on both UW and MIMIC datasets show that an explicit binding step (Hadamard product) is not strictly necessary. In many cases, simple bundling achieves comparable or better performance.

Table 5.2: Comparison of bindings on MIMIC and UW inpatient data. Each model was trained for at least 5 epochs with a 10% early-stop on validation loss.

MIMIC inpatient summary

	Graded Universal fixed, addition	Graded Universal fixed, Hadamard binding	Graded Concept-specific fixed, addition	Graded Concept-specific fixed, Hadamard binding	Graded Concept-specific, trainable, addition	Graded Concept-specific, trainable, Hadamard binding	Rotary Universal, trainable, addition	Rotary Universal, trainable, Hadamard binding	Rotary Concept-Specific, trainable, addition	Rotary Concept-Specific, trainable, Hadamard binding
OMOP + EU (82+, 101-)	0.623	0.612	0.634	0.605	0.613	0.650	0.655	0.664	0.618	0.610
UMN (58+, 16-)	0.651	0.509	0.587	0.571	0.601	0.669	0.639	0.574	0.531	0.578
AZCERT (33+, 215-)	0.655	0.610	0.649	0.627	0.574	0.641	0.581	0.563	0.629	0.613
Across all sets (173+, 332-)	0.688	0.642	0.684	0.663	0.636	**0.701**	0.666	0.671	0.653	0.654

UW inpatient summary

	Graded Universal fixed, addition	Graded Universal fixed, Hadamard binding	Graded Concept-specific fixed, addition	Graded Concept-specific fixed, Hadamard binding	Graded Concept-specific, trainable, addition	Graded Concept-specific, trainable, Hadamard binding	Rotary Universal, trainable, addition	Rotary Universal, trainable, Hadamard binding (running)	Rotary Concept-Specific, trainable, addition	Rotary Concept-Specific, trainable, Hadamard binding
OMOP + EU (74+, 86-)	0.650	0.640	0.650	0.623	0.665	0.642	0.591	0.607	0.659	0.605
UMN (49+, 13-)	0.630	0.619	0.589	0.612	0.583	0.597	0.561	0.597	0.615	0.575
AZCERT (32+, 211-)	0.640	0.638	0.646	0.646	0.583	0.641	0.613	0.624	0.634	0.641
Across all sets (155+, 340-)	**0.709**	0.700	0.702	0.696	0.702	0.687	0.659	0.660	0.702	0.685

Table 5.3: MSE results for the graded models. MSE was calculated for trainable models only. MIMIC

Model	**Mean MSE**	**Max MSE**	**Min MSE**	**Std MSE**	**Perplexity**
Graded vectors, universal, fixed, addition	-	-	-	-	1.610
Graded vectors, universal, fixed, Hadamard binding	-	-	-	-	6.762
Graded concept-specific, fixed, addition	-	-	-	-	1.612
Graded concept-specific, fixed, Hadamard binding	-	-	-	-	6.766
Graded concept-specific, trainable, addition	0.113	0.123	0.0651	0.0150	1.600
Graded concept-specific, trainable, Hadamard binding	0.0785	0.0888	0.0440	0.0115	1.887

Table 5.4: MSE results for the graded models. MSE was calculated for trainable models only. UW

Model	**Mean MSE**	**Max MSE**	**Min MSE**	**Std MSE**	**Perplexity**
Graded vectors, universal, fixed, addition	-	-	-	-	11.72
Graded vectors, universal, fixed, Hadamard binding	-	-	-	-	41.96
Graded concept-specific, fixed, addition	-	-	-	-	11.72
Graded concept-specific, fixed, Hadamard binding	-	-	-	-	41.97
Graded concept-specific, trainable, addition	0.0799	0.0888	0.0490	0.0097	11.70
Graded concept-specific, trainable, Hadamard binding	0.0611	0.0668	0.0290	0.0102	39.39

5.5 Discussion

In this chapter, I extended the previous chapter by examining both "bundling" (simple addition) and "binding" (Hadamard multiplication) for combining numeric embeddings with GPT-2's token and positional embeddings. I then compared universal vs. concept-specific approaches, including fully trainable demarcator vectors $\mathbf{D}_{\alpha}{}^{(c)}$ and $\mathbf{D}_{\omega}{}^{(c)}$. Our key finding is that binding does not necessarily outperform bundling (see Tables 5.2, 5.3, 5.4).

For the MIMIC set, the drop in performance with explicit binding is larger. For the UW set, they were compariable. One plausible explanation is that the MIMIC and UW datasets differ substantially in both scope and curation as discussed in the previous chapters. MIMIC is much larger (approximately five times) and contains primarily structured numeric or timestamp data with minimal free-text. In contrast, the UW dataset does include clinical notes and was filtered toward specific drug events and outcomes. Consequently, the numeric features (e.g., lab values, dosage amounts) in MIMIC may be more heterogeneous and span a wider range(because UW structured data is smaller). If the Hadamard product amplifies certain numeric outliers or magnifies variability in ways that hinder stable learning, it could explain why explicit binding is more detrimental on MIMIC than on a more constrained dataset like UW. Additionally, MIMIC's more precise timestamps may lead to a more accurate but noisier temporal embeddings.

Although Transformers do much more than pure linear addition, several studies have shown that even straightforward linear combinations (e.g.vector superposition) can be quite powerful in a sufficiently high-dimensional space[53, 94]. This ties in with the VSA perspective, where superpositions in high-dimensional vector spaces can represent multiple pieces of information compactly[53]. Knittel et al. observe that the GPT-2 token embeddings are relatively uncorrelated, approximating orthogonality in high-dimensional space[57]. This behavior aligns with certain VSA assumptions about distributing information across large vector spaces. In the VSA view, superposition (often called bundling) can combine multiple vectors into one, while binding (e.g. via the Hadamard product) encodes structured relationships. Transformers do both: they superimpose signals in residual streams, and also implement forms of binding-like operations via attention[57]. Hence GPT-2 might appear to utilize VSA-like principles even without explicit 'binding' tokens. Another abstract on VSA [96] also describes how linear operations in high-dimensional spaces provide a simple but of-

ten adequate approach. The same author later showed that while non-linear expansions such as a Hadamard product can encode more structure, they may distort original similarity relationships [97]. Our GPT-2 model, with an embedding dimension of 768, is not strictly hyperdimensional, but it still benefits from many properties of high-dimensional spaces, which helps explain why simple addition can be sufficient or surpass a more explicit binding.

In another paper published in 2024, it explores how LMs solve the binding problem. The paper focuses on understanding how entities are associated with their attributes. The authors propose that LLMs internally using binding IDs, which are abstract vectors added to entities and attributes to maintain associations. The paper showed some empirical evidence that by replacing or swapping binding ID vectors leads to predictable changes in associations, confirming that the binding IDs encode binding information. In addition, binding is not tied to the sequence position but to the binding IDs[31]. This also partially aligns with what I observed in chapter 3 that the presence of positional embedding does not significantly improve the predictability(Table 3.3). It could be another piece of evidence that the transformer model relies more on the specialized mechanism within the activation space.

For models that have trainable alpha and omega, I measured mean squared error (MSE) for numeric reconstruction on the MIMIC and UW datasets. Some configurations do not retain alpha–omega vectors because the graded vectors were pre-calculcated before the training, so in those cases I relied mainly on perplexity or other downstream metrics. Even so, these models perform similarly, that the MSE for Hadamard binding is lower than bundling. The perplexity was showing the opposite. This part of the result is inconclusive.

In real-world applications like medical records, small differences in numeric accuracy could matter; even minor reconstruction errors might affect predictions. But our results still indicate that both bundling and binding can encode numeric information without large distortions. Future work could be, for example, working on the same analysis but using some benchmark datasets for a more direct comparison.

I also examined hidden-state norms across layers, noticing that norms often grow in deeper layers as the transformer emphasizes task-relevant features [50, 103] Jawahar et al. suggested that lower layers in the transformer models capture more surface-level features, while deeper layers handle semantic features. In the genomic domain, shallower layer has a broader attention, and deep layer

focus on phenotypic genes [141]. In our experiments, I did not find a correlation between numeric MSE and final performance metrics (like AUROC) at the layer level.

Our approach has several limitations. First, there are still some unanswered open-ended questions. For example, more work needs to validate whether internal binding is happening in the customized value-aware GPT-2 models. Secondly, I tested only on MIMIC and UW inpatient datasets, which may not fully represent the breadth of clinical or general numeric data. Finally, some configurations did not preserve the alpha–omega vectors for direct numeric decoding, limiting our ability to measure numeric fidelity across all models.

Overall, our results suggest that a simple bundling approach can be competitive, and universal embeddings can match concept-specific embeddings—even if alpha–omega vectors are fully trainable. While certain models can do a direct alpha–omega numeric recovery, it does not necessarily translate into better perplexity or downstream performance. Future studies might address the limitations and explore how attention mechanisms or layer-level constraints can further stabilize numeric representations without sacrificing the transformer's broader language modeling performance.

5.6 Conclusion

I find that an explicit binding step (e.g., Hadamard product) does not consistently outperform simple addition for encoding numeric embeddings in GPT-2. Moreover, concept-specific embeddings do not reliably surpass universal approaches, indicating that a straightforward, universal addition strategy generally is sufficient for numeric handling in high-dimensional transformer models.

Chapter 6

IMPROVING DRUG DRUG INTERACTION DETECTION IN ELECTRONIC HEALTH RECORDS USING GPT2 MODELS

In this exploratory aim, I used the best performing model from the previous aims to further explore their utility for detecting harmful drug drug interactions.

6.1 Abstract

I introduce a transformer-based next-token prediction approach—built on a GPT-2 backbone with value-aware graded embeddings—and compare it against established pharmacovigilance methods (BCPNN, GPS, and logistic regression) for detecting drug–drug interactions (DDIs). Using only EHR data, our model significantly outperforms these baselines, achieving up to a 30% improvement in AUROC. By capturing temporal and clinical relationships that simpler regression or disproportionality methods often miss, this method effectively distinguishes true ADR-related drug pairs from negatives. These findings highlight the potential of generative language models in pharmacovigilance.

6.2 Introduction

Adverse drug reactions (ADRs) resulting from drug–drug interactions (DDIs) pose a significant threat to patient safety worldwide [47]. Although research in this area has been ongoing for decades, harmful interactions still frequently go undetected[92]. One reason is the exponential growth of possible drug combinations: with N drugs, there are $N(N-1)/2$ distinct pairs to assess(for drug-drug-disease, even more with more than 3 drugs working synergistically), making in-vitro studies expensive. Nonetheless, computational approaches have expanded the field by offering systematic ways to detect potential interactions at scale.

Early methods relied heavily on spontaneous reporting systems (SRS), where voluntary reports of suspected ADRs are collected. Techniques such as the proportional reporting ratio (PRR), reporting odds ratio (ROR), association rules, and logistic regression were introduced to flag possible DDIs [88, 12, 58]. More recent SRS-based methods include Bayesian Confidence Propagation Neural

Network (BCPNN) and Gamma Poisson Shrinkage (GPS), both of which use Bayesian frameworks to assess whether certain drug–event co-occurrences deviate from what would be expected by chance [11, 29, 113]. BCPNN calculates an Information Component (IC) to quantify deviation from independence [11], whereas GPS uses a Gamma–Poisson model to shrink noisy estimates toward more stable values [29]. Although these methods are effective, they often suffer from under-reporting and capture only a cross-sectional snapshot of the data.

Logistic regression (LR) has also been adapted to SRS data, sometimes incorporating covariates such as patient age, sex, and drug–drug interaction terms [122, 123, 41, 88, 134]. Bayesian logistic regression can handle high-dimensional data [2], but limitations persist, including collinearity issues and bias tied to overrepresentation of severe ADRs. More recently, deep learning approaches—including similarity-based heuristics, network diffusion, Bayesian networks, CNNs, and graph neural networks—have broadened DDI detection methods [143, 89]. Transformers, introduced by Vaswani et al.[124], can bring greater flexibility by leveraging self-attention to model complex relationships among drugs, clinical events, and patient characteristics.

Compared to SRS and in vitro studies, EHRs remain underutilized in DDI research—even though several studies attest to their value. Wu et al. (2021) proposed a "Drug–Drug Interaction Wide Association Study (DDIWAS)," mining drug pairs from EHR allergy lists and applying it to 616 drugs [134]. They identified 34 known interactions for simvastatin and amlodipine (with high positive predictive values of 0.85 and 0.86) and further revealed novel interactions such as simvastatin–hydrochlorothiazide. While impactful, their approach relies on allergy-list data, which may be unavailable. In addition, they focused on a limited number of drugs, requiring extensive expert review. Tatonetti et al. (2011) used profiles of eight major ADR categories in the FDA's Adverse Event Reporting System (FAERS) to find drug pairs with side effect patterns matching known profiles [115], identifying 171 previously unreported interactions. These were partially validated by multivariate models using EHR data. Iyer et al. (2014) performed large-scale EHR text mining on over 50 million clinical notes from two healthcare systems, using disproportionality ratios to highlight significant drug pair–event associations; they achieved AUROCs above 0.8 when using EHR alone which is comparable to SRS performance (using the same methods) [49]. Lorberbaum et al. (2016) investigated QT prolongation risk by combining EHR-based QT interval analysis with laboratory experiments, illustrating how observational data can be paired with in vitro validation

to strengthen evidence [75].

Moving beyond disproportionality and traditional regression, Transformer architectures present a promising path for enhanced DDI signal detection. They can integrate heterogeneous data: structured, unstructured, and longitudinal into one framework. Large-scale pretraining enables stronger generalization, while the self-attention mechanism captures complex temporal relationships among multiple drugs, ADRs, and patient-specific factors, making Transformers ideal for polypharmacy scenarios [141]. Although relatively few studies had applied self-attention to uncover DDIs in 2023, by 2025 the technique has become more widely accepted. Many new attention-based DDI approaches have emerged [137, 140, 130], reflecting the evolving capabilities of Transformer-based methods.

In this chapter, I propose a novel application of a value-aware GPT-2 model for DDI signal detection using only EHR data. GPT-2 was selected for its open-source availability and manageable computational requirements, in contrast to more resource-intensive models such as lLama [118, 38], which was originally trained on texts. Moreover, I incorporate graded embeddings for numerical clinical variables—an approach that our prior experiments indicate can represent diagnosis, medications and lab results effectively. I hypothesize that this Transformer-based method will surpass classical DDI detection methods and simpler machine learning baselines in uncovering clinically relevant DDIs.

6.3 Method

6.3.1 Reference Set and Data Construction

I obtained all positive and negative DDI pairs from the CRESCENDDI reference sets, which integrate multiple clinical resources—including the British National Formulary (BNF), the French National Drug Safety Institute (ANSM), and Micromedex [60]. In total, these reference sets contain 10,286 positive and 4,544 negative DDI–ADR pairs. Positive cases originate from DDI sources with overlapping evidence, whereas negative cases were verified through PubMed queries to ensure no documented association with the given ADR [60]. A separate publication further validates variations between BNF, Thesaurus, and Micromedex, which informed the development of the CRESCENDDI reference sets [59].

Our analysis began by linking each adverse drug reaction (ADR) to its corresponding ICD-9 and ICD-10 codes, then merging these codes with an extensive set of drug–drug–interaction-event (DDI-E) triplets. Out of more than 3,000 potential combinations, I filtered for those involving a common drug A that formed both a known ADR-associated pair $(A + B)$ and a non-ADR pair $(A + C)$. I further required that drugs appear *before* the relevant ICD code (diagnosis) in observed hospitalization sequences. Combinations with fewer than 10 occurrences in MIMIC or fewer than 2 occurrences in UW (a much smaller dataset) were discarded. This process yielded roughly 1,400 unique pairs in MIMIC and 280 pairs in UW. I selected the 10 most frequent conditions from the reference set and ultimately focused on nine conditions that had sufficient data: acute renal failure, increased blood pressure, bradycardia, depressed level of consciousness, hemorrhage, hyperkalemia, hyponatremia, hypotension, and respiratory depression.

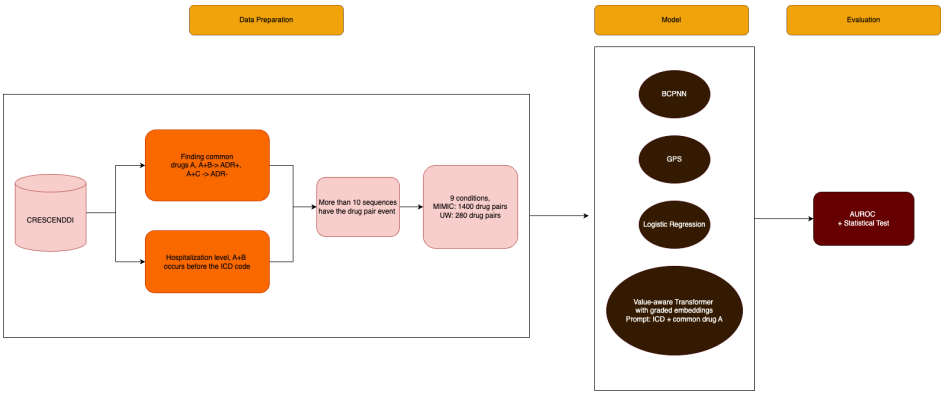

Figure 6.1: Overview of the method: from reference-set linkage to final model evaluations.

6.3.2 Baseline Models

Logistic Regression (LR) Logistic regression (LR) was chosen as a baseline because of its established use in DDI detection. The model estimates:

$$p = \Pr(Y = 1),$$

where $Y = 1$ indicates the presence of an ADR. I included three binary predictors (A, B, C) denoting whether each of three drugs is present. Specifically, A is the common drug, while B and C are the

other two drugs under consideration. The logistic equation is:

$$\ln\Big(\frac{p}{1-p}\Big) = \beta_0 + \beta_A A + \beta_B B + \beta_C C + \beta_{AB}(A \times B) + \beta_{AC}(A \times C),$$

where $\beta_A, \beta_B, \beta_C$ capture the main effects, and β_{AB}, β_{AC} capture additional interaction effects and I repeated the same procedure for each ADR. In alignment with Wu et al. [134], I also included age and gender for MIMIC. For UW, only age was available. I used an 80/20 train–test split and applied L2 regularization.

BCPNN I adopted the Bayesian Confidence Propagation Neural Network (BCPNN) to assess the disproportionality between a drug pair (exposure) and an ADR (outcome). For each drug-ADR combination, a 2×2 contingency table is constructed:

$$\begin{aligned} n_{11} &: \text{ drug pair + ADR both present,} \\ n_{10} &: \text{ drug pair present, ADR absent,} \\ n_{01} &: \text{ ADR present, drug pair absent,} \\ n_{00} &: \text{ neither drug pair nor ADR.} \end{aligned}$$

BCPNN then calculates:

$$\log_2\Big(\frac{(n_{11}+n_{10})(n_{11}+n_{01})}{n_{11}\,N}\Big).$$

and $N = n_{11} + n_{10} + n_{01} + n_{00}$, which reflects the ratio of observed to expected co-occurrence. Higher scores suggest stronger DDI signals.

GPS Alongside BCPNN, I computed the Gamma Poisson Shrinkage (GPS) score for each drug–event pair:

$$\text{GPS} = \frac{n_{11}}{E},$$

where E is the expected value

$$E = \frac{(n_{11}+n_{10})(n_{11}+n_{01})}{N},$$

Both BCPNN and GPS are standard tools in pharmacovigilance. I tested multiple cutoffs (e.g., GPS > 3, BCPNN > 2) but ultimately used continuous values to calculate AUROCs, which I found to be more principled than selecting arbitrary thresholds.

6.3.3 GPT-2 Value-Aware Model

I next developed a custom inference pipeline to measure whether a GPT-2 model assigns higher probabilities to positively labeled DDEs (drug–drug–event) than to negatives. For each record in our dataset, I constructed a prompt by concatenating the ICD code tokens with the "common drug" tokens. I then calculated the softmax and sigmoid likelihood of the previous token would being a known positive or negative drug from the reference set. By averaging these probabilities across the top three token variants from our custom vocabulary, I obtained a single probability indicating how "logical" GPT-2 deemed each candidate drug, given the ICD code and the common drug.

This next-token prediction approach is based on the best-performing GPT-2 model from previous chapters. It leverages value-aware graded embeddings (universal variant) to encode numeric clinical variables (e.g., lab results and drug dosages).

I did two versions of the analysis: 1) Full: including full dataset with as many DDEs as possible. For MIMIC, there are 5872 unique drug-drug-ICD triplets. After aggregating by ADR, there were 322 unique DDEs included in the GPT-based prediction. For UW, there are 1136 unique drug-drug-ICD triplets, 126 unique DDEs made to the analysis 2) Bounded: the bounded version that the DDEs appear in BCPNN, GPS, LR(excluding the cases where A=1 for every observation), and the GPT based method. For MIMIC data, there were 74 unique DDEs met the selection criteria. For UW data, there were 27 unique DDEs.

6.3.4 Evaluation

I compared four methods—BCPNN, GPS, LR, and the value-aware GPT-2—by calculating the area under the receiver operating characteristic curve (AUROC) for their continuous scores. The previous value-aware based GPT-2 model was trained in a reverse order. Each model outputs a probability (or score) representing the $P(drugPair|ADR)$. I generated ROC curves and summarized their performance with AUROC. Additionally, I conducted Welch's t-tests to verify whether GPT-2 scores for positive pairs were significantly higher than those for negative pairs, providing a supplementary check on the model.

6.4 Results

For the baseline models, I identified approximately 74 unique drug–event pairs in MIMIC and 24 in UW that met our inclusion criteria. For both logistic regression (LR) and the GPT-2–based next-token predictor, I also evaluated each model on the entire dataset (the full version, see Appendices A.3–A.10). In MIMIC, the weighted average AUROC for LR across all target conditions was 0.67; in UW, it was 0.59(Appendix Figures 5-8). By contrast, our GPT-2 model with value-aware embeddings outperformed all baselines(Figures 6.2, 6.4). When applied to the full dataset in MIMIC, it achieved an AUROC of 0.89 using softmax-based probabilities and 0.79 using sigmoid-based probabilities (Appendix Figure A.8, A.10). For UW, the respective AUROCs were 0.79 and 0.77 (Appendix Figure A.9, A.11). As illustrated Figures.6.4, 6.5), the GPT-2 model effectively distinguished ADR-associated drug pairs from those without documented ADRs.

I further tested the bounded subset of the data in which all methods could be applied consistently (e.g., removing drug–event pairs with insufficient sample sizes or complete collinearity). In MIMIC, the baseline methods' AUROCs were around 0.60, whereas GPT-2 surpassed them by at least 0.25 on average. In UW, baseline results ranged from about 0.40 to 0.60, and again the GPT-2 model demonstrated a significantly higher AUROC (Figure 6.3, 6.5). These results confirm that our approach generalizes across two distinct EHR datasets.

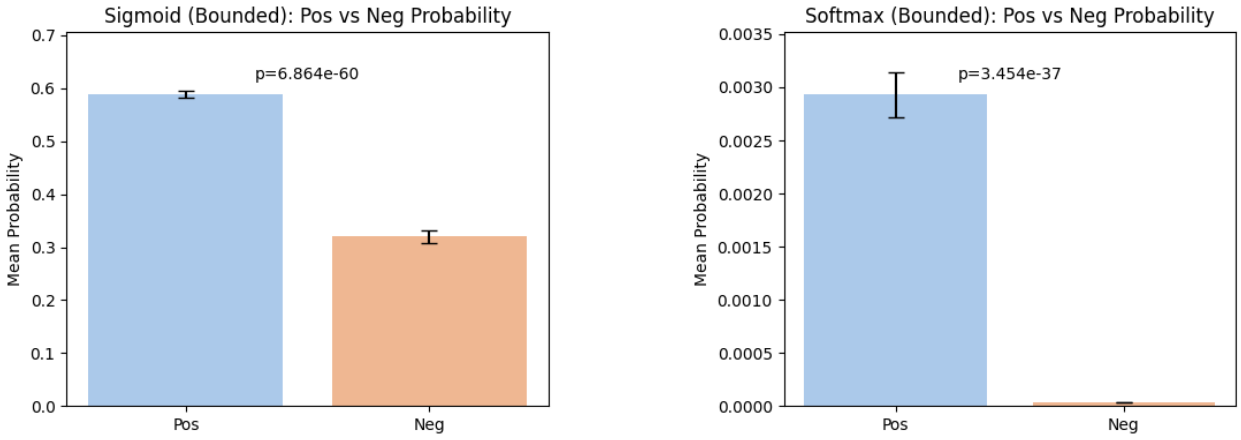

Figure 6.2: Comparison of the sigmoid (left) and softmax (right) probabilities of positive versus negative drug pairs using data that are bounded by all methods(MIMIC, N=74 DDEs).

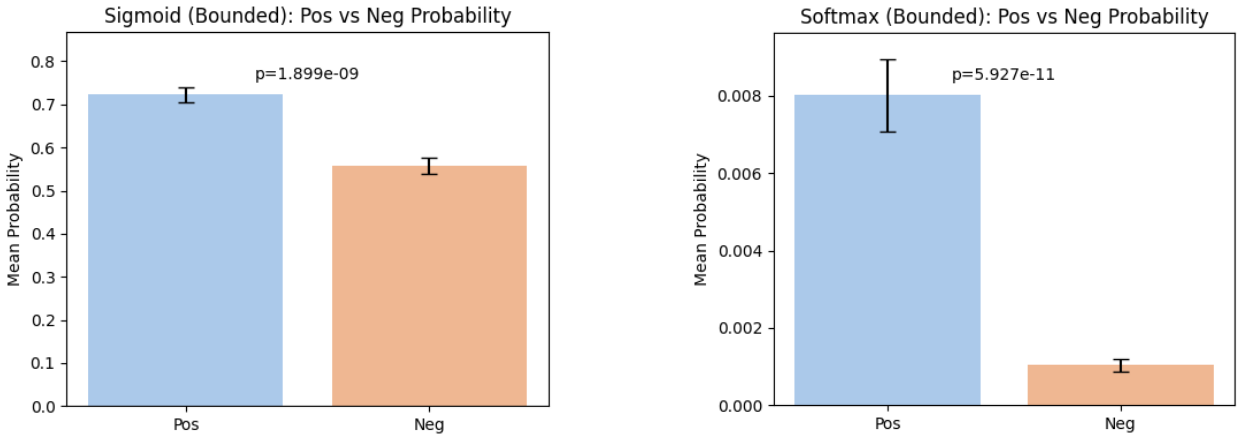

Figure 6.3: Comparison of the sigmoid (left) and softmax (right) probabilities of positive versus negative drug pairs using data that are bounded by all methods(UW, N=27 DDEs).

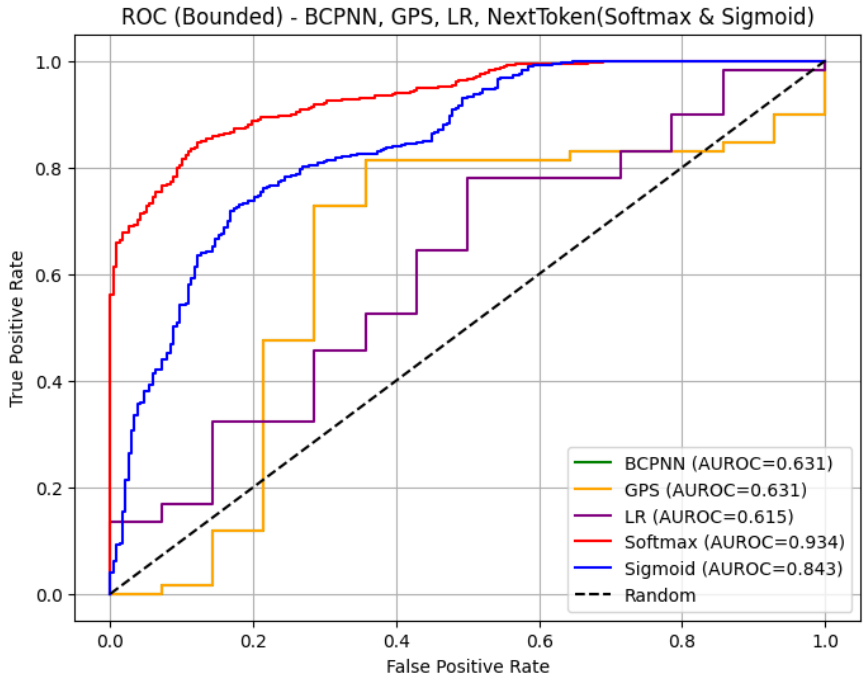

Figure 6.4: AUROC for BCPNN, GPS, Logistic Regression and GPT-2 based next token prediction (softmax and sigmoid), MIMIC

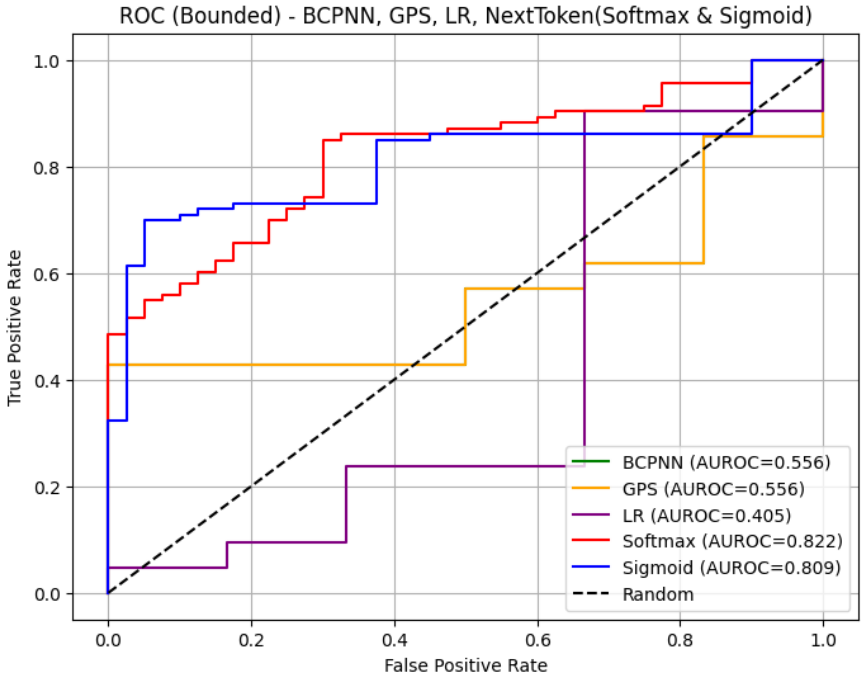

Figure 6.5: AUROC for BCPNN, GPS, Logistic Regression and GPT-2 based next token prediction (softmax and sigmoid), UW

6.5 Discussion

In this chapter, I evaluated our GPT-2–based model, which was described in the previous chapter, for its ability to detect harmful DDIs that lead to ADRs. By leveraging the model's generative capabilities, I compared its performance against conventional DDI detection methods—including BCPNN, GPS, and logistic regression. I found that GPT-2 outperformed the baselines in both MIMIC and UW datasets. This is a compelling outcome given that EHR-based DDI detection, in contrast to SRS, remains relatively uncommon due to challenges like incomplete documentation and heterogeneous data structures.

Both BCPNN and GPS are widely accepted methods in pharmacovigilance for detecting single-drug and multi-drug ADR signals [11, 113, 19]. They use shrinkage estimates to mitigate high false-positive rates [19], but neither has been definitively shown to be superior in every context [88]. Prior studies report average AUROCs of 0.5–0.7 on OMOP data for ADR detection, and no single metric

was consistently best [19]. In our analysis, the BCPNN and GPS methods gave the output with the same rank ordering, which is not surprising. Because BCPNN is simply $log_2(GPS)$In keeping with findings that machine-learning methods often outperform disproportionality-based metrics in EHR data [51], I saw that our Transformer-based approach was more effective at distinguishing truly harmful drug pairs.

I included LR as another baseline because disproportionality metrics are primarily designed for SRS, whereas LR can incorporate more granular hospitalization-level data. In our analysis, LR showed performance comparable to BCPNN and GPS in MIMIC but performed worse in UW. As previously noted, EHR-based DDI detection is still uncommon, with most existing methods developed for SRS. Notably, as described in the method, if a "common drug" appears in every row, that predictor column becomes a constant (all ones). Including such a column—alongside the model intercept—introduces complete collinearity, so it must be excluded. Consequently, any interaction term involving that drug (A $\times$ B) simply reduces to the effect of B under the condition that A is always present, rather than reflecting a true interaction. Numerically, in a binary dataset where A = 1 for every observation, the column for A $\times$ B is identical to that for B. Although the model can still be fit, the coefficient for this "interaction" effectively represents the effect of B for patients who always receive Drug A. In MIMIC, this collinearity issue arose in LR models for bradycardia and hyperkalaemia, so they were dropped in the bounded version. For UW, only depressed level of consciousness, acute renal failure, respiratory depression, and blood pressure increased had sufficient data to include both A $\times$ B and A $\times$ C interaction terms. Nevertheless, LR still did not outperform the Transformer-based method. I speculate that LR's weaker performance in UW than in MIMIC stems from its smaller sample size and the absence of additional variables such as gender.

Initially, I intended to scrutinize the interpretability of Transformer attention layers, similar to previous work on gene–gene interactions [141, 125]. However, the attention weights varied widely across different random seeds, prompting us instead to adopt the next-token prediction probability as a more stable measure. I also enforced strict chronological ordering of ICD codes and drugs, ensuring that diagnoses preceded medication usage. This sequence ordering minimized temporal confounding and aligned with the real-world timing of clinical events. While I attempted to construct hypothetical sequences (ICD code + common drug + positive drug + negative drug) to

measure attention directly, the model's output fluctuated too significantly with different initializations. Restricting our approach to realistic hospital sequences ultimately produced the most reliable results.

Despite these promising findings, our study has several limitations. First, only a subset of CRESCENDDI's thousands of reference DDI–ADR pairs could be matched to EHR data, limiting the number of conditions I could include. Expanding to a greater range of ADRs would offer further evidence of model robustness. Second, while GPT-2 was chosen here for its superior performance in earlier experiments, other Transformer architectures or additional hyperparameter tuning might yield even better results. Third, our study was limited to MIMIC and UW; validating on other EHR systems, preferably with richer clinical documentation (e.g., progress notes, medication administration records, claims data), could enhance model performance. Lastly, I didn't include other transformer based methods as a baseline.

For future analysis, I will address the limitation discussed above. The CRESCENDDI reference set can be extended to incorporate severity indicators and mechanistic features from systems pharmacology [58, 61], potentially supporting new baselines such as modified Bayesian or ML-based methods. Deeper exploration of attention-based DDI models [142, 107, 89] or fine-tuning with external data sources like FAERS or knowledge graphs [137] also represents a promising path forward. Beyond pairwise interactions, future research should evaluate whether Transformer-based models can predict more complex polypharmacy risks. Such extensions would further align with real-world prescribing environments, where patients commonly take multiple medications simultaneously.

6.6 Conclusion

Our findings underscore the promise of applying Transformer-based language models— with value-aware embeddings—to detect adverse drug reactions directly from EHR. Even with limited ICD-based condition labeling and a modest number of reference DDI pairs, the model surpassed established methods in discriminating known harmful drug combinations from negative samples. In doing so, it reaffirmed the importance of harnessing the rich temporal and contextual information embedded in EHR data for monitoring drug safety.

Chapter 7

SUMMARY, CONCLUSIONS AND VISION

This chapter follows the outline of the last chapter of "The Organization and content of informatics doctoral dissertations" by Shortliffe 2016 [109].

7.1 Abstract

This final chapter summarizes how transformer-based models, leveraging electronic health record (EHR) data, address important gaps in post-market drug surveillance. Traditional pharmacovigilance methods often neglect temporal patterns, numeric data (e.g., dosage, lab values), and the complexities of polypharmacy. Generative transformers trained on two distinct EHR datasets (MIMIC-IV and UW), surpassed standard disproportionality metrics in adverse drug reaction (ADR) detection, incorporating value-aware embeddings to capture numeric details. In addition, this approach effectively models drug–drug interactions (DDIs), demonstrating superior predictive performance to simpler machine learning baselines. These results show the potential for transformer architectures to enhance both ADR signal detection and patient safety.

7.2 Summary of accomplishments and Contributions

In this dissertation, I addressed the challenges in post-market drug surveillance by focusing on three limitations of existing methods:

- Lack of robust temporal modeling
- Insufficient representation of numeric data (e.g., dosage, lab values)
- Limited characterization of drug drug interactions in polypharmacy

I proposed transformer-based methods that rely exclusively on EHR data rather than spontaneous reporting systems (SRS) data, demonstrating that:

- A transformer model trained from scratch can surpass conventional disproportionality metrics in ADR signal detection.

- Incorporating value-aware embeddings enhances prediction by encoding numeric information (drug dose, lab measurements), addressing a major gap in existing sequence models.

- The trained models can effectively predict drug–drug interactions (DDIs) across diverse conditions

All experiments were conducted on two EHR datasets, MIMIC-IV and UW, with consistent improvement reinforcing the feasibility of using clinical records (instead of SRSs) to improve ADR signal detection.

7.2.1 Innovation

Most EHR-based ADR detection studies remain cross-sectional, capturing only drug–event co-occurrences over narrow intervals and neglecting the temporal dimension essential for causal inference. In this dissertation, I introduced several innovations to address these gaps. First, I employed GPT-based architectures to represent structured EHR data as time series. Although GPT has proven powerful in general NLP tasks, few peer-reviewed publications had explored applying GPT-like transformers to structured EHR data for ADR detection. Second, I integrated dose and lab information directly into the sequence modeling, which is a feature rarely seen in drug safety studies, where dose is more often examined in pharmacokinetics than in observational post-market drug surveillance. By incorporating graded semantic vectors and other value-aware approaches, I retained granular numeric details such as creatinine fluctuations or daily drug dose changes. Third, I modeled drug–drug interactions. Prior EHR-based ADR methods did not address drug interactions or relied on narrow, manually selected sets of features. These contributions extend beyond pharmacovigilance by providing value-aware embedding strategies for any sequence-to-sequence model requiring numeric or continuous inputs. From a clinical standpoint, improving ADR and DDI detection has direct implications for patient safety, particularly since individuals might not recognize that newly experienced symptoms arise from harmful drug interactions.

7.3 Assessment of hypotheses-from Chapter 1

Chapter 1 laid out three primary hypotheses:

- H1: Transformer-based models would outperform disproportionality metrics in identifying ADR signals.
- H2: Numeric data (both dose and lab information) would further boost predictive performance of the developed model.
- H3: Modeling DDIs implicitly would reveal additional valuable signal, improving detection beyond single-drug models.

Results described in Chapters 3 and 4 clearly support Hypothesis 1. The transformer model substantially improved predictive accuracy compared to disproportionality-based approaches, reflecting a better capture of temporal and contextual patterns. For Hypothesis 2, results confirmed that models incorporating value-aware embeddings outperformed baseline architectures without numeric representations, and these gains were reproducible across two different EHR datasets (Chapter 4, 5). Finally, Hypothesis 3 was validated by the significant boost in performance when exploring polypharmacy-driven ADRs in Chapter 6. The transformer framework was effective at capturing disease-related DDIs, exceeding the performance of simpler ML baselines (e.g., logistic regression).

7.4 Generalizability of the results

Although our experiments spanned two distinct EHR repositories (MIMIC-IV and UW), further work is necessary to confirm the model's robustness across broader health systems with different populations, coding practices, and EHR structures. I also acknowledge that numeric embeddings were limited to select labs and drug dosages. In principle, any continuous metric—such as vital signs or biomarkers—could be encoded within the value-aware framework, potentially enhancing generalizability for a wider range of clinical tasks.

The proposed approach is not limited to a single disease type or diagnostic code set; in principle, it can be applied to any condition reflected in the EHR data, which underscores its high generalizability. Although I did not perform fine-tuning for specific clinical domains(clinical trial data or

data from a comprehensive disease registry), our findings indicate strong performance on gastric bleeding and acute renal failure in particular. By fine-tuning on domain-specific datasets, the model could encode more clinical relationships and perhaps achieve even better predictive accuracy. The tool could be customized to support a wide range of diseases and practice settings, fulfilling the broader goal of scalable, data-driven decision support within healthcare systems.

7.5 Range of applicability

In this study, the method I proposed outperformed method used in practice for phamracovigilance signal detection. Given its flexibility, the proposed approach can theoretically be used to build clinical decision support systems for EHR, flagging high-risk drug regimens in near real-time. However, domain expertise would be vital to calibrate or fine-tune the model for specific medical areas, where frequent ADR or DDIs have been observed.

Despite these potential applications, the method is not yet ready for direct deployment into clinical settings. There is still space for performance improvement. Furthermore, regulatory scrutiny and thorough prospective validation trials would be necessary to ensure safety and reliability of such approaches. Transformer-based algorithms show promise for post-market drug surveillance, but we are not there yet. More work is needed to deploy such a tool into clinical practice.

7.6 Limitations

There are several limitations. First, the UW dataset was originally developed with specific reference sets (e.g., OMOP and EU-ADR). Although I introduced unstructured data and mapped Unified Medical Language System (UMLS) concepts to expand the vocabulary, the dataset may still omit clinical signals(for instance, clinical terms might not be fully captured). Second, timestamp granularity differs between MIMIC-IV (which tracks events by the minute) and UW (which often records data by day). Yet, performance remained consistent despite these discrepancies.

A further limitation is that I only tested our method on MIMIC-IV and UW EHR data. Institution-specific coding conventions and local workflows might affect direct transfer of our models to other healthcare institutes. Lastly, while I highlight the potential utility of numeric embeddings, I restricted them to drug dose and certain lab values. Other numeric data, such as vital signs or imaging biomarkers, might further refine or complicate the model's predictive power.

Because each model was trained from scratch, I have presented the results from a single training run for clarity. In practice, a more robust evaluation would include multiple training runs with different random seeds to capture the variability in training. In our case, while I did conduct multiple runs using various seeds(not reported here), the observed differences in performance for the value-aware models were minimal, and our main conclusions remained consistent across all experiments.

7.7 Future work

Although the proposed method consistently shows promising results for acute renal failure and gastric bleeding, it could be further customized for specific clinical specialties. With appropriate domain knowledge and fine-tuning, the model might respond to additional nuances relevant to specific conditions, thereby offering stronger predictive performance in the targeted areas.

Another direction for future work is to experiment with different open-source LLM models, like LlaMa, Deepseek etc., since GPT-2 is still a relative small model (but we trained from scratch). More advanced models(with more parameters) might show even more impressive results.

Recent publications also suggest directions for enhancing our approach. One 2024 study trained a generative transformer model to predict the next medical event— a disorder, a medication, or a symptom, and used GPT-2 to forecast if a particular code would appear within the next 10(or 2 or 1) tokens within various timeframe [63]. However, when the model encountered sequences it had never seen, a drop in accuracy occurred. Nevertheless, this approach could potentially be integrated into our existing framework. In another preprint authored by researchers at the UW, an architecture similar to ours was applied to the eMERGE Network, with a few differences: it excluded positional embeddings and used Sentence-BERT to flatten transformer-generated patient embeddings [136]. The authors did not train the model from scratch. They initialized the pretrained embeddings from an autoencoder that they trained previously on ICD and medical procedure codes. They included a logistic regression classifier for predicting future medical events and demonstrated high accuracy, though they did not incorporate numeric data. Integrating our value-aware embeddings could improve these approaches and extend their usefulness to broader EHR settings.

For DDI detection, multiple opportunities remain. Incorporating knowledge graphs or ontologies [137](e.g., SNOMED-CT, RxNorm) may enhance the algorithm's ability to infer more complex

drug–drug interactions. Additionally, developing gold standard or benchmark datasets—as discussed by Zhang et al. [143], would facilitate more uniform comparisons across DDI studies. Also, extending our method to discover unknown DDIs is of interest for future work.

7.8 Vision and Close Remarks

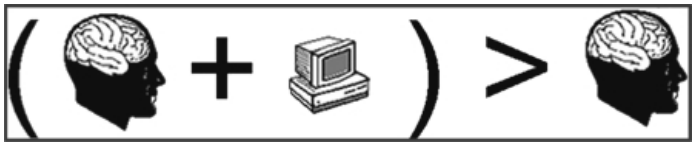

Figure 7.1: Human plus a computer > human alone [48].

My journey into biomedical informatics was first started with the "central theorem of informatics" illustrated in Figure 7.1[48], which suggests that a clinician's mind plus a computer can exceed the capacities of a clinician alone(thinking in the context of medicine). This principle resonates strongly today, as advanced generative models (e.g., GPT-o1, LLaMA variants) rapidly emerge as collaborative tools in biomedicine. Recent reviews underscore the growing breadth of LLM applications in healthcare—from training models on PubMed abstracts [77] to fine-tuning them on patient–physician conversations [72], or using instruction prompt tuning for specialized medical tasks [110, 117]. While the promise of these systems is immense, concerns around hallucinations(generative AIs create misleading information), privacy, and misapplied outputs persist [116].

Nevertheless, the synergy depicted in Figure 7.1 illustrates the potential of carefully integrating human expertise with computational intelligence. Early studies show that LLMs can outperform physicians on certain diagnostic tasks, yet combining clinician insight and AI recommendations can yield unpredictable results[35]. Larger-scale prospective evaluations and regulatory frameworks will be imperative for safe adoption.

Over time, it is likely that generative models will become inseparable from clinical practice, supporting tasks like summarizing patient histories, suggesting therapeutic alternatives, or identifying high-risk medication interactions. Although the rapid development of LLMs may seem

overwhelming, it offers unprecedented opportunities to augment how clinicians reason, communicate, and make decisions. Embracing these capabilities responsibly, by ensuring robust validation and transparency, will realize the vision of "(mind + computer) > mind alone."

7.9 Conclusions

The work presented in this dissertation demonstrates how transformer-based architectures, when integrated with value-aware embeddings and longitudinal EHR data, can tackle critical gaps in ADR and DDI detection. A GPT-2 model trained from scratch on EHRs outperformed the current FDA ADR signal detection. With the addition of value-aware embeddings derived from lab and drug dose, the performance was enhanced even further. The best performing model was utilized on detecting harmful DDI and showed superior performance when comparing to the conventional methods used by FDA. As with any new technology, challenges surrounding data quality, interpretability, and clinical integration must be met with collaboration between informaticians, clinicians, and regulatory bodies. By addressing these complexities, the next generation of models can become powerful tools to improve patient safety across diverse healthcare systems.

BIBLIOGRAPHY

[1] Julia Adler-Milstein and Ashish K Jha. HITECH act drove large gains in hospital electronic health record adoption. *Health Aff. (Millwood)*, 36(8):1416–1422, August 2017.

[2] David D Lewis Alexander Genkin and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[3] Stephanie Archer, Louise Hull, Tayana Soukup, Erik Mayer, Thanos Athanasiou, Nick Sevdalis, and Ara Darzi. Development of a theoretical framework of factors affecting patient safety incident reporting: a theoretical review of the literature. *BMJ Open*, 7(12):e017155, December 2017.

[4] A R Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, pages 17–21, 2001.

[5] Dewi Susanti Atmaja, Yulistiani, Suharjono, and Elida Zairina. Detection tools for prediction and identification of adverse drug reactions in older patients: a systematic review and meta-analysis. *Sci. Rep.*, 12(1):13189, August 2022.

[6] Ar Kar Aung, Steven Walker, Yin Li Khu, Mei Jie Tang, Jennifer I Lee, and Linda Velta Graudins. Adverse drug reaction management in hospital settings: review on practice variations, quality indicators and education focus. *Eur. J. Clin. Pharmacol.*, 78(5):781–791, May 2022.

[7] Francesco Bagattini, Isak Karlsson, Jonathan Rebane, and Panagiotis Papapetrou. A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records. *BMC Med. Inform. Decis. Mak.*, 19(1):7, January 2019.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[9] Juan M Banda, Alison Callahan, Rainer Winnenburg, Howard R Strasberg, Aurel Cami, Ben Y Reis, Santiago Vilar, George Hripcsak, Michel Dumontier, and Nigam Haresh Shah. Feasibility of prioritizing drug-drug-event associations found in electronic health records. *Drug Saf.*, 39(1):45–57, January 2016.

[10] A Bate, M Lindquist, I R Edwards, S Olsson, R Orre, A Lansner, and R M De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *Eur. J. Clin. Pharmacol.*, 54(4):315–321, June 1998.

[11] Andrew Bate. Bayesian confidence propagation neural network. *Drug Saf.*, 30(7):623–625, 2007.

[12] Vera Battini, Marianna Cocco, Maria Antonietta Barbieri, Greg Powell, Carla Carnovale, Emilio Clementi, Andrew Bate, and Maurizio Sessa. Timing matters: A machine learning method for the prioritization of drug-drug interactions through signal detection in the FDA adverse event reporting system and their relationship with time of co-exposure. *Drug Saf.*, 47(9):895–907, September 2024.

[13] David R. Bellamy, Bhawesh Kumar, Cindy Wang, and Andrew Beam. Labrador: Exploring the limits of masked language modeling for laboratory data, 2023.

[14] Roberto Carvalho, Mariana Lobo, Mariana Oliveira, Ana Raquel Oliveira, Fernando Lopes, Júlio Souza, André Ramalho, João Viana, Vera Alonso, Ismael Caballero, João Vasco Santos, and Alberto Freitas. Analysis of root causes of problems affecting the quality of hospital administrative data: A systematic review and ishikawa diagram. *International Journal of Medical Informatics*, 156:104584, 2021.

[15] Center for Drug Evaluation and Research. Fda's sentinel initiative. `https://www.fda.gov/safety/fdas-sentinel-initiative`, 2024. Accessed: 2024-10-18.

[16] Trevor Cohen and Dominic Widdows. Bringing order to neural word embeddings with embeddings augmented by random permutations (EARP). In Anna Korhonen and Ivan Titov, editors, *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 465–475, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[17] Trevor Cohen, Dominic Widdows, Jason Heiden, Namita Gupta, and Steven Kleinstein. Graded vector representations of immunoglobulins produced in response to west nile virus. volume 10106, pages 135–148, 01 2017.

[18] Jamie J Coleman and Sarah K Pontefract. Adverse drug reactions. *Clin. Med.*, 16(5):481–485, October 2016.

[19] Astrid Coste, Angel Wong, Marleen Bokern, Andrew Bate, and Ian J. Douglas. Methods for drug safety signal detection using routinely collected observational electronic health care data: A systematic review. *Pharmacoepidemiology and Drug Safety*, 32(1):28–43, 2023.

[20] Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, Alexander Michel, Seleen Ong, Jill P Pell, Mary Ross Southworth, Wendy Gattis Stough, Martin Thoenes, Faiez Zannad, and Andrew Zalewski. Electronic health records to facilitate clinical research. *Clin. Res. Cardiol.*, 106(1):1–9, January 2017.

[21] Arghya Datta, Noah R Flynn, Dustyn A Barnette, Keith F Woeltje, Grover P Miller, and S Joshua Swamidass. Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from a retrospective electronic health record (EHR) cohort. *PLoS Comput. Biol.*, 17(7):e1009053, July 2021.

[22] D. G. Dauner, E. Leal, T. J. Adam, R. Zhang, and J. F. Farley. Evaluation of four machine learning models for signal detection. *Ther Adv Drug Saf*, 14:20420986231219472, 2023.

[23] S. E. Davis, L. Zabotka, R. J. Desai, S. V. Wang, J. C. Maro, K. Coughlin, J. J. oz, D. Stojanovic, N. H. Shah, and J. C. Smith. Use of Electronic Health Record Data for Drug Safety Signal Identification: A Scoping Review. *Drug Saf*, 46(8):725–742, Aug 2023.

[24] Sharon E Davis, Luke Zabotka, Rishi J Desai, Shirley V Wang, Judith C Maro, Kevin Coughlin, José J Hernández-Muñoz, Danijela Stojanovic, Nigam H Shah, and Joshua C Smith. Use of electronic health record data for drug safety signal identification: A scoping review. *Drug Saf.*, 46(8):725–742, August 2023.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[26] Bruce A Donzanti. Pharmacovigilance is everyone's concern: Let's work it out together. *Clin. Ther.*, 40(12):1967–1972, December 2018.

[27] Hossein Estiri, Zachary H Strasser, Jeffery G Klann, Thomas H McCoy, Jr, Kavishwar B Wagholikar, Sebastien Vasey, Victor M Castro, Marykate E Murphy, and Shawn N Murphy. Transitive sequencing medical records for mining predictive and interpretable temporal representations. *Patterns (N. Y.)*, 1(4):100051, July 2020.

[28] Hossein Estiri, Zachary H Strasser, and Shawn N Murphy. High-throughput phenotyping with temporal sequences. *J. Am. Med. Inform. Assoc.*, 28(4):772–781, March 2021.

[29] S J Evans, P C Waller, and S Davis. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol. Drug Saf.*, 10(6):483–486, October 2001.

[30] Kate Faasse, Jarry T Porsius, Jonathan Faasse, and Leslie R Martin. Bad news: The influence of news coverage and google searches on gardasil adverse event reporting. *Vaccine*, 35(49 Pt B):6872–6878, December 2017.

[31] Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context?, 2024.

[32] Stephen I. Gallant and T. Wendy Okaywe. Representing objects, relations, and sequences, 2015.

[33] Ross Gayler. Multiplicative binding, representation operators analogy (workshop poster). 01 1998.

[34] Mor Geva, Ankit Gupta, and Jonathan Berant. Injecting numerical reasoning skills into language models, 2020.

[35] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969, 10 2024.

[36] Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, Francois Lanusse, Michael McCabe, Ruben Ohana, Liam Parker, Bruno Régaldo-Saint Blancard, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. xval: A continuous number encoding for large language models, 2023.

[37] Jan Gosmann and Chris Eliasmith. Vector-derived transformation binding: An improved binding operation for deep symbol-like processing in neural networks. *Neural Computation*, 31:849–869, 2019.

[38] Aaron Grattafiori, ..., and Zhiyu Ma. The llama 3 herd of models, 2024.

[39] Bruce Guthrie, Boikanyo Makubate, Virginia Hernandez-Santiago, and Tobias Dreischulte. The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995-2010. *BMC Med.*, 13(1):74, April 2015.

[40] K Haerian, D Varn, S Vaidya, L Ena, H S Chase, and C Friedman. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin. Pharmacol. Ther.*, 92(2):228–234, August 2012.

[41] Manfred Hauben. Artificial intelligence and data mining for the pharmacovigilance of drug–drug interactions. *Clinical Therapeutics*, 45(2):117–133, 2023.

[42] Seok-Jae Heo and Inkyung Jung. Extended multi-item gamma poisson shrinker methods based on the zero-inflated poisson model for postmarket drug safety surveillance. *Stat. Med.*, 39(30):4636–4650, December 2020.

[43] A B Hill. The environment and disease: Association or causation? *Proc. R. Soc. Med.*, 58(5):295–300, May 1965.

[44] Jan Horsky, Elizabeth A Drucker, and Harley Z Ramelson. Accuracy and completeness of clinical coding using ICD-10 for ambulatory visits. *AMIA Annu. Symp. Proc.*, 2017:912–920, 2017.

[45] Qiaozhi Hu, Zhou Qin, Mei Zhan, Zhaoyan Chen, Bin Wu, and Ting Xu. Validating the chinese geriatric trigger tool and analyzing adverse drug event associated risk factors in elderly chinese patients: A retrospective review. *PLoS One*, 15(4):e0232095, April 2020.

[46] Qiaozhi Hu, Bin Wu, Jinhui Wu, and Ting Xu. Predicting adverse drug events in older inpatients: a machine learning study. *Int. J. Clin. Pharm.*, 44(6):1304–1311, December 2022.

[47] Sara Hult, Daniele Sartori, Tomas Bergvall, Sara Hedfors Vidlin, Birgitta Grundmark, Johan Ellenius, and G Niklas Norén. A feasibility study of drug-drug interaction signal detection in regular pharmacovigilance. *Drug Saf.*, 43(8):775–785, August 2020.

[48] J Stuart Hunter. Enhancing friedman's "fundamental theorem of biomedical informatics". *J. Am. Med. Inform. Assoc.*, 17(1):112; author reply 112–3, January 2010.

[49] Srinivasan V Iyer, Rave Harpaz, Paea LePendu, Anna Bauer-Mehren, and Nigam H Shah. Mining clinical text for signals of adverse drug-drug interactions. *J. Am. Med. Inform. Assoc.*, 21(2):353–362, March 2014.

[50] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.

[51] Eugene Jeong, Namgi Park, Young Choi, Rae Woong Park, and Dukyong Yoon. Correction: Machine learning model combining features from algorithms with different analytical methodologies to detect laboratory-event-related adverse drug reaction signals. *PLoS One*, 14(4):e0215344, April 2019.

[52] Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-Wei H Lehman, Leo A Celi, and Roger G Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data*, 10(1):1, January 2023.

[53] Pentti Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognit. Comput.*, 1(2):139–159, June 2009.

[54] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training, 2021.

[55] Najoung Kim and Sebastian Schuster. Entity tracking in language models, 2023.

[56] Denis Kleyko, Mike Davies, E Paxon Frady, Pentti Kanerva, Spencer J Kent, Bruno A Olshausen, Evgeny Osipov, Jan M Rabaey, Dmitri A Rachkovskij, Abbas Rahimi, and Friedrich T Sommer. Vector symbolic architectures as a computing framework for emerging hardware. *Proc. IEEE Inst. Electr. Electron. Eng.*, 110(10):1538–1571, October 2022.

[57] Johannes Knittel, Tushaar Gangavarapu, Hendrik Strobelt, and Hanspeter Pfister. Gpt-2 through the lens of vector symbolic architectures, 2024.

[58] Elpida Kontsioti, Simon Maskell, Isobel Anderson, and Munir Pirmohamed. Identifying drug-drug interactions in spontaneous reports utilizing signal detection and biological plausibility aspects. *Clin. Pharmacol. Ther.*, 116(1):165–176, July 2024.

[59] Elpida Kontsioti, Simon Maskell, Amina Bensalem, Bhaskar Dutta, and Munir Pirmohamed. Similarity and consistency assessment of three major online drug-drug interaction resources. *Br. J. Clin. Pharmacol.*, 88(9):4067–4079, September 2022.

[60] Elpida Kontsioti, Simon Maskell, Bhaskar Dutta, and Munir Pirmohamed. A reference set of clinically relevant adverse drug-drug interactions. *Sci. Data*, 9(1):72, March 2022.

[61] Elpida Kontsioti, Simon Maskell, and Munir Pirmohamed. Exploring the impact of design criteria for reference sets on performance evaluation of signal detection algorithms: The case of drug-drug interactions. *Pharmacoepidemiol. Drug Saf.*, 32(8):832–844, August 2023.

[62] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfie Baston, Jack Ross, Esther Idowu, James T Teo, and Richard J Dobson. Foresight – generative pretrained transformer (gpt) for modelling of patient timelines using ehrs, 2023.

[63] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, James T Teo, and Richard J B Dobson. Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit. Health*, 6(4):e281–e290, April 2024.

[64] M. Kulldorff, I. Dashevsky, T. R. Avery, A. K. Chan, R. L. Davis, D. Graham, R. Platt, S. E. Andrade, D. Boudreau, M. J. Gunter, L. J. Herrinton, P. A. Pawloski, M. A. Raebel, D. Roblin, and J. S. Brown. Drug safety data mining with a tree-based scan statistic. *Pharmacoepidemiol Drug Saf*, 22(5):517–523, May 2013.

[65] M. Kulldorff, Z. Fang, and S. J. Walsh. A tree-based scan statistic for database disease surveillance. *Biometrics*, 59(2):323–331, Jun 2003.

[66] Martin Kulldorff, Zixing Fang, and Stephen J Walsh. A tree-based scan statistic for database disease surveillance. *Biometrics*, 59(2):323–331, June 2003.

[67] Hervé Le Louët and Peter J Pitts. Twenty-first century global ADR management: A need for clarification, redesign, and coordinated action. *Ther. Innov. Regul. Sci.*, 57(1):100–103, January 2023.

[68] Matthew E Levine, David J Albers, and George Hripcsak. Methodological variations in lagged regression for detecting physiologic drug effects in EHR data. *J. Biomed. Inform.*, 86:149–159, October 2018.

[69] Y. Li, J. Li, J. He, and C. Tao. AE-GPT: Using Large Language Models to extract adverse events from surveillance reports-A use case with influenza vaccine adverse events. *PLoS One*, 19(3):e0300919, 2024.

[70] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Sci. Rep.*, 10(1):7155, April 2020.

[71] Ying Li, Patrick B Ryan, Ying Wei, and Carol Friedman. A method to combine signals from spontaneous reporting systems and observational healthcare data to detect adverse drug reactions. *Drug Saf.*, 38(10):895–908, October 2015.

[72] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, 2023.

[73] Shenggeng Lin, Yanjing Wang, Lingfeng Zhang, Yanyi Chu, Yatong Liu, Yitian Fang, Mingming Jiang, Qiankun Wang, Bowen Zhao, Yi Xiong, and Dong-Qing Wei. MDF-SA-DDI: predicting drug-drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. *Brief. Bioinform.*, 23(1), January 2022.

[74] Fang Liu and Demosthenes Panagiotakos. Correction: Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med. Res. Methodol.*, 23(1):109, May 2023.

[75] Tal Lorberbaum, Kevin J Sampson, Jeremy B Chang, Vivek Iyer, Raymond L Woosley, Robert S Kass, and Nicholas P Tatonetti. Coupling data mining and laboratory experiments to discover drug interactions causing QT prolongation. *J. Am. Coll. Cardiol.*, 68(16):1756–1764, October 2016.

[76] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[77] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), September 2022.

[78] Scott A. Malec, Peng Wei, Elmer V. Bernstam, Richard D. Boyce, and Trevor Cohen. Using computable knowledge mined from the literature to elucidate confounders for ehr-based pharmacovigilance. *Journal of Biomedical Informatics*, 117:103719, 2021.

[79] J C McElnay, C R McCallion, F Al-Deagi, and M G Scott. Development of a risk model for adverse drug events in the elderly. *Clin. Drug Investig.*, 13(1):47–55, January 1997.

[80] Sean McLeish, Arpit Bansal, Alex Stein, Neel Jain, John Kirchenbauer, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, Jonas Geiping, Avi Schwarzschild, and Tom Goldstein. Transformers can do arithmetic with the right embeddings, 2024.

[81] Gina Mejía, Miriam Saiz-Rodríguez, Beatriz Gómez de Olea, Dolores Ochoa, and Francisco Abad-Santos. Urgent hospital admissions caused by adverse drug reactions and medication errors-a population-based study in spain. *Front. Pharmacol.*, 11:734, May 2020.

[82] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.

[83] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[84] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.*, 6:26094, May 2016.

[85] J. Mower, E. Bernstam, H. Xu, S. Myneni, D. Subramanian, and T. Cohen. Improving Pharmacovigilance Signal Detection from Clinical Notes with Locality Sensitive Neural Concept Embeddings. *AMIA Jt Summits Transl Sci Proc*, 2022:349–358, 2022.

[86] Rachel M Murphy, Joanna E Klopotowska, Nicolette F de Keizer, Kitty J Jager, Jan Hendrik Leopold, Dave A Dongelmans, Ameen Abu-Hanna, and Martijn C Schut. Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. *PLoS One*, 18(1):e0279842, January 2023.

[87] Azadeh Nikfarjam, Julia D Ransohoff, Alison Callahan, Erik Jones, Brian Loew, Bernice Y Kwong, Kavita Y Sarin, and Nigam H Shah. Early detection of adverse drug reactions in social health networks: A natural language processing pipeline for signal detection. *JMIR Public Health Surveill.*, 5(2):e11264, June 2019.

[88] Yoshihiro Noguchi, Tomoya Tachi, and Hitomi Teramachi. Detection algorithms and attentive points of safety signal using spontaneous reporting systems as a clinical data source. *Briefings in Bioinformatics*, 22(6):bbab347, 08 2021.

[89] Shanchen Pang, Ying Zhang, Tao Song, Xudong Zhang, Xun Wang, and Alfonso Rodriguez-Patón. AMDE: a novel attention-mechanism-based multidimensional feature encoder for drug-drug interaction prediction. *Brief. Bioinform.*, 23(1), January 2022.

[90] G. Park, H. Jung, S. J. Heo, and I. Jung. Comparison of Data Mining Methods for the Signal Detection of Adverse Drug Events with a Hierarchical Structure in Postmarketing Surveillance. *Life (Basel)*, 10(8), Aug 2020.

[91] Consuelo Pedrós, Beatriz Quintana, Mireia Rebolledo, Núria Porta, Antoni Vallano, and Josep Maria Arnau. Prevalence, risk factors and main features of adverse drug reactions leading to hospital admission. *Eur. J. Clin. Pharmacol.*, 70(3):361–367, March 2014.

[92] Bethany Percha, Yael Garten, and Russ B Altman. Discovery and explanation of drug-drug interactions via text mining. *Pac. Symp. Biocomput.*, pages 410–421, 2012.

[93] M Pirmohamed, A M Breckenridge, N R Kitteringham, and B K Park. Adverse drug reactions. *BMJ*, 316(7140):1295–1298, April 1998.

[94] T.A. Plate. Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6(3):623–641, 1995.

[95] T.A. Plate. Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6(3):623–641, 1995.

[96] Tony A. Plate. Vector representations + addition + multiplication = conceptual reasoning, 2020. Presented at the Plate Workshop, July 2020.

[97] Tony A. Plate. Mapping from ordinary embeddings to hyperdimensional spaces. YouTube Video, 2023. Available at `https://www.youtube.com/watch?v=k4ChL0dqI9c`.

[98] Tony Alexander Plate and Geoffrey Hinton. *Distributed representations and nested compositional structure*. PhD thesis, CAN, 1994. AAINN97247.

[99] J. Portanova, N. Murray, J. Mower, D. Subramanian, and T. Cohen. : Distributed Representations of Adverse Event Reporting System Data as a Means to Identify Drug/Side-Effect Associations. *AMIA Annu Symp Proc*, 2019:717–726, 2019.

[100] Sara K Quinney. Opportunities and challenges of using big data to detect drug-drug interaction risk. *Clin. Pharmacol. Ther.*, 106(1):72–74, July 2019.

[101] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018.

[102] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.*, 4(1):86, May 2021.

[103] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how BERT works. *CoRR*, abs/2002.12327, 2020.

[104] Patrick B. Ryan, David Madigan, Paul E. Stang, J. Marc Overhage, Judith A. Racoosin, and Abraham G. Hartzema. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership. *Statistics in Medicine*, 31(30):4401–4415, 2012.

[105] Jon Sánchez-Valle, Rion Brattig Correia, Marta Camacho-Artacho, Rosalba Lepore, Mauro M Mattos, Luis M Rocha, and Alfonso Valencia. Prevalence and differences in the co-administration of drugs known to interact: an analysis of three distinct and large populations. *BMC Med.*, 22(1):166, April 2024.

[106] Kenny Schlegel, Peer Neubert, and Peter Protzel. A comparison of vector symbolic architectures. *Artificial Intelligence Review*, 55(6):4523–4555, December 2021.

[107] Kyriakos Schwarz, Ahmed Allam, Nicolas Andres Perez Gonzalez, and Michael Krauthammer. AttentionDDI: Siamese attention-based deep learning method for drug-drug interaction predictions. *BMC Bioinformatics*, 22(1):412, August 2021.

[108] Yijun Shao, Yan Cheng, Stuart J. Nelson, Peter Kokkinos, Edward Y. Zamrini, Ali Ahmed, and Qing Zeng-Treitler. Hybrid value-aware transformer architecture for joint learning from longitudinal and non-longitudinal clinical data. *medRxiv*, 2023.

[109] Edward H Shortliffe. The organization and content of informatics doctoral dissertations. *J. Am. Med. Inform. Assoc.*, 23(4):840–843, July 2016.

[110] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022.

[111] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.

[112] Xiaorui Su, Lun Hu, Zhuhong You, Pengwei Hu, and Bowei Zhao. Attention-based knowledge graph representation learning for predicting drug-drug interactions. *Brief. Bioinform.*, 23(3), May 2022.

[113] Ana Szarfman, Stella G Machado, and Robert T O'Neill. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Saf.*, 25(6):381–392, 2002.

[114] Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting?, 2024.

[115] N P Tatonetti, J C Denny, S N Murphy, G H Fernald, G Krishnan, V Castro, P Yue, P S Tsao, I Kohane, D M Roden, and R B Altman. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clin. Pharmacol. Ther.*, 90(1):133–142, July 2011.

[116] Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493, 01 2024.

[117] Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding, 2023.

[118] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[119] Craig A Umscheid, David J Margolis, and Craig E Grossman. Key concepts of clinical trials: a narrative review. *Postgrad. Med.*, 123(5):194–204, September 2011.

[120] Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, and Laura I. Furlong. The eu-adr corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884, 2012. Text Mining and Natural Language Processing in Pharmacogenomics.

[121] Gail A Van Norman. Drugs, devices, and the FDA: Part 1: An overview of approval processes for drugs. *JACC Basic Transl. Sci.*, 1(3):170–179, April 2016.

[122] E P Van Puijenbroek, A C Egberts, R H Meyboom, and H G Leufkens. Signalling possible drug-drug interactions in a spontaneous reporting system: delay of withdrawal bleeding during concomitant use of oral contraceptives and itraconazole. *Br. J. Clin. Pharmacol.*, 47(6):689–693, June 1999.

[123] Eugène P van Puijenbroek, Andrew Bate, Hubert G M Leufkens, Marie Lindquist, Roland Orre, and Antoine C G Egberts. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol. Drug Saf.*, 11(1):3–10, January 2002.

[124] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[125] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[126] Michael A Veronin, Robert P Schumaker, and Rohit Dixit. The irony of MedWatch and the FAERS database: An assessment of data input errors and potential consequences. *J. Pharm. Technol.*, 36(4):164–167, August 2020.

[127] Ulrich Vogel, John van Stekelenborg, Brian Dreyfus, Anju Garg, Marian Habib, Romana Hosain, and Antoni Wisniewski. Investigating overlap in signals from EVDAS, FAERS, and VigiBase®. *Drug Saf.*, 43(4):351–362, April 2020.

[128] E. A. Voss, R. D. Boyce, P. B. Ryan, J. van der Lei, P. R. Rijnbeek, and M. J. Schuemie. Accuracy of an automated knowledge base for identifying drug adverse reactions. *J Biomed Inform*, 66:72–81, Feb 2017.

[129] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do nlp models know numbers? probing numeracy in embeddings, 2019.

[130] Jing Wang, Shuo Zhang, Runzhi Li, Gang Chen, Siyu Yan, and Lihong Ma. Multi-view feature representation and fusion for drug-drug interactions prediction. *BMC Bioinformatics*, 24(1):93, March 2023.

[131] Shirley V Wang, Judith C Maro, Elande Baro, Rima Izem, Inna Dashevsky, James R Rogers, Michael Nguyen, Joshua J Gagne, Elisabetta Patorno, Krista F Huybrechts, Jacqueline M Major, Esther Zhou, Megan Reidy, Austin Cosgrove, Sebastian Schneeweiss, and Martin Kulldorff. Data mining for adverse drug events with a propensity score-matched tree-based scan statistic. *Epidemiology*, 29(6):895–903, November 2018.

[132] Dominic Widdows and Trevor Cohen. Graded semantic vectors: An approach to representing graded quantities in generalized quantum models. volume 9535, pages 231–244, 01 2016.

[133] R. L. Woosley and K. Romero. Assessing cardiovascular drug safety for clinical decision-making. *Nat Rev Cardiol*, 10(6):330–337, Jun 2013. [DOI:https://dx.doi.org/10.1038/nrcardio.2013.5710.1038/nrcardio.2013.57] [PubMed:https://www.ncbi.nlm.nih.gov/pubmed/86019238601923].

[134] Patrick Wu, Scott D Nelson, Juan Zhao, Cosby A Stone, Jr, Qiping Feng, Qingxia Chen, Eric A Larson, Bingshan Li, Nancy J Cox, C Michael Stein, Elizabeth J Phillips, Dan M Roden, Joshua C Denny, and Wei-Qi Wei. DDIWAS: High-throughput electronic health record-based screening of drug-drug interactions. *J. Am. Med. Inform. Assoc.*, 28(7):1421–1430, July 2021.

[135] Yifan Wu, Justin Mower, Xiruo Ding, Oliver Li, Devika Subramanian, and Trevor Cohen. Predicting drug blood-brain barrier penetration with adverse event report embeddings. *AMIA Annu. Symp. Proc.*, 2022:1163–1172, 2022.

[136] Su Xian, Monika E Grabowska, Iftikhar J Kullo, Yuan Luo, Jordan W Smoller, Wei-Qi Wei, Gail Jarvik, Sean Mooney, and David Crosslin. Language-model-based patient embedding using electronic health records facilitates phenotyping, disease forecasting, and progression analysis. *Res. Sq.*, September 2024.

[137] Chengqi Xu, Krishna C. Bulusu, Heng Pan, and Olivier Elemento. Ddi-gpt: Explainable prediction of drug-drug interactions using large language models enhanced with knowledge graphs. *bioRxiv*, 2024.

[138] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. A large language model for electronic health records. *NPJ Digit. Med.*, 5(1):194, December 2022.

[139] Izak A R Yasrebi-de Kom, Dave A Dongelmans, Nicolette F de Keizer, Kitty J Jager, Martijn C Schut, Ameen Abu-Hanna, and Joanna E Klopotowska. Electronic health record-based prediction models for in-hospital adverse drug event diagnosis or prognosis: a systematic review. *J. Am. Med. Inform. Assoc.*, 30(5):978–988, April 2023.

[140] Liyi Yu, Zhaochun Xu, Meiling Cheng, Weizhong Lin, Wangren Qiu, and Xuan Xiao. MSEDDI: Multi-scale embedding for predicting drug-drug interaction events. *Int. J. Mol. Sci.*, 24(5), February 2023.

[141] Ting-He Zhang, Md Musaddaqul Hasib, Yu-Chiao Chiu, Zhi-Feng Han, Yu-Fang Jin, Mario Flores, Yidong Chen, and Yufei Huang. Transformer for gene expression modeling (T-GEM): An interpretable deep learning model for gene expression-based phenotype predictions. *Cancers (Basel)*, 14(19):4763, September 2022.

[142] Yang Zhang, Caiqi Liu, Mujiexin Liu, Tianyuan Liu, Hao Lin, Cheng-Bing Huang, and Lin Ning. Attention is all you need: utilizing attention in ai-enabled drug discovery. *Briefings in Bioinformatics*, 25(1):bbad467, 01 2024.

[143] Yuanyuan Zhang, Zengqian Deng, Xiaoyu Xu, Yinfei Feng, and Shang Junliang. Application of artificial intelligence in drug-drug interactions prediction: A review. *J. Chem. Inf. Model.*, 64(7):2158–2173, April 2024.

[144] Jing Zhao, Aron Henriksson, Maria Kvist, Lars Asker, and Henrik Boström. Handling temporality of clinical events for drug safety surveillance. *AMIA Annu. Symp. Proc.*, 2015:1371–1380, November 2015.

[145] Jing Zhao, Panagiotis Papapetrou, Lars Asker, and Henrik Boström. Learning from heterogeneous temporal data in electronic health records. *J. Biomed. Inform.*, 65:105–119, January 2017.

[146] Wei Zhao, Kris Sachsenmeier, Lanju Zhang, Erin Sult, Robert E Hollingsworth, and Harry Yang. A new bliss independence model to analyze drug combination data. *J. Biomol. Screen.*, 19(5):817–821, June 2014.

[147] Yunpeng Zhao, Xing He, Zheng Feng, Sarah Bost, Mattia Prosperi, Yonghui Wu, Yi Guo, and Jiang Bian. Biases in using social media data for public health surveillance: A scoping review. *International Journal of Medical Informatics*, 164:104804, 2022.

Appendix A

APPENDIX

The parameters used to generate the pretrained embeddings. The semvecpy can be found on Github[1].

Table A.1: Parameters for Building a Positional Index with Semantic Vectors.

Parameter	**Value**
-luceneindexpath	semanticvectors/positional_index
-encodingmethod	embeddings
-windowradius	5
-minfrequency	1
-dimension	768
-seedlength	768
-docindexing	none
-numthreads	64
-trainingcycles	4
-positionalmethod	directional
-notnormalized	true

[1]https://github.com/semanticvectors/semvecpy

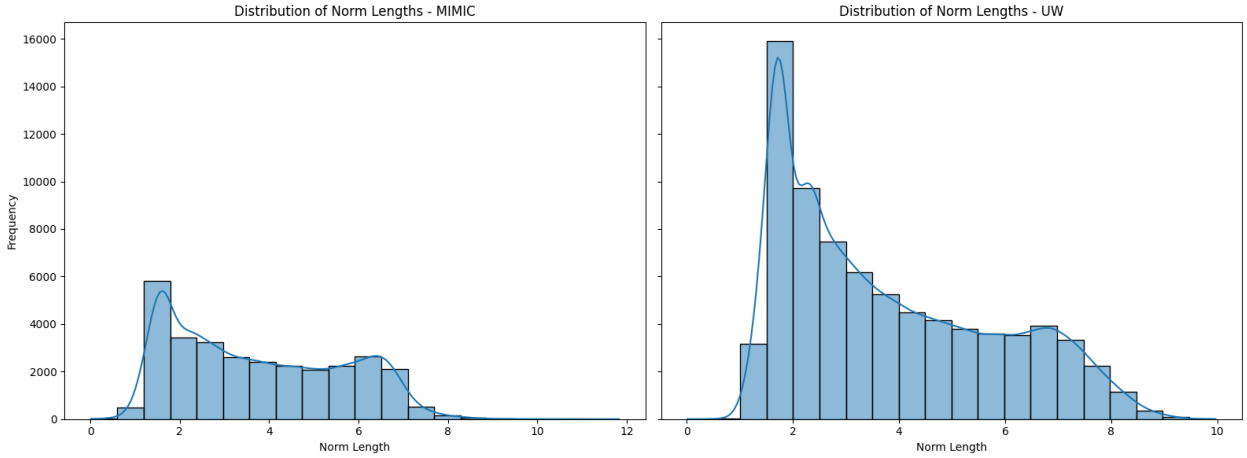

Figure A.1: Distribution of Norm Length of UW and MIMIC

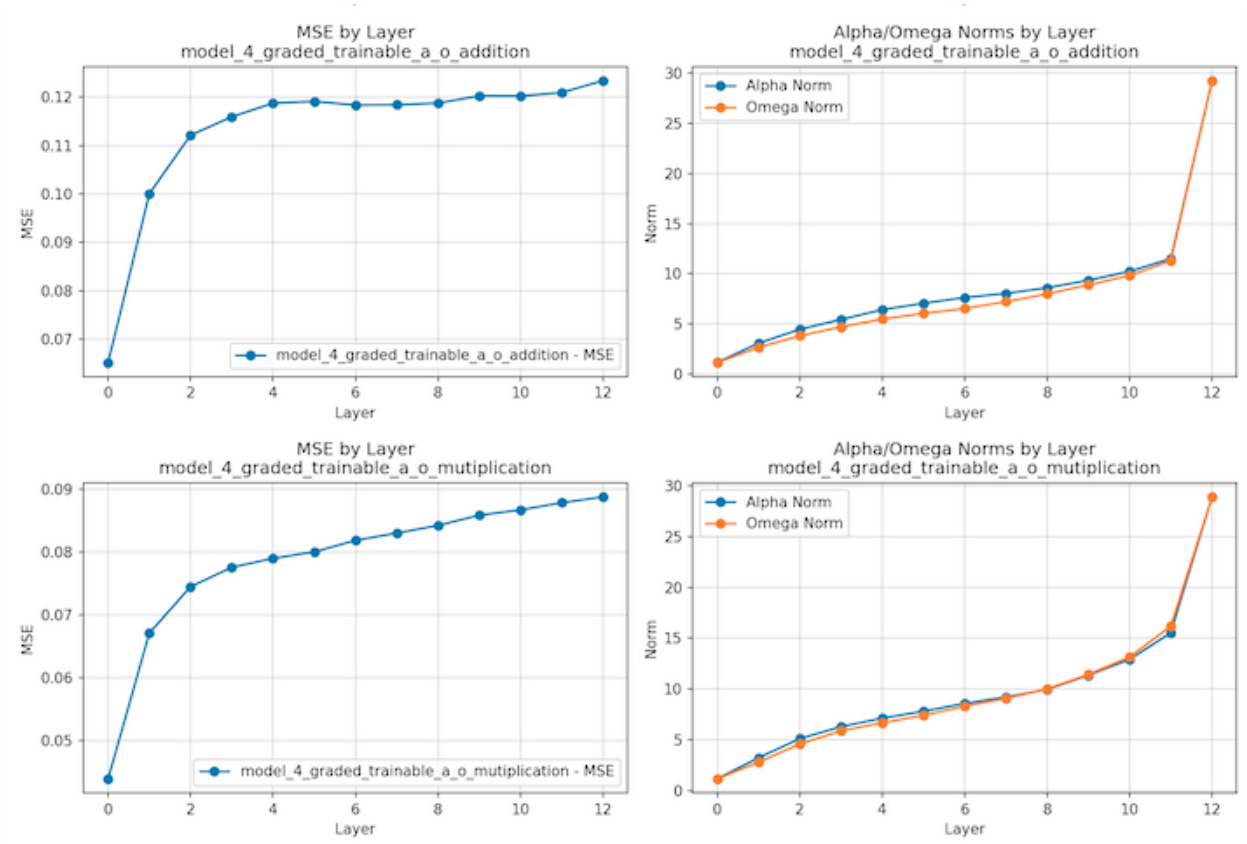

Figure A.2: MSE change across different layers for models with graded value-aware embeddings, MIMIC

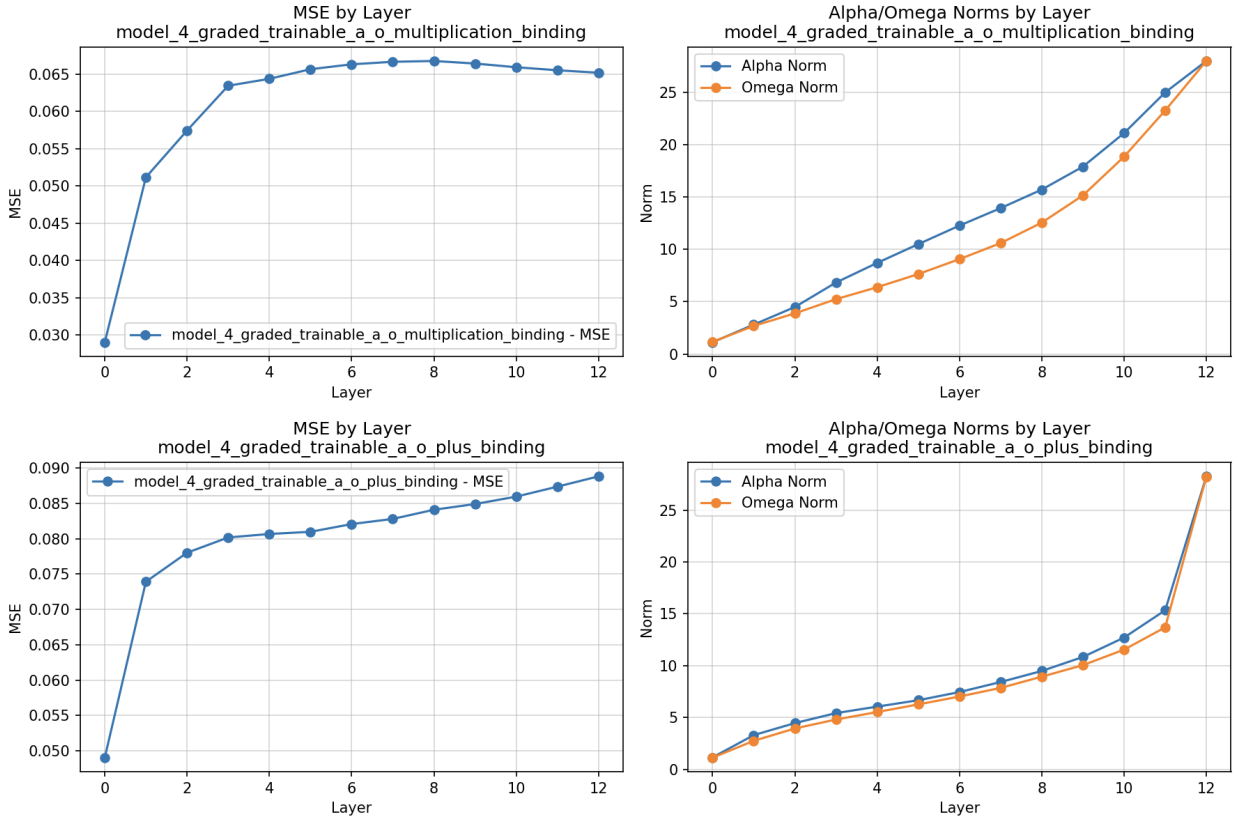

Figure A.3: MSE change across different layers for models with graded value-aware embeddings, UW

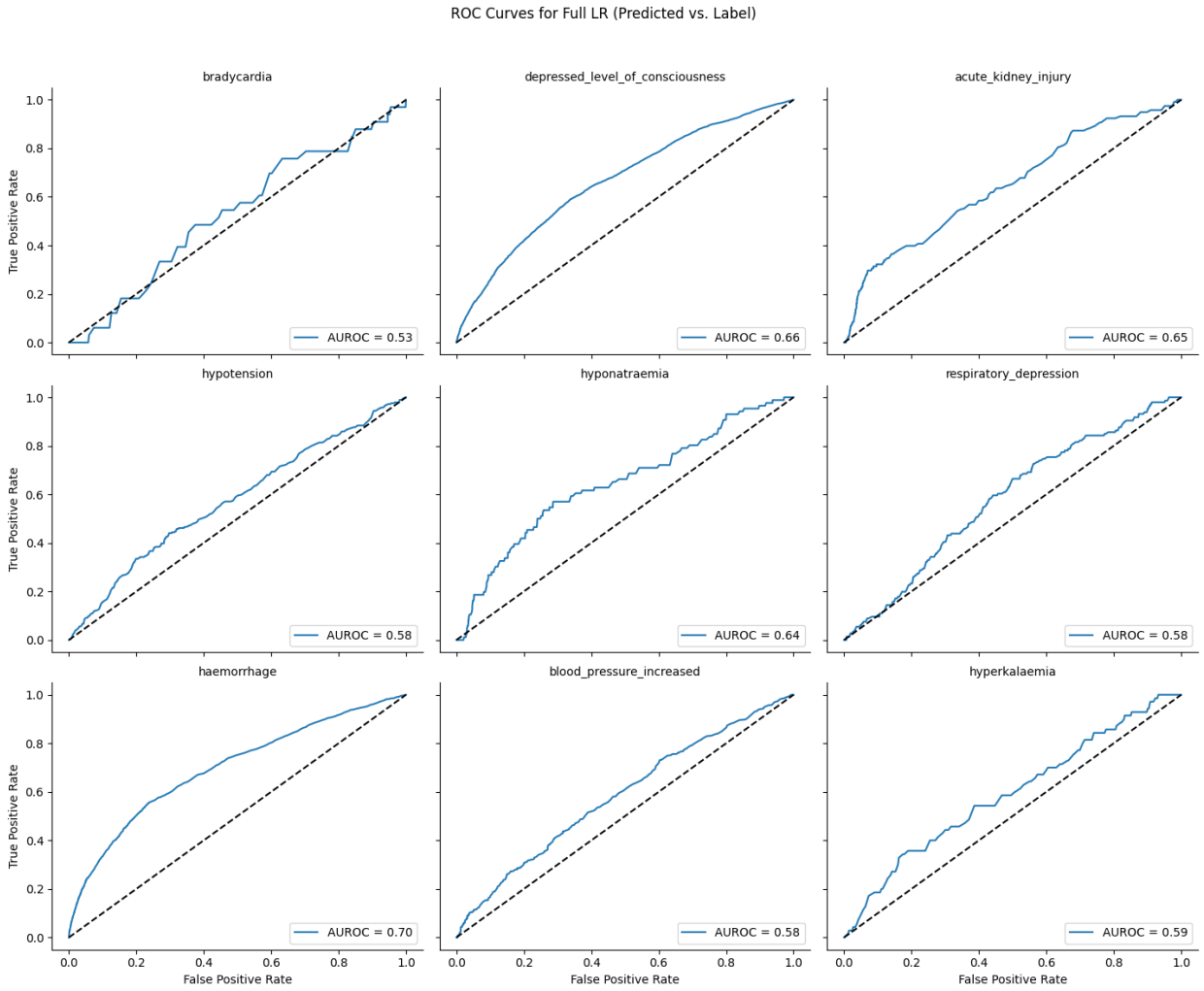

Figure A.4: Logistic Regression with full dataset, breakdown, MIMIC

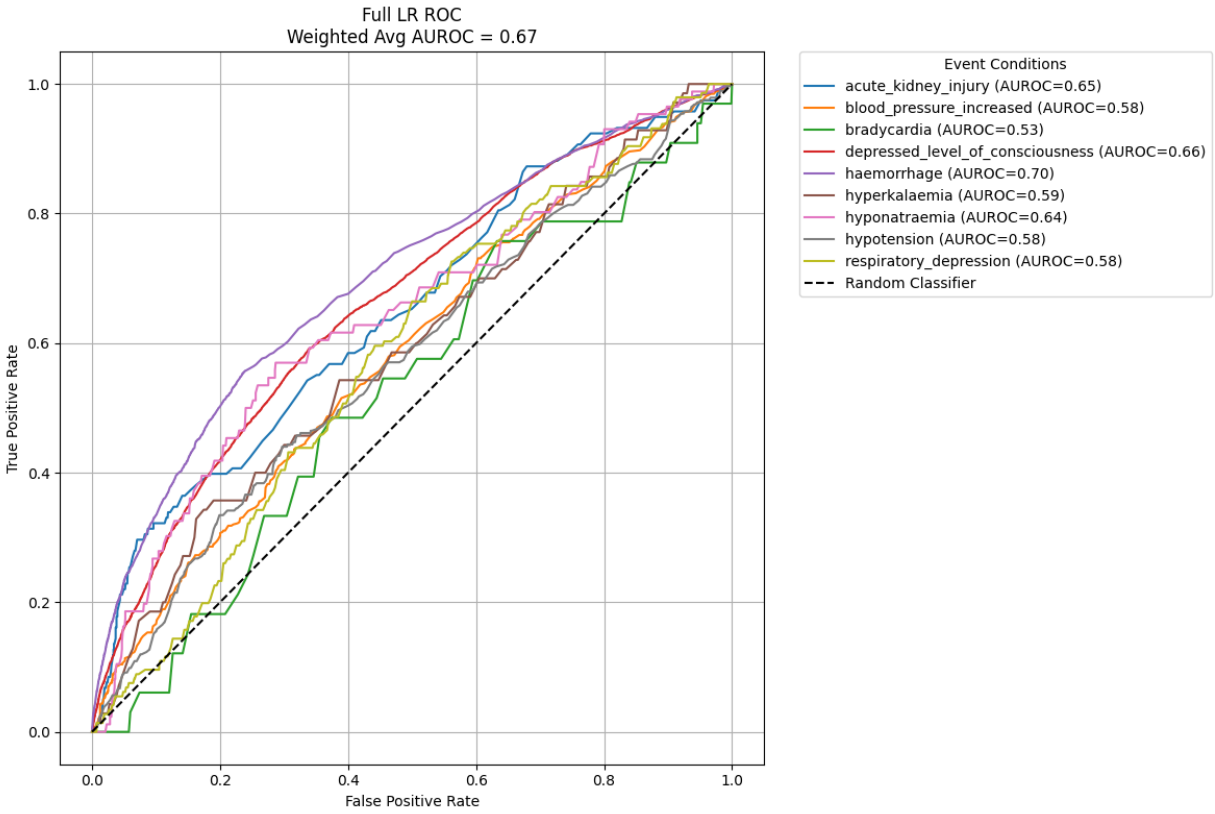

Figure A.5: Logistic Regression with full dataset, overall, MIMIC

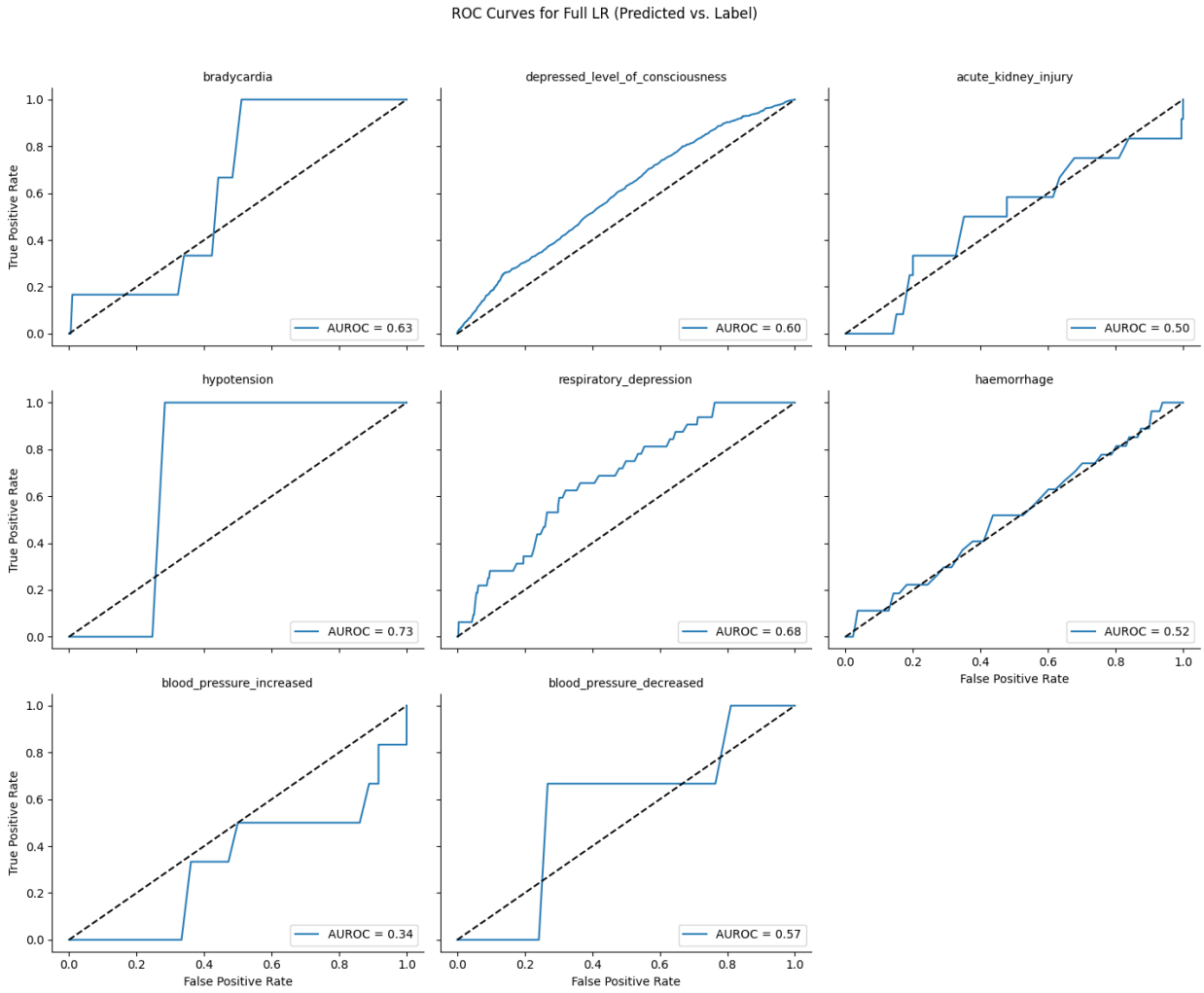

Figure A.6: Logistic Regression with full dataset, breakdown, UW

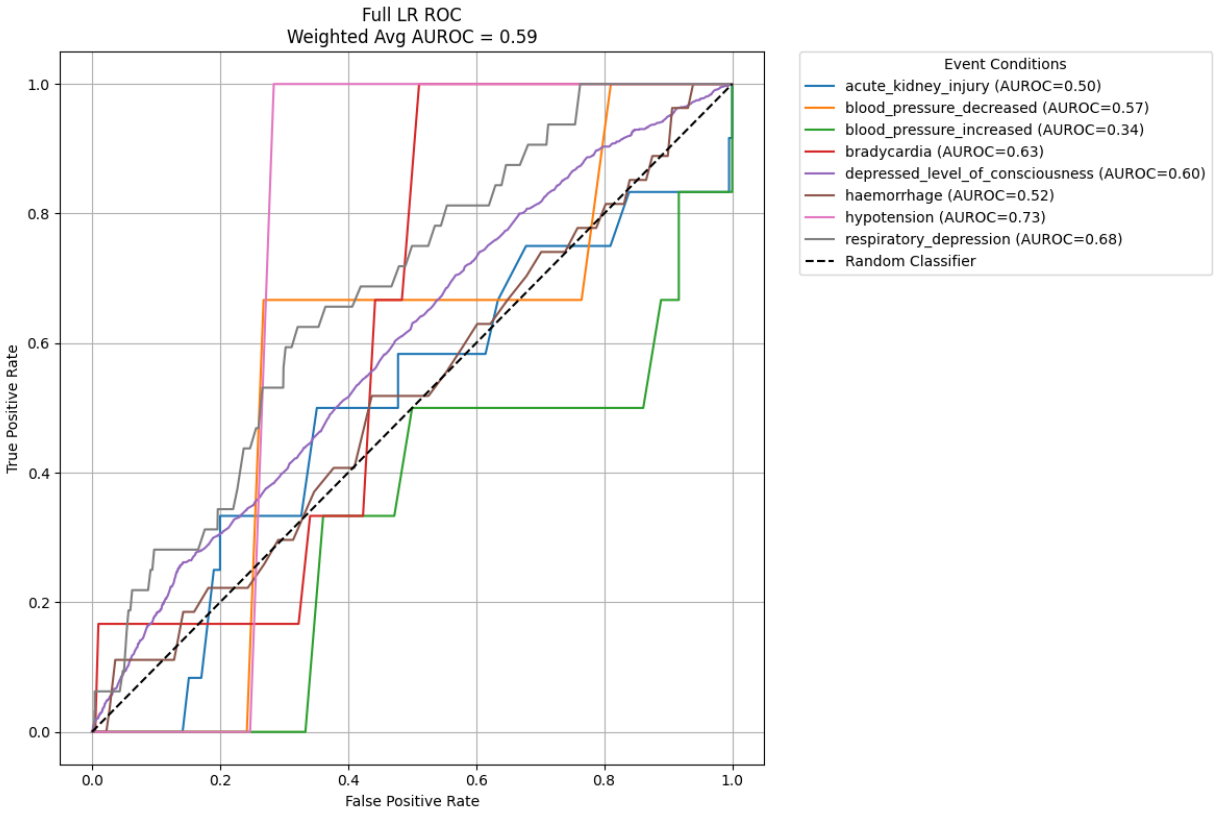

Figure A.7: Logistic Regression with full dataset, overall, UW

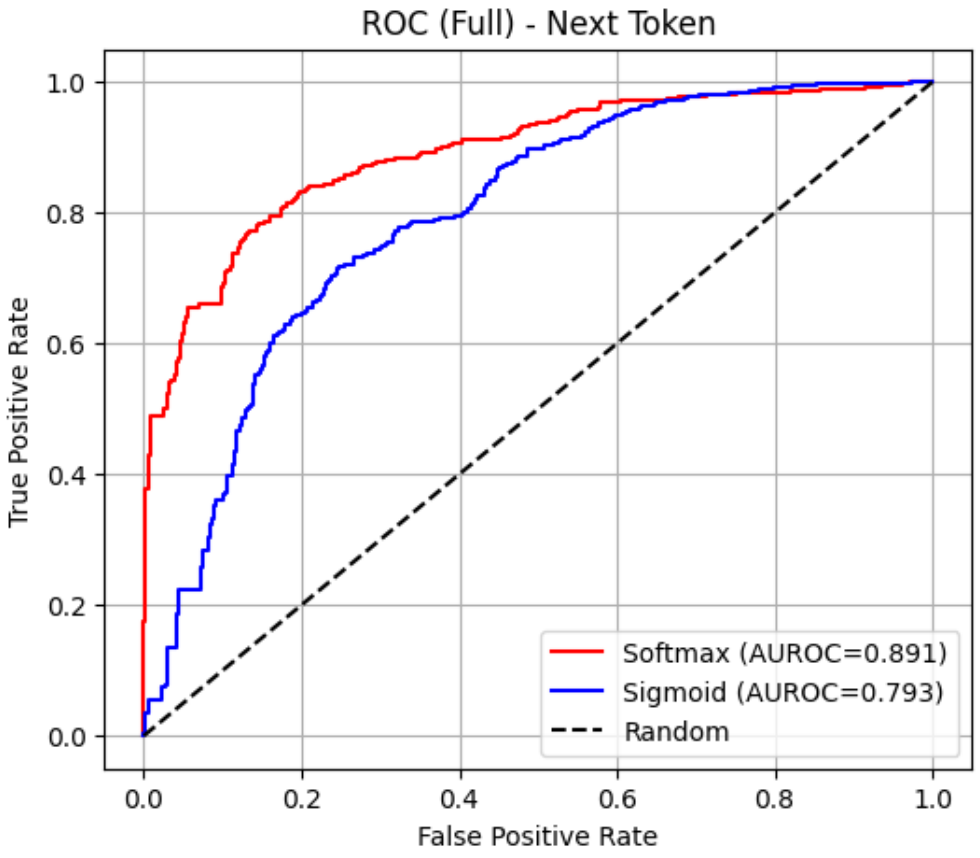

Figure A.8: GPT-2 based model, next token prediction method, Softmax vs Sigmoid, MIMIC

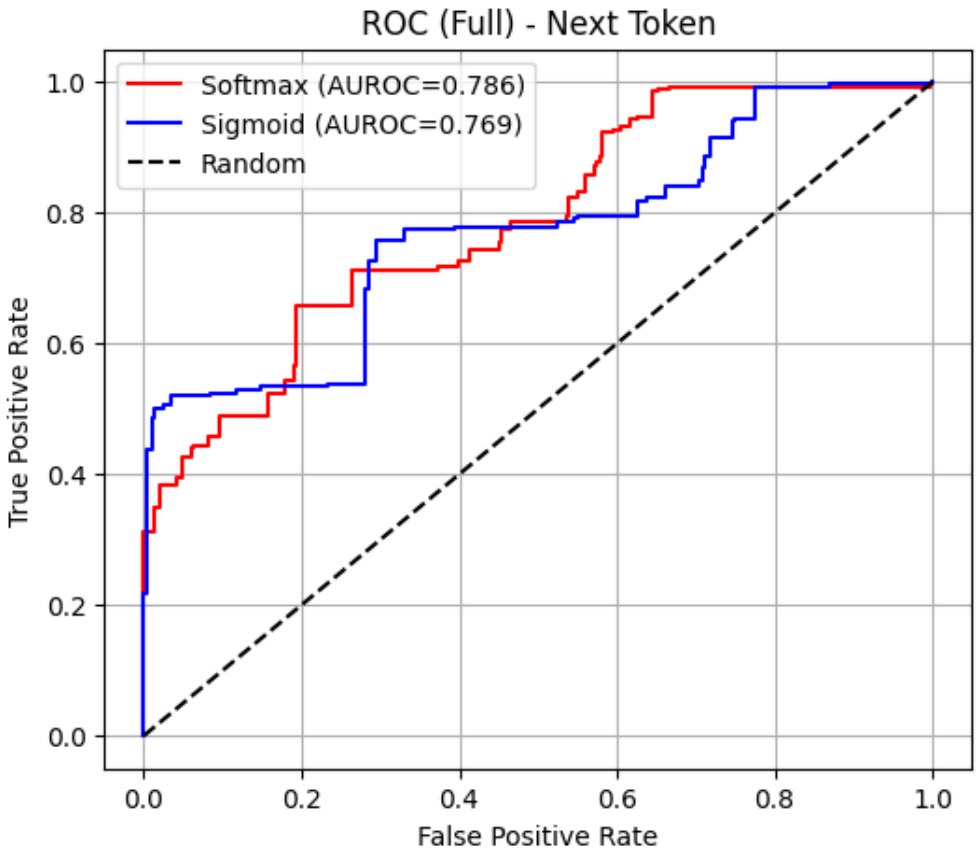

Figure A.9: GPT-2 based model, next token prediction method, Softmax vs Sigmoid, UW

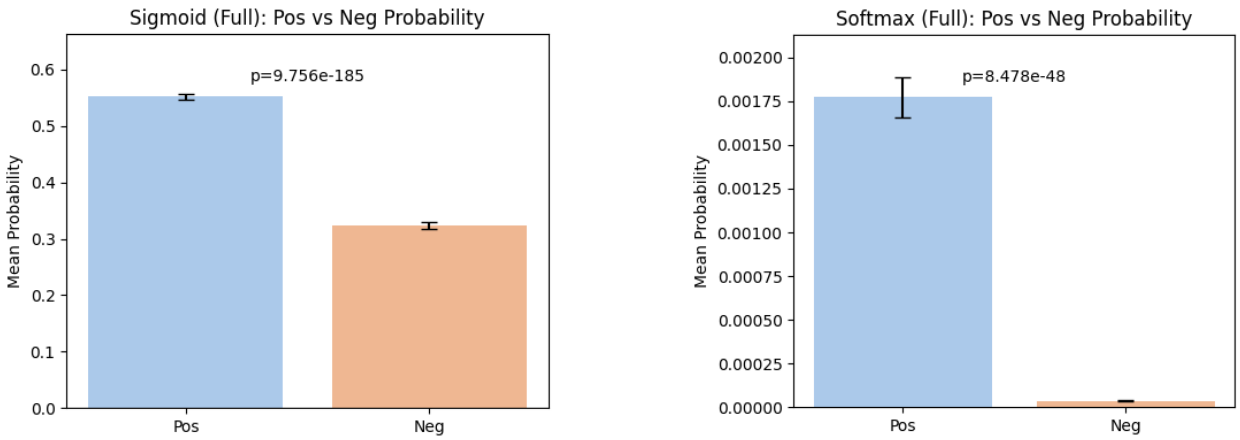

Figure A.10: Comparison of the sigmoid (left) and softmax (right) probabilities of positive versus negative drug pairs using full MIMIC dataset.

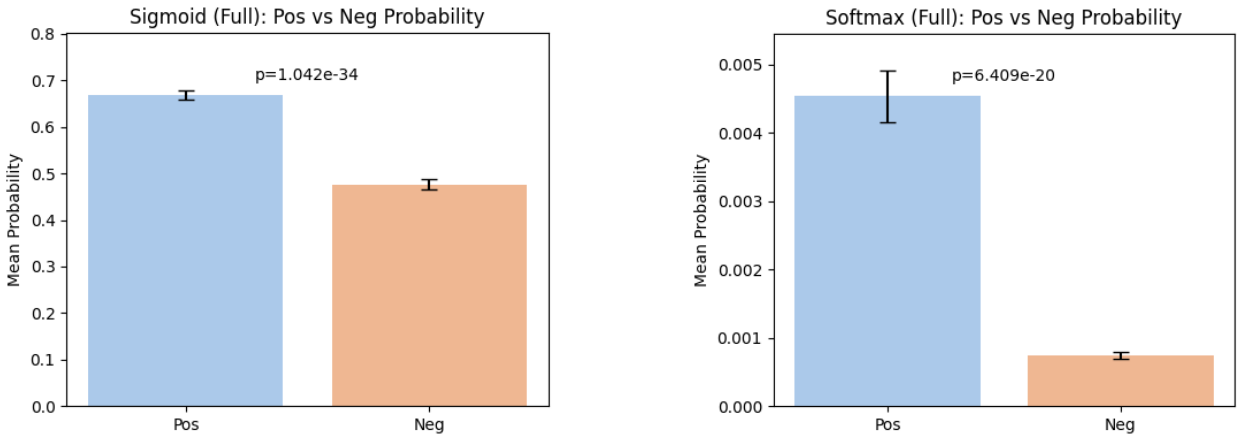

Figure A.11: Comparison of the sigmoid (left) and softmax (right) probabilities of positive versus negative drug pairs using full UW dataset.

Table A.2: Hyperparameters Used for Training the GPT-2 Model

Parameter	**Value/Description**
Model architecture	GPT-2 (custom, trained from scratch)
Embedding type	Word Token Embeddings/Permutation embeddings
Sequence length	Last 512 tokens (processed in reverse order)
Batch size (effective)	16 (gradient accumulation over 16 steps)
Learning rate	6×10^{-6}
Weight decay	0.1
Warm-up steps	2000
Number of epochs	8
Optimizer	AdamW
Training precision	Mixed precision (FP16)
Early stopping patience	2 evaluation steps
Padding token ID	0
Maximum sequence length (tokenized)	1024 tokens
Evaluation metric	Validation loss

VITA

YiFan Wu grew up in Shanghai, China. She went to Shanghai Foreign Language School before moving to the USA. She studied biology and math at St Mary's college of Maryland. She earned a master's degree in biostatistics from University of Michigan, Ann Arbor. She moved to Seattle to pursue a PhD degree in biomedical and health informatics at the University of Washington.