'Effects Or Efficacy?' Enhancing Cancer Pharmacovigilance with Multi-Label Corpora and Generative Models

***Abstract*—Cancer encompasses a diverse range of conditions, each necessitating a customized treatment approach for individual patients. While cancer treatment drugs are effective, their adverse drug reactions (ADR) are often underreported. With the rapid expansion of data sources such as social media, biomedical literature, and Electronic Medical Records, it is essential to extract ADR-related information from these unstructured texts efficiently. In this work, we introduce a novel multi-labeled dataset, *CanExpance*, sourced from healthcare forums and blogs, comprising 3,011 records of ADR associated with cancer treatment. Additionally, we propose a multi-task generative pipeline, the *Cancer Adverse Drug Identification Framework (CanADI)*, designed to generate ADR, severity, and adversity labels from user posts and drug usage during treatment. Our end-to-end approach is two-fold: (i) We benchmark the *CanExpance* dataset using various large language models (LLMs) by evaluating the impact of instruction tuning through zero-shot, few-shot, and chain-of-thought prompting. (ii) We develop *CanADI*, a multi-task generative framework, incorporating Low-Rank Adaptation (LoRA) modules on the Query, Key, and Value matrices of the multi-head Attention module in LLaMA 2. We conducted extensive evaluations, both automated and manual, of the performance of various LLMs using BLEU and ROUGE scores to assess the generation of ADR responses related to specific drugs. Furthermore, we examined three different prompting strategies (zero-shot, few-shot, and chain-of-thought) to evaluate the generative capabilities of the *CanADI* framework. This multifaceted approach provides a comprehensive understanding of LLM capabilities in cancer pharmacovigilance.**

***Impact Statement*—The limited focus on adverse drug reactions (ADR) in cancer treatment, compared to drug efficacy, has created a significant gap in pharmacovigilance. Addressing this key issue, we introduced *CanExpance*, a multi-labeled dataset thoroughly curated from healthcare forums, documenting patient-reported drug reactions linked to various chemotherapy drugs. Leveraging this dataset, we created the *Cancer Adverse Drug Identification framework (CanADI)*, which detects adverse drug reactions from user posts and assesses their adversity and severity levels. We believe that our work advances healthcare by improving the detection and analysis of ADR, with broad implications across several levels. The adverse drug identification framework has the potential to influence regulatory policies by providing more detailed information on drug safety. Furthermore, it could reduce healthcare costs associated with managing severe drug reactions. In the long run, our work establishes a foundation for safer and more effective cancer treatment strategies, making a lasting impact on oncology pharmacovigilance.**

***Index Terms*—Pharmacovigilance, Adverse Drug Reactions, Classification, Large Language Models (LLMs), Prompting Techniques**

I. Introduction

PHARMACOVIGILANCE is the scientific field and set of activities that concentrate on detecting, evaluating, comprehending, and preventing any adverse effects or additional concerns associated with drugs. Over the past decade, cancer has emerged as the world's second-leading cause of death, causing a significant global health burden. As reported by the National Center for Health Statistics, figures for 2023 reported 1,958,310 new cancer cases and 609,820 cancer-related deaths in the United States alone. The increasing number of cases and mortality rates associated with cancer have resulted in rapid developments in treatment options. As we progress forward in an era of rapid medical advancements and an expanding range of treatment choices, it is essential to thoroughly address the potential risks associated with pharmaceutical treatments thereby enhancing cancer care and outcomes in the fight against this chronic disease. **Adverse drug reaction (ADR)** has a critical role in pharmacovigilance (PCV). An ADR is any harmful or unintended response to a medication that occurs at doses normally used for treatment. As a result, ADR is one of the key outcomes monitored within the scope of PCV.

Pharmacovigilance agencies, responsible for monitoring drug safety post-market release, utilize various surveillance systems to identify potential adverse drug events (ADEs). However, the primary dependence on passive spontaneous reporting system databases, such as the Federal Drug Administration's Adverse Event Reporting System (FAERS) [1], raises concerns about potential under-reporting, bias, and delayed data collection [2] within existing Adverse Drug Event (ADE) surveillance systems. Gathering and maintaining clinical evidence for PCV can be challenging because it requires labor-intensive manual collection of data on drugs [3]. To address the limitations of existing approaches, there are ADR monitoring methods that are constantly looking for updated ADR data sources [4]. Unstructured textual data, such as medical literature, EHR entries, and social media posts, contain a significant amount of relevant information. Natural language processing (NLP) methods for identifying and extracting adverse drug events from unstructured text have the potential to improve the efficiency of monitoring these different sources of information.

In our research, we have taken the initiative to bridge the gap between available data and cancer PCV by introducing *CanExpance*, a novel dataset containing critical information about drugs used in cancer therapy as well as patient experiences. Furthermore, we have built a generative framework capable of generating a multitasking output, such as identifying severity, adversity, and side effects of cancer drugs based on a user post and a specified drug name, which provided insight into the various aspects of cancer treatment. In addition, inspired by the introduction of large language models (LLMs), we have also utilized their capabilities through different prompting

techniques and built a multi-feature classification model using LLMs capable of predicting multiple labels from the input data.

a) Motivation and Research Questions: The demand for drugs in various medical fields is huge, and considering the significant toxicity of antineoplastic (medications used to treat cancer) agents, conducting PCV studies in oncology is critical. Moreover, PCV in oncology patients assists healthcare providers in identifying additional ADR outside of clinical trial settings. Despite their importance, oncology ADR are frequently underreported, often because patients believe that such side effects are unavoidable [5]. Large language models can understand and generate text similar to human language, indicating a significant advancement in natural language processing. Despite their potential, they are currently underutilized in pharmacovigilance (PCV). Researchers can use LLM's predictive and analytical capabilities to streamline the cancer drug development process, analyze complex patient data, and forecast potential adverse events in oncology more efficiently. In this work, we primarily investigate the following three research questions.

RQ1: What advantages do LLMs offer over traditional deep learning models in the context of pharmacovigilance?

RQ2: Does the clinical language model possess knowledge regarding ADR associated with cancer drugs?

RQ3: In the context of LLMs, do changing prompts influence the efficacy of models in generating accurate and relevant responses?

RQ4: How will collecting data from sources through self-reporting mechanisms for patients on medications improve PCV practices?

Societal Impact: Despite significant advances in drug research and development, the potential side effects of these medications frequently get underestimated. Cancer is a life-threatening illness that affects people of all ages and cultures. ADR can further compromise a patient's health and overall quality of life. Understanding the prevalence and severity of these drug reactions is essential for developing safer and more effective treatments. Our research aims to contribute to the field of cancer care by proposing a multi-feature classification and multi-task generation model. We believe that our current work will benefit both patients and healthcare professionals by providing valuable insights that can help develop safer and more effective cancer treatment strategies. To sum up, our **key contributions** include:

- We introduce ***CanExpance***, the first multi-labeled dataset created exclusively for **CAN**cer pharmacovigil**ANCE**, establishing a new benchmark for multi-feature classification and multitask generation.
- We propose a multi-task generative framework for cancer PCV that harnesses the unique capabilities of LLMs to generate adverse effects, severity, and adversity within the specified dataset.
- We validate our framework by assessing various prompting techniques utilized in LLMs, intending to determine the most efficient approach for addressing cancer PCV-related concerns.

II. Background

In recent years, the detection and assessment of drug reactions associated with cancer treatments have received a lot of attention because of their potential impact on patient safety and treatment outcomes. Prior research in this area has focused on a variety of methodologies in order to improve patient care and quality of life.

1) Social Media for PCV: The introduction of social media increased the spread of information across a wide variety of domains, including the medical sciences, by providing a platform for sharing insights and experiences. In contrast to clinical information retrieval from electronic health records (EHR), which has limited access [6], web data is freely available and provides an easy and attractive source of medical data. Patients are increasingly reaching out to social media platforms as a reliable source of health-related data, where they can learn about the common health-related experiences of other individuals. One of the most popular social media platforms is Twitter, where data and health-related information are shared in the form of posts and tweets between users [7]–[11]. Users can benefit from these posts by learning about other individuals diagnosed with the same disease.

Clinical Datasets : The current datasets, such as the PSB 2016 social media shared task dataset [12], the Medline ADE corpus [13], the CADEC dataset [14], and the BioDEX dataset [15], consist of adverse drug events (ADEs) across a wide range of clinical fields that are not specific to any particular medical condition or treatment.

2) Frameworks: Prior research in pharmacovigilance has focused mainly on solving individual tasks such as classification [20], extraction [21], and labeling of ADR [22], [23]. In recent years, with the introduction of deep learning techniques, the majority of studies have switched to deep learning models for predicting ADR. Lee et al. [24] created a semi-supervised deep learning model using a convolutional neural network (CNN) for classification based on the Twitter corpus. Similarly, Chowdhury et al. [25] proposed a multitask encoder-decoder framework for simultaneously modeling three ADR detection tasks: ADR classification, ADR labeling, and indication labeling. Furthermore, Biseda et al. [26] evaluated the performance of BERT [27] with two variants trained on biomedical papers: BioBERT [28] and Clinical BERT [29]. In this work, eight different BERT models were fine-tuned and tested on three different tasks to evaluate their relative performance in ADR tasks such as sentiment classification of drug reviews, identification of ADR presenct in Twitter posts, and named entity recognition of ADR within Twitter posts.

3) Pharmacovigilance in Oncology: Anti-cancer drugs, although extensively researched and proven highly effective in cancer treatment, should be approached with caution due to their high toxicity and narrow treatment window [30]. ADR in the field of oncology are common and often predictable, to an extent where they are recognized as an inherent part of the treatment process [31]. While the cancer drugs effectively target and treat various types of cancer, they also have the potential to cause ADR. These reactions range in severity, from mild and easily manageable to severe and requiring

TABLE I
STATISTICS OF EXISTING DATASETS FOR PHARMACOVIGILANCE

Dataset Name	Corpus Size	Language	Disease Specific	Severity	ADR Specific	Text Type
Gurulingappa et al. [16]	2972	English	✗	✗	✓	MEDLINE case reports
Oronoz et al. [17]	75	Spanish	✗	✗	✗	Clinical Discharge Reports
Patki et al. [18]	10,822	English	✗	✗	✓	Social Media Tweets
Li et al. [19]	1500	English	✗	✗	✗	PubMed Articles
CanExpance(Ours)	2996	English	✓	✓	✓	Medical Discussion Forums

hospitalization [32]. According to a review conducted in 2010 [33], which analyzed 95 articles, it was found that the failure to report adverse events accurately could lead to an increase in hospitalizations. Despite the growing interest in artificial intelligence (AI) and its application in safety data analysis, as proven by numerous studies and review articles [34], [35] there is a notable lack of systematic reviews that critically evaluate the potential improvement of PCV through AI within this framework.

As evident from the literature review, existing works have made significant contributions to the community but have certain limitations in addressing oncology-related PCV challenges. Previous research has primarily focused on social media data, frequently collected from Twitter. However, platforms like Twitter present inherent challenges due to users' diverse backgrounds and linguistic preferences, leading to ambiguous expressions that hinder PCV efforts. Additionally, the existing pharmacovigilance datasets are not specifically designed for cancer, which limits their effectiveness in monitoring ADR in cancer treatment. To address this limitation, we propose ***CanExpance***, the first multi-labeled dataset specific to cancer for enhancing pharmacovigilance in cancer care. Moreover, while machine learning and deep learning techniques are widely used in PCV, the application of LLMs in oncology remains largely unexplored. To bridge this gap, we propose the ***Cancer Adverse Drug Identification Framework (CanADI)***. *CanADI* generates information on adverse drug reactions associated with specific cancer drugs when users post queries, and it identifies the severity and adversity levels of the reported ADR.

III. CORPUS DESCRIPTION

Data plays a pivotal role, particularly in the field of medical sciences. Considering the severity of cancer, accurate data on drug reactions is essential because it helps in the prevention, detection, and management of drug-induced adverse reactions, minimizing avoidable medical prescription errors. In the existing literature, underreporting of drug reactions is a common issue in oncology, which could potentially be addressed through proactive cancer forums. To the best of our knowledge, there is no readily available dataset comprising all cancer types for PCV practices. Therefore, to enhance cancer care, we have curated a novel dataset named *CanExpance* that includes drugs and their effects on various cancer types. We will now explain the steps taken to create the dataset.

A. Data Collection

A recent qualitative analysis of online discussion forums was conducted to investigate cancer patients' perspectives on ADR caused by chemotherapy drugs. We conducted an internet search to find relevant online forums, focusing on drug reactions associated with all types of cancers to help generalize pharmacovigilance (PCV) in oncology. After reviewing numerous websites, two government-funded virtual health services were identified as the most relevant for our research: Cancer Research UK (CRU) [1] and Cancer Survival Network (CSN) [2]. These forums provided access to content without requiring membership and directly promoted discussion among patients about side effects and experiences with drug consumption or discontinuation.

Internal searches on each forum were conducted using keywords such as "side effects", "adverse drug reactions", "adverse drug events" and "drug reactions". We used Python's Selenium to web scrape the data for pharmacovigilance. The CSN website returned 1,100 results, while CRU returned around 1,900 results. The information retrieved included the topic of discussion, the type of cancer, patient posts about drugs and their effects, and the date of posting. To protect user's identities, usernames on these websites were anonymized. The resulting threads were sorted by the date of the most recent post. We have provided different statistics of the *CanExpance* dataset in Table II.

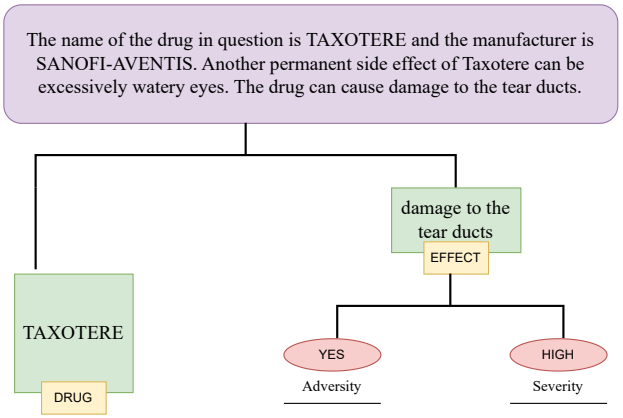

Fig. 1. Flowchart depicting the annotation process for the *'CanExpance'* dataset

1) Data Annotation: To ensure detailed annotation in accordance with ethical standards, we recruited the help of two medical students and one Ph.D. student chosen based on specific criteria. These criteria included being at least 25 years old, fluent in English (reading, writing, and speaking), and willing to deal with sensitive material. The annotation

[1] https://www.cancerresearchuk.org/
[2] https://csn.cancer.org/

process was completed in two months, and participants were compensated for their efforts. [3]. We structured the annotation process as shown in Figure 1. To verify the quality of the annotated data, we established rigorous standards that each sample had to meet:

- For each post mentioning multiple drugs and numerous effects (positive and negative), extract only those drug names linked to adverse drug events (negative effects).
- Each data instance's adversity of the drug event is assessed using specific terms indicating adversity, such as "bad," "worse," "unbearable," "irrecoverable," "permanent," or similar expressions conveying similar sentiments.
- Each data instance's severity of the drug event is assessed based on explicit mentions of congenital anomalies, life-threatening situations, disabilities, or hospitalizations (initial or prolonged). If these criteria are not explicitly stated, the severity is categorized as not applicable to that specific data point.
- Every data point includes a URL link. For each data instance, access the content at that URL to gain insight and context about the data.

TABLE II
STATISTICS OF *CanExpance* DATASET

Measures	**Size**
No of Samples	2996
Number of true labels (Adversity)	922
Number of false labels (Adversity)	2074
Number of High labels (Severity)	292
Number of Moderate labels (Severity)	876
Number of Mild labels (Severity)	553
Number of Not Mentioned labels (Severity)	1275

To maintain consistency among annotators, final labels were assigned via majority voting. Annotators were instructed to remain objective without bias related to demographics or other factors. The quality of annotations was determined by calculating inter-annotator agreement (IAA) using Cohen's Kappa [36]. The agreement score of 0.75 confirms that the annotations are acceptable and of high quality.

B. Ethics and Broader Impact

1) User Privacy: Our dataset includes ADR , drug names and corresponding posts, each annotated with labels, while ensuring no personal user information is included.

2) Biases: Any biases identified within the dataset are unintentional. We emphasize our commitment to not causing harm to any individual or group. Recognizing the subjectivity involved in assessing whether a post contains ADR, we have obtained consensus from all annotators before finalizing the data.

[3]The The medical students were remunerated with gift vouchers and honorariums in accordance with https://www.minimum-wage.org/international/india.

3) Intended Use: We provide this dataset to encourage further research in the detection of Adverse Drug Events. It is shared exclusively for research purposes and is not licensed for commercial use.

IV. PROPOSED METHODOLOGY

A. Problem Definition

Given a dataset of medical instances $\{(P_i, D_i, E_i, A_i, S_i)\}_{i=1}^{N}$, where P_i represents the user post, D_i denotes the drug name(s), E_i is the adverse drug reaction (ADR), A_i indicates adversity level, and S_i signifies severity level, our objective is to develop a multi-task generative framework M to predict the triplet (E_i, A_i, S_i) from the input pair (P_i, D_i). We aim to model the joint probability:

$$P(E_i, A_i, S_i \mid P_i, D_i, M)$$

which captures the likelihood of generating the effects (E_i), adversity (A_i), and severity (S_i) given the input post (P_i) and drug name(s) (D_i) within the multitask generative framework (M). Furthermore, our objective is to conduct multi-feature classification on the adversity and severity parameters of the *CanExpance* dataset. We evaluate the performance of deep learning models, which includes CNN+LSTM, RNN, and GRU, along with Clinical LLMs, on this dataset for the task of multi-feature classification.

Our objective is to conduct multi-feature classification on the adversity and severity parameters of the *CanExpance* dataset. We evaluate the performance of deep learning models, which includes CNN+LSTM, RNN, and GRU, along with Clinical LLMs, on this dataset for the task of multi-feature classification.

To train our multitask generation model, we incorporated clinical LLMs with the *CanExpance* dataset and used contextual examples of user posts, drug names, effects, adverse labels, and severity labels. We have fine-tuned the LLaMA 2 [37] model to develop the Cancer Adverse Drug Identification Framework (*CanADI*). As demonstrated in the Figure 2, the input consists of user posts and drug names, which are processed through embedding layer, multi-head attention, and LoRA modules. During the validation phase, user posts and drug names were provided as input to the *CanADI* Framework to generate effects, adversity labels, and severity labels. Furthermore, considering the prevalence of large language models (LLMs) we conducted a comparative analysis on the *CanExpance* dataset to evaluate the performance of various clinical LLMs like Flan T5 [38], PMC-LLaMA [39], MedAlpaca [40], Bio Clinical BERT [41] for pharmacovigilance in cancer because of their training on medical corpora. This comparison detailed each LLM's effectiveness and suitability in addressing the complex task of generating ADR associated with each drug and its severity and adversity levels, thereby contributing to the refinement and optimization of our multitask generative framework.

B. Alignment for Cancer Vigilance : Implementation Details

Clinical language models are known for their effectiveness in medical analysis because they are trained on large

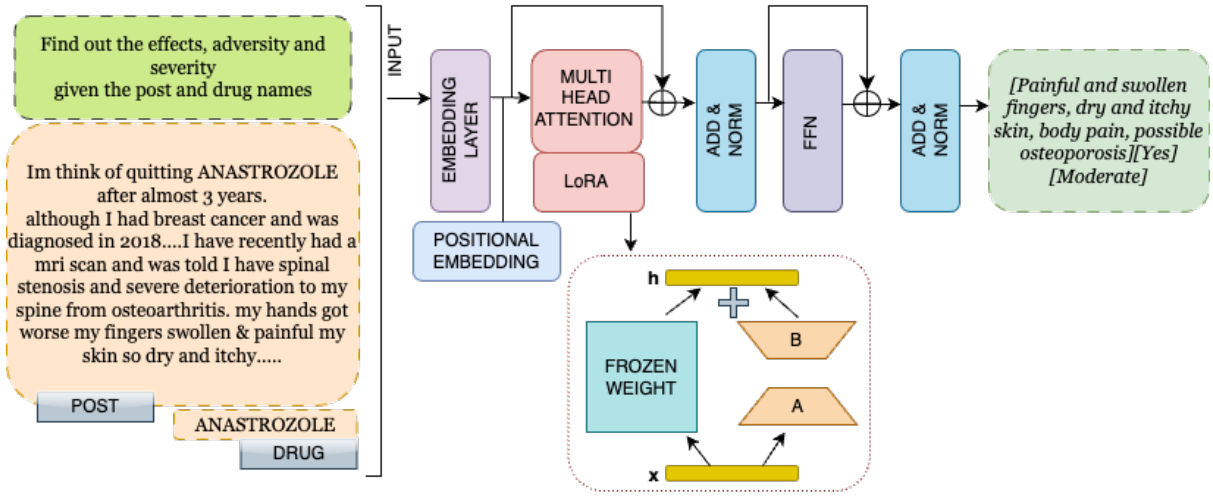

Fig. 2. Proposed Model Architecture of *Cancer Adverse Drug Identification Framework (CanADI)*. A and B denote the LoRA modules. Frozen weight refers to the fixed weights of Multi-head attention. x and h are hidden representations before and after applying the LoRA module, respectively.

medical corpora. However, when we tested on questions related to cancer drugs, our clinical LLMs failed to generate any relevant answers. To address this shortcoming, we fine-tuned the pre-trained Llama2 model with QLoRA using the *CanExpance* dataset to develop *CanADI*, which aligns with our pharmacovigilance goals as demonstrated in Figure 2. For the generative task of predicting side effects, adversity, and severity based on cancer drug-related posts and drug names, we used low-rank adaptation for QLoRA (rank 32) on a variety of LLMs, including the general domain *LLaMA2* (13B parameters) and biomedical LLMs such as *PMC-LLaMA* (13B), *MedAlpaca*, and *Flan-T5 Base*.

$$\hat{y} = \text{QLoRA}(\text{LLM}_{\text{adapted}}, \text{input})$$

where, $\hat{y}$ are predicted output for side effects, adversity, and severity, $\text{LLM}_{\text{adapted}}$ is adapted Language Model (LLM) *LLaMA2*; input is the input data comprising cancer drug-related posts and drug names.

For QLoRA's configuration details, we included parameters such as learning rate (5×10^{-5}), adapter rank (32), batch size (8), epochs (3), optimizer (AdamW with 8-bit), scaling factor (Alpha) set to 32, and dropout (0.05). These modified models were used in 'Few-Shot', 'Zero-Shot', and 'Chain-Of-Thought' prompting setups to generate side effects.

C. Generation Components

- Input Layer: The input, consisting of posts and drug names, is tokenized into words or subwords for processing.
- Positional Embedding: Positional embedding is used to encode the position of each token in the sequence. The positional embeddings are added to the token embeddings, providing the model with information about the relative positions of tokens in the sequence.
- Embedding Layer: The embedding layer transforms each token (from the post and drug names) into a dense vector with a fixed size. These vectors capture the semantic properties of the tokens.
- Multi-Head Attention: Multi-head attention allows the model to focus on different parts of the input sequence simultaneously. It consists of several attention heads, each of which performs a scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are the query, key, and value matrices, and d_k is the dimension of the key vectors. The outputs of all attention heads are concatenated and linearly transformed to form the final attention output.
- Add & Norm: Residual connections (Add) help train deep networks by allowing gradients to flow directly through the network. The layer normalization (Norm) approach regulates the learning process by normalizing the input to the next layer.

$$\text{LayerNorm}(x) = \frac{x - \text{mean}(x)}{\sqrt{\text{var}(x) + \epsilon}} \cdot \gamma + \beta$$

where γ and β are learnable parameters, and ϵ is a small constant.
- Feed-Forward Network (FFN): The FFN consists of two linear transformations with a ReLU activation in between:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

This network captures non-linear relationships in the data, further processing the output from the attention mechanism.
- LoRA Modules (A and B): LoRA (Low-Rank Adaptation) modules introduce trainable parameters in the form of low-rank matrices. These modules adapt the model for specific tasks without significantly altering the pre-trained weights:

$$\Delta W = AB$$

where A and B are low-rank matrices. The original weights W are modified by adding ΔW, allowing the model to learn task-specific features.

- Output Layer: The final layer generates predictions for the ADR, including their effects, adversity, and severity. This layer typically uses a combination of softmax (for categorical outputs like severity) and sigmoid (for binary outputs like adversity) activation functions to produce probabilities:

$$P(E_i, A_i, S_i | P_i, D_i, M)$$

where E_i, A_i, and S_i denote the effects, adversity, and severity, respectively, given the input paragraph (P_i) and drug name(s) (D_i) using the multitask generative framework (M).

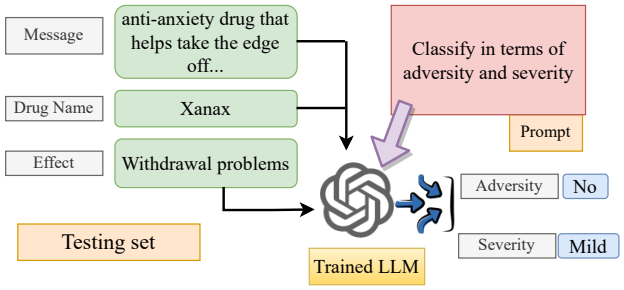

Fig. 3. An instance of the Multi-Feature Classification model classifying ADR in terms of the presence of adversity (yes/no) and severity levels, tested on the *CanExpance* dataset.

D. Multi Feature Classification

1) ADR Classification: : The primary objective of this task is to distinguish between ADR assertive posts using binary classification. In this classification scheme, the two distinct classes are denoted as 'Yes', demonstrating that the user's experience is adverse, and 'No', indicating that the user's experience is not adverse, even though the post included other medical terms such as drugs and indications.

2) Severity Classification: : Based on the binary classification methodology, we extended our approach to multi-class classification for the severity feature as demonstrated in Figure 3. To achieve this, we utilized the capabilities of various LLMs that are capable of processing and understanding clinical data. In addition, we conducted a comparative study of deep learning classification frameworks and language models. For the development of multi-feature classification models for cancer PCV, the following deep learning frameworks were used for classification: CNN + LSTM, RNN, and GRU.

CNN + LSTM:

$$\hat{y}_{\text{binary}} = \sigma(h_{\text{fc}} \times W_{\text{binary}})$$

$$\hat{y}_{\text{multi}} = \text{softmax}(h_{\text{fc}} \times W_{\text{multi}})$$

where, $\hat{y}_{\text{binary}}$ represents the binary classification output (adverse effects), $\hat{y}_{\text{multi}}$ represents the multi-class classification output (severity), h_{fc} represents the output of the fully connected layer, W_{binary} represents the weight matrix for binary classification, W_{multi} represents the weight matrix for multi-class classification, σ represents the sigmoid activation function, softmax represents the softmax activation function.

RNN:

$$\hat{y}_{\text{binary}} = \sigma(w^T x + b)$$

where, $\hat{y}_{\text{binary}}$ is predicted output for adverse effects, σ is sigmoid activation function, w is weight vector, x is the input vector, b represents Bias.

$$\hat{y}_i = \sigma(z_i) \quad \text{where} \quad \sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{4} e^{z_j}}, \quad i = 1, 2, 3, 4$$

where, $\hat{y}_i$ is predicted output for class i in multi-class classification (severity prediction), $\sigma(z_i)$ is the softmax activation function applied to z_i, and $i = 1, 2, 3, 4$ is the index for severity classes (mild, moderate, high, not mentioned).

GRU:

Binary Classification Output Layer:

$$o_t = \sigma(W_o h_t + b_o)$$

Multiclass Classification Output Layer:

$$o_t = \text{softmax}(W_o h_t + b_o)$$

where, W_o is the weight matrix, b_o is the bias, σ is the sigmoid activation function and softmax is the softmax activation function.

V. Experiments, Results, and Analysis

A. Benchmarking the dataset

We have undertaken an initiative to contribute resources in the field of pharmacovigilance for cancer through the introduction of a novel dataset named *CanExpance*. To evaluate the utility and effectiveness of this dataset, we conducted extensive benchmarking experiments using both specialized clinical LLMs, such as PMC Llama [39] and Med Alpaca [40], and general-purpose LLMs like Flan T5 [38], alongside deep learning architectures for classification and generative framework tasks. The performance of these frameworks was assessed using evaluation metrics like the Rouge (R) score [42] and Bleu (B) score [43].

B. Qualitative Evaluation

A qualitative analysis was conducted by a team of medical students under the guidance of a doctor. The team randomly selected 10% of the dataset for evaluation and rated the ADR generated, taking user posts and drugs as context. As shown in Figure 4, the team conducted a thorough evaluation of the outputs generated by both the *CanADI* framework and LLMs pre trained on medical data. Despite being trained for medical analysis, the clinical LLMs failed to produce relevant responses for cancer pharmacovigilance. In contrast, the *CanADI* framework accurately identified ADR and associated them with severity and adversity labels, highlighting our proposed framework's superior performance in the domain of cancer pharmacovigilance.

C. Quantitative Evaluation

1) Multi Feature Classification: In our work on the *CanExpance* dataset, we implemented multi-feature classification to evaluate and compare the effectiveness of deep learning models and clinical language models in dealing with classification tasks. Table III indicates that language models outperformed deep learning models in both adversity and severity classification tasks. Looking at the adversity classification scores, we can observe that the GRU performs better than CNN + LSTM and RNN, achieving an accuracy of 84.05%, an F1 score of 74.33%, a recall of 71.28%, and a precision of 77.65%. However, language models like *Flan-T5 small-med*, *LLaMA 2*, *PMC-LLaMA*, and *MedAlpaca* consistently outperform deep learning models in all metrics. In the severity classification task, deep learning models perform inconsistently. While RNN and GRU achieve reasonable accuracy, F1 score, recall, and precision scores, the CNN + LSTM model performs significantly worse, indicating limitations in capturing the complexity of severity classification. On the other hand, language models such as *Flan T5 small-med*, *LLaMA 2*, *PMC-LLaMA*, and *MedAlpaca* consistently surpass deep learning models in all metrics. Notably, *LLaMA 2* outperforms in accuracy and precision, while Flan T5 small med has the highest F1 score and recall. Language models outperform deep learning models due to their ability to capture intricate patterns and semantic variations that exist in text data. Language models use large-scale pre-training on various textual data to capture rich contextual information and semantic relationships. Furthermore, language models excel at understanding the underlying structure of language, making them ideal for tasks such as text classification. As a result, due to their superior performance and robustness in dealing with text data, we use language models to build our generative frameworks for cancer PCV.

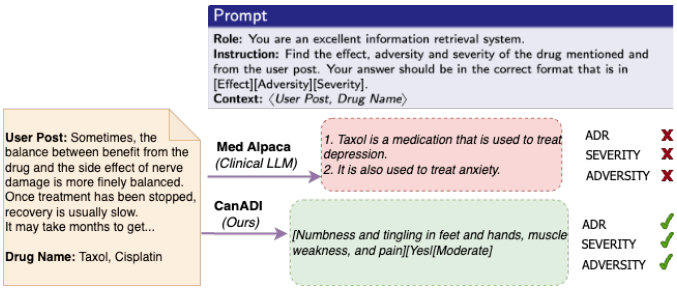

Fig. 4. A qualitative comparison of our proposed *CanADI* framework with a clinical LLM, focusing on its performance in detecting adverse drug reactions (ADR), adversity and severity labels.

2) Multi Task Generation: The evaluation of various clinical language models using three distinct prompting techniques (‘Zero Shot‘, ‘Few Shot‘, and ‘Chain of Thought‘) for our Multi-Task Generative Framework demonstrates the superior performance and robustness of our *CanADI* model, positioning it as the standout solution for cancer ADR identification. In the ‘Zero Shot‘ prompting strategy, the *CanADI* model exhibited performance with a ROUGE-1 score of 0.266 and a BLEU-1 score of 0.414. While other models like Flan T5 Base and PMC LLaMA achieved slightly higher scores, *CanADI*’s performance without any prior examples demonstrates its fundamental robustness and potential for further improvement with more context-specific training.

The Few Shot prompting strategy, as illustrated in Table V, demonstrated a notable improvement in model performance. *CanADI*, in particular, showed the highest enhancement with a ROUGE-1 score of 0.502 and a BLEU-1 score of 0.637, highlighting its ability to utilise limited examples effectively. PMC LLaMA maintained strong performance with a ROUGE-1 score of 0.482 and a BLEU-1 score of 0.603, further validating its robustness in few-shot scenarios. Flan T5 Base displayed stable performance, with a ROUGE-1 score of 0.454 and a BLEU-1 score of 0.541, although slightly lower than PMC LLaMA and *CanADI*. Med Alpaca showed significant improvements, achieving a ROUGE-1 score of 0.479 and a BLEU-1 score of 0.601, suggesting enhanced generative capabilities with few-shot examples. Bio Clinical BERT, while improved from the zero-shot setting, still lagged with a ROUGE-1 score of 0.054 and a BLEU-1 score of 0.144, indicating its relative ineffectiveness in few-shot prompting compared to other LLMs. The significant increases in both ROUGE and BLEU scores demonstrate *CanADI*’s ability to generate relevant and coherent content, distinguishing it from other models.

The Chain of Thought prompting strategy yielded the highest performance across all models, as shown in Table VI. *CanADI* achieved the best results, with a ROUGE-1 score of 0.719, ROUGE-2 score of 0.607, and BLEU-1 score of 0.774, demonstrating *CanADI*’s superior task alignment and advanced generative abilities and effective utilization of the Chain of Thought approach. PMC LLaMA followed closely with a ROUGE-1 score of 0.704 and BLEU-1 score of 0.752, reflecting its strong generative performance. Flan T5 Base also showed substantial improvement, with a ROUGE-1 score of 0.622 and a BLEU-1 score of 0.674, indicating its capability to generate coherent and relevant content. Med Alpaca, with a ROUGE-1 score of 0.672 and a BLEU-1 score of 0.723, demonstrated significant enhancement in its generative abilities. Bio Clinical BERT, while improved, still lagged behind the other models, with a ROUGE-1 score of 0.194 and a BLEU-1 score of 0.226, highlighting its relative inefficiency even in the Chain of Thought setting. The results obtained validate *CanADI*’s ability to understand and generate complex medical information, making it an invaluable tool for assessing ADR.

Findings and Observations Based on the experimental findings, we report the following answers (with evidence) to our investigated research questions (RQs).

RQ 1: What advantages do LLMs offer over traditional deep learning models in the context of pharmacovigilance? Based on the classification scores demonstrated in Table III for Multi Feature Classification of Adversity and Severity in the *CanExpance* dataset, large language models (LLMs) demonstrate clear advantages over traditional deep learning models (RNN, GRU, CNN+LSTM). LLaMA 2, for instance, achieves the highest accuracy for adversity classification at 88.17% and severity classification at 82.16%, outperforming traditional models like GRU, which achieve 84.05% and 79.57% accuracy

TABLE III
CLASSIFICATION SCORES FOR THE MULTI FEATURE CLASSIFICATION OF ADVERSITY AND SEVERITY FOR *CanExpance* DATASET

	Adversity				Severity			
Models↓	***Accuracy***	***f1_score***	***Recall***	***Precision***	***Accuracy***	***f1_score***	***Recall***	***Precision***
RNN	81.89	67.85	58.97	79.86	75.58	74.22	75.58	75.46
GRU	84.05	74.33	71.28	77.65	79.57	79.13	79.57	79.28
CNN + LSTM	81.33	78.25	78.09	78.42	68.00	45.40	48.64	43.82
Flan T5 small-med	83.06	82.62	**83.06**	82.73	81.23	**80.75**	81.23	**81.72**
LLaMA 2	**88.17**	**81.70**	76.30	**87.80**	**82.16**	75.70	79.60	74.20
PMC LLaMA	86.50	79.70	76.80	82.80	81.33	74.30	79.20	72.30
Med Alpaca	87.33	81.10	78.70	83.60	81.16	73.50	**81.50**	71.50

TABLE IV
EVALUATION OF VARIOUS CLINICAL LLMS USING 'ZERO SHOT' PROMPTING STRATEGY FOR THE MULTI TASK GENERATIVE FRAMEWORK

		Models				
Metrics		***Flan T5 Base***	***PMC LLaMA***	***Med Alpaca***	***Bio Clinical Bert***	***CanADI***
ROUGE Score	*ROUGE-1*	0.455	0.495	0.322	0.034	**0.498**
	ROUGE-2	0.397	0.421	0.249	0.011	**0.471**
	ROUGE-L	0.455	**0.491**	0.315	0.034	0.431
BLEU Score	*BLEU-1*	0.599	0.621	0.444	0.100	**0.652**
	BLEU-2	0.512	0.544	0.370	0.051	**0.601**
	BLEU-3	0.463	0.501	0.324	0.026	**0.558**
	BLEU-4	0.445	0.480	0.300	0.016	**0.506**

TABLE V
EVALUATION OF VARIOUS CLINICAL LLMS USING 'FEW SHOT' PROMPTING STRATEGY FOR THE MULTI TASK GENERATIVE FRAMEWORK

		Models				
Metrics		***Flan T5 Base***	***PMC LLaMA***	***Med Alpaca***	***Bio Clinical Bert***	***CanADI***
ROUGE Score	*ROUGE-1*	0.454	0.482	0.479	0.054	**0.502**
	ROUGE-2	0.287	0.410	0.401	0.029	**0.411**
	ROUGE-L	0.398	0.479	0.473	0.053	**0.496**
BLEU Score	*BLEU-1*	0.541	0.603	0.601	0.144	**0.637**
	BLEU-2	0.487	0.527	0.531	0.081	**0.557**
	BLEU-3	0.456	0.484	0.489	0.046	**0.511**
	BLEU-4	0.397	0.463	0.467	0.029	**0.486**

TABLE VI
EVALUATION OF VARIOUS CLINICAL LLMS USING 'CHAIN OF THOUGHT' PROMPTING STRATEGY FOR THE MULTI TASK GENERATIVE FRAMEWORK

		Models				
Metrics		***Flan T5 Base***	***PMC LLaMA***	***Med Alpaca***	***Bio Clinical Bert***	***CanADI***
ROUGE Score	*ROUGE-1*	0.622	0.704	0.672	0.194	**0.719**
	ROUGE-2	0.486	0.606	0.578	0.138	**0.607**
	ROUGE-L	0.621	0.699	0.668	0.187	**0.713**
BLEU Score	*BLEU-1*	0.674	0.752	0.723	0.226	**0.774**
	BLEU-2	0.609	0.712	0.682	0.166	**0.731**
	BLEU-3	0.570	0.688	0.658	0.122	**0.704**
	BLEU-4	0.543	0.672	0.642	0.095	**0.687**

scores, respectively. Additionally, LLMs like Flan T5 small-med and LLaMA 2 exhibit superior F1 scores, with Flan T5 small-med reaching F1 score as 82.62% for adversity and 80.75% for severity. This indicates a balanced performance in precision and recall, with LLaMA 2 achieving a precision of 87.80% for adversity and Flan T5 small-med excelling in recall at 83.06% for adversity and 81.23% for severity. Furthermore, LLMs maintain high performance across tasks, showcasing their robustness and versatility, unlike traditional models, which display greater variability. This analysis highlights the efficacy and reliability of LLMs in accurately classifying ADR, making them highly suitable for pharmacovigilance applications.

RQ 2: How proficiently can clinical LLMs identify ADR, assess their severity, and estimate their adversity to cancer-related drugs?

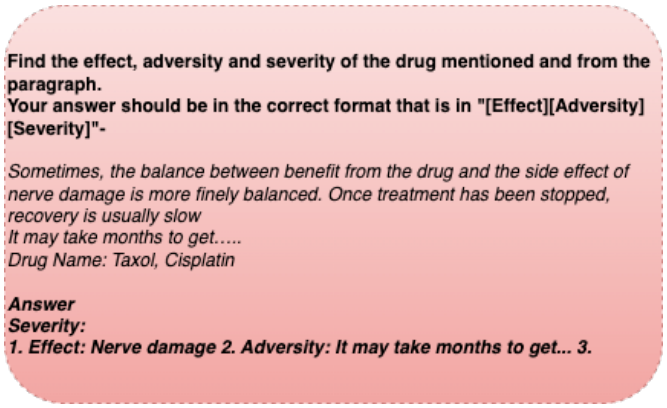

Fig. 5. An instance of PMC-LLaMA tested on *CanExpance* dataset without Fine Tuning and Training

In our research, we used clinical LLMs to create a generative framework that can generate answers when presented with a user post and a drug. An illustrative example featuring PMC-LLaMA, a clinical LLM demonstrated in Figure 5 shows that without fine-tuning the model, clinical LLMs fail to detect any potential effects of cancer-related drugs. However, after fine-tuning the model and training it on the *CanExpance* dataset, the generative framework can effectively identify ADR, assess their severity, and predict their adversity about cancer-related drugs as shown in Figure 6. Moreover, our results suggest that dataset addition and model adaptation play an essential role in improving the efficacy of clinical LLMs for addressing major challenges in pharmacovigilance and drug safety assessment.

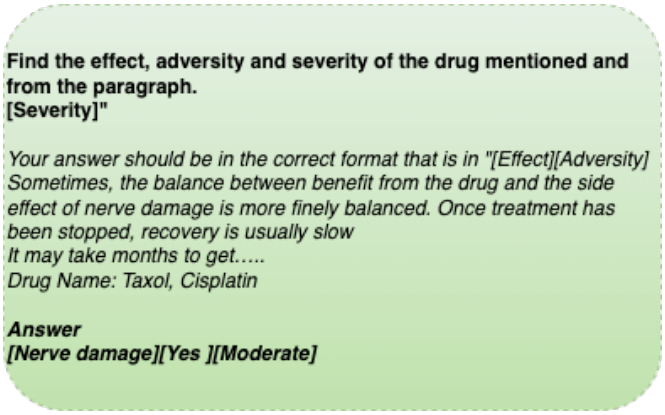

Fig. 6. An instance of PMC-LLaMA tested on *CanExpance* dataset after Fine Tuning and Training

RQ 3: In the context of LLMs, does changing prompts influence the efficacy of models in generating accurate and relevant responses?

In our evaluation of generative PCV frameworks, we used three different prompting techniques to evaluate clinical LLMs' ability for generating cancer-related responses. The prompting strategies used were 'Zero Shot', 'Few Shot', and 'Chain of Thought'. Our findings demonstrated varying levels of performance across prompting techniques. For instance, using the 'Few Shot' prompting technique, as shown in Table V, the *CanADI* model achieved a BLEU-1 score of 0.637. However, using the 'Chain of Thought' prompting strategy, as shown in Table VI, the same model achieved an even higher BLEU-1 score of 0.774. In Table IV, we can see that PMC Llama performs better than all the models, while in 'Chain of Thought' and 'Few Shot' techniques, it demonstrates poorer performance compared to the our *CanADI* Model. Overall, it was observed that when the model was prompted with the 'Chain of Thought' technique, it quickly understood the patterns of questions and produced more relevant responses than the other two prompting techniques. Therefore, adapting prompting strategies influences the efficacy of models in generating accurate and relevant responses.

RQ 4: How will collecting data from sources through self-reporting mechanisms for patients on medications improve pharmacovigilance practices?

Self-reporting mechanisms allow patients to directly communicate their medication experiences, including adverse reactions, to healthcare professionals, regulatory agencies, and other patients undergoing the same treatment. This real-world data provides valuable insights into drug safety profiles, allowing for the early detection of adverse events and potential medication risks. By using patient-reported data, pharmacovigilance practices can become more proactive, responsive, and inclusive, resulting in improved patient safety and better-informed healthcare decisions. Based on such observations, oncologists may decide to adjust medication doses, change the course of treatment, or discontinue therapy. In addition, patient participation in pharmacovigilance promotes a sense of confidence and engagement in their healthcare, leading to a patient-centered approach to medication safety.

VI. Conclusion and Limitations

Cancer patients already endure a life-threatening illness, making active pharmacovigilance practices essential to protect them from unnecessary harm. Given the scarcity of pharmacovigilance data in cancer care, we introduced a novel dataset containing crucial information about drugs used in cancer therapy and patient experiences. Additionally, we propose the *Cancer Adverse Drug Identification Framework (CanADI)* for generating ADR, along with adversity and severity labels, based on user posts and drug information. We conducted a comprehensive comparative evaluation of various clinical LLMs on the *CanExpance* dataset in zero-shot, few-shot, and chain-of-thought settings, using performance metrics such as BLEU and ROUGE. Furthermore, we implemented deep learning classification frameworks to underscore the importance of incorporating large language models (LLMs) into cancer pharmacovigilance practices. Our multifaceted approach addresses the critical need for improved pharmacovigilance in cancer treatment by introducing the *CanExpance* dataset and leveraging advanced language models to enhance patient safety, optimize treatment outcomes, and contribute to a healthier world. While this study provides valuable insights into cancer pharmacovigilance (PCV), it is important to note several limitations. The *CanExpance* dataset comprises approximately 3011 samples of drug reactions, but it lacks any associated pictures for visual comprehension. Furthermore, the study is limited to the English language, possibly excluding valuable observations and variations found in multilingual contexts.

References

[1] H. Li, X.-J. Guo, X.-F. Ye, H. Jiang, W.-M. Du, J.-F. Xu, X.-J. Zhang, and J. He, "Adverse drug reactions of spontaneous reports in shanghai pediatric population," *PLoS One*, vol. 9, no. 2, p. e89829, 2014.

[2] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez, "Utilizing social media data for pharmacovigilance: a review," *Journal of biomedical informatics*, vol. 54, pp. 202–212, 2015.

[3] P. Thompson, S. Daikou, K. Ueno, R. Batista-Navarro, J. Tsujii, and S. Ananiadou, "Annotation and detection of drug effects in text for pharmacovigilance," *Journal of cheminformatics*, vol. 10, no. 1, pp. 1–33, 2018.

[4] R. E. Behrman, J. S. Benner, J. S. Brown, M. McClellan, J. Woodcock, and R. Platt, "Developing the sentinel system—a national resource for evidence development," *New England Journal of Medicine*, vol. 364, no. 6, pp. 498–499, 2011.

[5] P. Baldo and P. De Paoli, "Pharmacovigilance in oncology: evaluation of current practice and future perspectives," *Journal of Evaluation in Clinical Practice*, vol. 20, no. 5, pp. 559–569, 2014.

[6] S. Wu, S. Liu, Y. Wang, T. Timmons, H. Uppili, S. Bedrick, W. Hersh, and H. Liu, "Intrainstitutional ehr collections for patient-level information retrieval," *Journal of the Association for Information Science and Technology*, vol. 68, no. 11, pp. 2636–2648, 2017.

[7] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez, "Utilizing social media data for pharmacovigilance: a review," *Journal of biomedical informatics*, vol. 54, pp. 202–212, 2015.

[8] V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox, "Assessing the impact of a health intervention via user-generated internet content," *Data Mining and Knowledge Discovery*, vol. 29, pp. 1434–1457, 2015.

[9] B. Biseda and K. Mo, "Enhancing pharmacovigilance with drug reviews and social media," *arXiv preprint arXiv:2004.08731*, 2020.

[10] E. Yom-Tov, "Predicting drug recalls from internet search engine queries," *IEEE journal of translational engineering in health and medicine*, vol. 5, pp. 1–6, 2017.

[11] B. Zou, V. Lampos, R. Gorton, and I. J. Cox, "On infectious intestinal disease surveillance using social media content," in *Proceedings of the 6th International Conference on Digital Health Conference*, 2016, pp. 157–161.

[12] A. Sarker, A. Nikfarjam, and G. Gonzalez, "Social media mining shared task workshop," in *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific, 2016, pp. 581–592.

[13] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo, "Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports," *Journal of biomedical informatics*, vol. 45, no. 5, pp. 885–892, 2012.

[14] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang, "Cadec: A corpus of adverse drug event annotations," *Journal of biomedical informatics*, vol. 55, pp. 73–81, 2015.

[15] K. D'Oosterlinck, F. Remy, J. Deleu, T. Demeester, C. Develder, K. Zaporojets, A. Ghodsi, S. Ellershaw, J. Collins, and C. Potts, "Biodex: Large-scale biomedical adverse drug event extraction for real-world pharmacovigilance," *arXiv preprint arXiv:2305.13395*, 2023.

[16] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo, "Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports," *Journal of biomedical informatics*, vol. 45, no. 5, pp. 885–892, 2012.

[17] M. Oronoz, K. Gojenola, A. Pérez, A. D. De Ilarraza, and A. Casillas, "On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions," *Journal of biomedical informatics*, vol. 56, pp. 318–332, 2015.

[18] A. Patki, A. Sarker, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. O'Connor, K. Smith, and G. Gonzalez, "Mining adverse drug reaction signals from social media: going beyond extraction," *Proceedings of BioLinkSig*, vol. 2014, pp. 1–8, 2014.

[19] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "Biocreative v cdr task corpus: a resource for chemical disease relation extraction," *Database*, vol. 2016, 2016.

[20] T. Huynh, Y. He, A. Willis, and S. Rüger, "Adverse drug reaction classification with deep neural networks," in *Proceedings of COLING 2016: Technical Papers*. Coling, 2016.

[21] A. Nikfarjam and G. H. Gonzalez, "Pattern mining for extraction of mentions of adverse drug reactions from user comments," in *AMIA annual symposium proceedings*, vol. 2011. American Medical Informatics Association, 2011, p. 1019.

[22] A. Nikfarjam, A. Sarker, K. O'connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *Journal of the American Medical Informatics Association*, vol. 22, no. 3, pp. 671–681, 2015.

[23] A. Cocos, A. G. Fiks, and A. J. Masino, "Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts," *Journal of the American Medical Informatics Association*, vol. 24, no. 4, pp. 813–821, 2017.

[24] K. Lee, A. Qadir, S. A. Hasan, V. Datla, A. Prakash, J. Liu, and O. Farri, "Adverse drug event detection in tweets with semi-supervised convolutional neural networks," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 705–714.

[25] S. Chowdhury, C. Zhang, and P. S. Yu, "Multi-task pharmacovigilance mining from social media posts," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 117–126.

[26] B. Biseda and K. Mo, "Enhancing pharmacovigilance with drug reviews and social media," *arXiv preprint arXiv:2004.08731*, 2020.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[28] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[29] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[30] T. K. Gandhi, S. B. Bartel, L. N. Shulman, D. Verrier, E. Burdick, A. Cleary, J. M. Rothschild, L. L. Leape, and D. W. Bates, "Medication safety in the ambulatory chemotherapy setting," *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 104, no. 11, pp. 2477–2483, 2005.

[31] P. M. Lau, K. Stewart, and M. Dooley, "The ten most common adverse drug reactions (adrs) in oncology patients: do they matter to you?" *Supportive care in cancer*, vol. 12, pp. 626–633, 2004.

[32] S. P. Shaikh and R. Nerurkar, "Adverse drug reaction profile of anticancer agents in a tertiary care hospital: An observational study," *Current Drug Safety*, vol. 17, no. 2, pp. 136–142, 2022.

[33] A. J. Leendertse, D. Visser, A. C. Egberts, and P. M. van den Bemt, "The relationship between study characteristics and the prevalence of medication-related hospitalizations: a literature review and novel analysis," *Drug Safety*, vol. 33, pp. 233–244, 2010.

[34] J. Schmider, K. Kumar, C. LaForest, B. Swankoski, K. Naim, and P. M. Caubel, "Innovation in pharmacovigilance: use of artificial intelligence in adverse event case processing," *Clinical pharmacology & therapeutics*, vol. 105, no. 4, pp. 954–961, 2019.

[35] E. Surendra and R. GARLAPATI, "Development of pharmacovigilance culture," *International Journal of Pharmacy Research & Technology (IJPRT)*, vol. 10, no. 1, pp. 1–4, 2020.

[36] A. J. Viera, J. M. Garrett *et al.*, "Understanding interobserver agreement: the kappa statistic," *Fam med*, vol. 37, no. 5, pp. 360–363, 2005.

[37] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[38] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[39] C. Wu, W. Lin, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Pmc-llama: Towards building open-source language models for medicine," 2023.

[40] T. Han, L. C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Löser, D. Truhn, and K. K. Bressem, "Medalpaca–an open-source collection of medical conversational ai models and training data," *arXiv preprint arXiv:2304.08247*, 2023.

[41] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[42] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[43] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.