THERAPEUTIC ADVANCES in
*Drug Safety*

*Original Research*

# Transformer-based models for ADR detection: cross-drug validation and benchmarking against large language models

*Ther Adv Drug Saf*

2025, Vol. 16: 1–13

DOI: 10.1177/20420986251405082



**Minjung Kim, Kyoung Eun Kim, Jae-Hee Kwon, Ja-Young Han, Jae Hyun Kim and Myeong Gyu Kim**

***Abstract***

**Background:** Adverse drug reactions (ADRs) are harmful side effects of medications. Social media provides real-time, patient-generated data, though its unstructured format presents challenges. Natural language processing and transfer learning offer promising solutions.

**Objective:** This study aimed to evaluate whether transformer-based models fine-tuned on a general ADR dataset can effectively classify ADRs from tweets related to glucagon-like peptide-1 (GLP-1) receptor agonists and to benchmark their performance against state-of-the-art large language models (LLMs).

**Design:** This study employed a machine learning approach using transformer-based language models to classify ADRs in social media.

**Methods:** BERT (bidirectional encoder representations from transformers)-base, BERTweet-base, and GPT-2 (Generative Pre-Trained Transformer-2) models were fine-tuned using Sarker and SIDER (Side Effect Resource) datasets for ADR classification. The test dataset comprised 396 tweets mentioning GLP-1 receptor agonists that were categorized as personal experiences. Model performance was primarily evaluated using the F1 score, which was used to select the optimal model. In addition, the fine-tuned transformer models were benchmarked against state-of-the-art LLMs, including ChatGPT 4o, ChatGPT 4o-mini, and Gemini 2.5 Flash.

**Results:** Among 396 tweets, 116 (29.3%) were classified as ADRs and 280 (70.7%) as non-ADRs. Among the transformer-based models, BERTweet-base achieved the highest performance (accuracy: 0.835, F1: 0.729), outperforming both BERT-base (accuracy: 0.826, F1: 0.679) and GPT-2 (accuracy: 0.766, F1: 0.628). Among the LLMs, ChatGPT 4o-mini demonstrated the best results (accuracy: 0.970, F1: 0.948), followed by Gemini 2.5 Flash (accuracy: 0.954, F1: 0.919) and ChatGPT 4o (accuracy: 0.936, F1: 0.895). Overall, LLMs substantially outperformed the fine-tuned transformer-based models.

**Conclusion:** Fine-tuned transformer-based models demonstrated reasonable performance in ADR detection from GLP-1 receptor agonist tweets, with BERTweet-base performing best. However, state-of-the-art LLMs, particularly ChatGPT 4o-mini, substantially outperformed these models, highlighting their potential for pharmacovigilance tasks.

## Plain language summary

### Using AI to detect drug side effects in tweets

Why was the study done?

Medicines can sometimes cause harmful side effects, known as adverse drug reactions (ADRs). Traditional systems that track these side effects rely on people reporting them,

Correspondence to:
**Myeong Gyu Kim**
College of Pharmacy and Graduate School of Pharmaceutical Sciences, Ewha Womans University, 52 ewhayeodae-gil, Seoul 03760, Republic of Korea
**kimmg@ewha.ac.kr**

**Minjung Kim**
**Kyoung Eun Kim**
**Jae-Hee Kwon**
**Ja-Young Han**
College of Pharmacy and Graduate School of Pharmaceutical Sciences, Ewha Womans University, Seoul, Republic of Korea

**Jae Hyun Kim**
School of Pharmacy and Institute of New Drug Development, Jeonbuk National University, Jeonju, Republic of Korea

but this doesn't always happen quickly or often enough. Social media, like Twitter, can provide real-time insights from people sharing their experiences with medications, but because the language is informal and unstructured, it's hard for computers to understand. This study explored whether advanced AI models could help automatically detect these side effects in tweets.

What did the researchers do?

The researchers tested different transformer-based language models (like BERT and GPT) to see how well they could detect ADRs in tweets about a specific group of diabetes and obesity drugs called GLP-1 receptor agonists. They trained the models using existing datasets about drug side effects and tested them on 396 real tweets where people talked about their personal experiences with the drugs.

What did the researchers find?

Out of the 396 tweets, 116 were identified as describing side effects. Among the transformer-based models, BERTweet-base achieved the best performance. However, the overall best-performing model was GPT-4o-mini, which outperformed both transformer-based and other language models. This finding suggests that while domain-specific transformer models are effective, advanced large language models can achieve even greater performance in detecting ADRs from social media posts about a specific drug group.

What do the findings mean?

This study shows that AI models, especially those designed to understand social media language, can be used to detect drug side effects in tweets. This could lead to faster and more accurate monitoring of medication safety using real-time data from patients themselves, especially for specific drugs like GLP-1 receptor agonists.

***Keywords:*** adverse drug reaction, BERT, GLP-1 receptor agonists, GPT, social media, transfer learning

Received: 30 March 2025; revised manuscript accepted: 20 November 2025.

## Introduction

An adverse drug reaction (ADR) refers a harmful and unintended response to a drug.[1] ADRs represent a significant public health concern, contributing to hospital admissions, prolonged hospital stays, visits to emergency departments, and an increased risk of mortality. ADRs are evaluated in clinical trials; however, rare ADRs may not be detected due to limited sample sizes and controlled study environments. Therefore, it is essential to monitor adverse events in real-world settings, where a diverse patient population can reveal potential risks that may not have been identified during the clinical trial phase.[2]

The voluntary reporting system, such as FDA's Adverse Event Reporting System (FAERS), aggregates data on adverse events and medication errors from various sources, thereby supporting post-market surveillance.[3] However, it heavily relies on reports by healthcare professionals and captures only a fraction of adverse events.[4] This limitation can lead to underreporting or delayed recognition of emerging safety issues. While some studies suggest that patient chart reviews by pharmacists are more comprehensive, such methods are resource-intensive and not sustainable for continuous monitoring.[5] A key challenge in improving medication safety is the lack of a routine, rapid, and scalable adverse event monitoring system.[6] Therefore, there is a critical need for innovative, automated systems capable of efficiently processing large volumes of clinical data to promptly detect and mitigate adverse events.

Complementary surveillance methods are essential to provide timely insights into drug safety concerns and address the limitations of current reporting systems.

The combination of social media platforms, such as X (formerly Twitter), and natural language processing can serve as a dynamic source for adverse drug events. Conventional natural language processing techniques commonly used for ADR detection include Term Frequency-Inverse Document Frequency, Bag of Words, Word Embeddings (such as Word2Vec and GloVe), n-gram models, and statistical models like Naive Bayes, Logistic Regression, and Support Vector Machines.[7] However, these methods have several limitations in ADR detection. Conventional approaches often fail to capture the contextual relationships between words, resulting in potential misinterpretation of ADR mentions.[8] Ambiguities such as polysemy or the presence of synonyms can further challenge these models. Moreover, their reliance on explicitly mentioned ADRs in the training data restricts their ability to identify novel or nuanced cases.[9] These shortcomings underscore the need for more sophisticated models capable of addressing these challenges effectively.

To address these challenges, researchers have adopted advanced architectures such as the transformer, introduced in 2017 by Google.[10] The self-attention mechanism, a core feature of this architecture, enables models to assign varying importance to different parts of an input sentence.[10] Multi-head self-attention further enhances this by allowing simultaneous focus on multiple parts of the input.[10] Leveraging this design, pre-trained models such as BERT (bidirectional encoder representations from transformers) and GPT (generative pre-trained transformer) have demonstrated exceptional performance across various natural language processing tasks and can be fine-tuned for ADR classification. Fan et al. demonstrated the feasibility of using BERT word embeddings for ADR extraction from social media forum data, outperforming non-pretrained and other pretrained embeddings.[11] Huang et al. introduced a batch-wise adaptive strategy with BERT for ADR detection.[12] Dong et al. utilized publicly available labeled adverse event data to construct the BERT-based model and evaluated through generalized external social media dataset.[13]

Challenges remain in applying transfer learning to ADR detection. In previous studies, a single ADR dataset (such as Sarker's study,[7] Nikfarjam's study,[14] the attention deficit hyperactivity disorder dataset,[15] or the SMM4H (Social Media Mining for Health) dataset[16]) was typically divided into fine-tuning and test datasets.[12,17–19] In our previous work, we found that a model fine-tuned on a general COVID-19 rumor dataset performed worse in classifying COVID-19 rumors related to garlic than a model specifically fine-tuned on garlic-related rumors.[8] This highlights that fine-tuning on more specific subsets of data can significantly improve performance in niche classification tasks. However, creating labeled datasets for individual drugs is labor-intensive and often impractical, as some drugs may lack sufficient data. Therefore, cross-drug validation is necessary to determine whether a model fine-tuned on a general annotated dataset can effectively detect ADRs for specific drugs outside its fine-tuning scope.

Thus, this study aimed to apply BERT and GPT-2-based language models to classify ADRs from tweets related to glucagon-like peptide-1 (GLP-1) receptor agonists and evaluate whether a model fine-tuned on a general annotated dataset can effectively detect ADRs for specific drugs. In addition, we benchmarked these models against large language models (LLMs) to assess their relative performance. Drugs suitable for adverse event detection through social media are primarily those used by populations that frequently engage on these platforms. GLP-1 receptor agonists, originally developed for type 2 diabetes, have recently gained U.S. approval for obesity treatment, with liraglutide and semaglutide emerging as promising options. While these medications are generally safe and associated primarily with gastrointestinal adverse events, their increasing use has the potential to reveal rare ADRs. Furthermore, the significant public interest in GLP-1 receptor agonists has driven a surge in online searches and social media discussions.[20,21] Given this heightened activity, GLP-1 receptor agonists are particularly well-suited for adverse event monitoring using social media. A key strength of our study lies in the development of a well-curated GLP-1 receptor agonist dataset, created through filtering and fine-tuning models to classify personal experiences and adverse events, ensuring its reliability for external

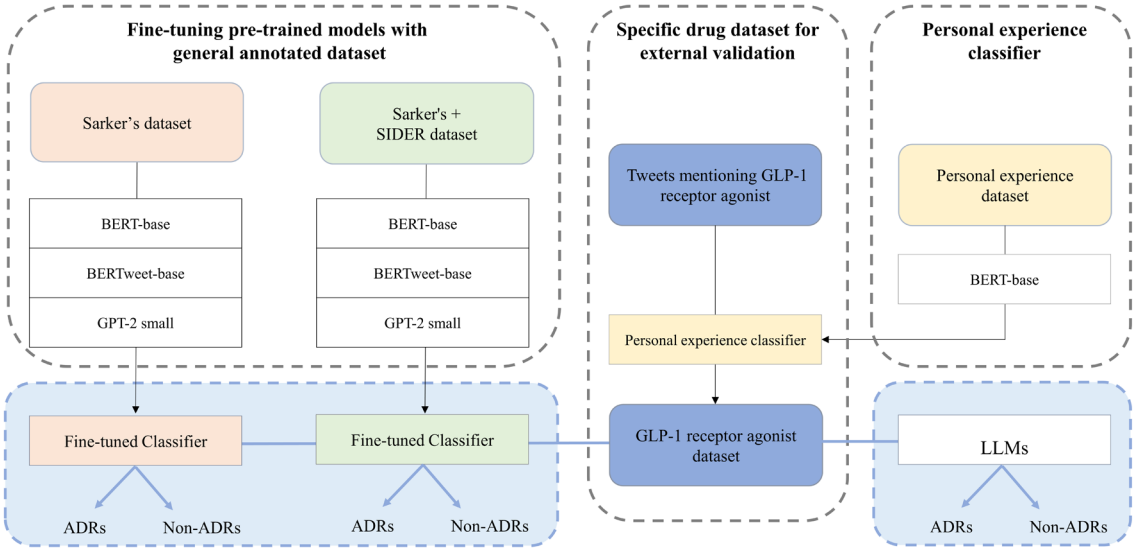


**Figure 1.** Study flow diagram.

validation. Unlike prior studies that utilized domains other than text, we focus solely on text to showcase the models' ability to handle real-world data challenges.

## Methods

### *Study design*

This study involves fine-tuning pre-trained models for two tasks: one for classifying personal experiences and the other for classifying adverse events (Figure 1). In addition, their performance was benchmarked against LLMs, including ChatGPT and Gemini, to assess relative effectiveness in ADR detection. The study was exempted from Institutional Review Board review (ewha-202203-0027-01). The reporting of this study conforms to the Minimum Information about Clinical Artificial Intelligence Modeling (MI-CLAIM) checklist.[22]

### *Fine-tuning datasets*

For fine-tuning, we used Sarker's dataset, which was annotated for the presence of ADRs in a previous study.[7] In that study, 10,822 tweets, randomly selected based on searches for drug generic and brand names, were annotated by two experts. The Cohen's Kappa for inter-annotator agreement was 0.71.[7] Tweet IDs and annotations were freely available on GitHub.[7] Among 10,822 tweets, 5514 tweets were accessible on July 6, 2022. The dataset was imbalanced, with 614 ADR tweets (11.1%; label = 1) and 4900 non-ADR tweets (88.9%; label = 0).

To address the data imbalance problem in Sarker's dataset, another dataset was added. The SIDER (Side Effect Resource) 4.1 dataset, which contains information on marketed medicines (DRUG NAME) and their recorded ADRs (LIST OF ADRs) extracted from public documents and package inserts, was used. All data are freely available on GitHub.[23] We modified the SIDER dataset to 841 sentences in the format of "DRUG NAME side effects are LIST OF ADRS," such as "liraglutide side effects are nausea, diarrhea, vomiting, constipation, headache" and "ramipril side effects are headache, hypotension, cough, cough increased, dizziness, vertigo, nausea, vomiting, asthenia, angina pectoris." All were labeled as 1 since they pertained to ADRs (Supplemental File 1).

Therefore, in our study, we compared two fine-tuning datasets: Sarker's dataset ($n = 5514$; 614 ADR and 4900 non-ADR) and Sarker's + SIDER dataset ($n = 6355$; 1455 ADR and 4900 non-ADR). In both cases, the number of non-ADR instances exceeded ADR instances, with ratios of approximately 8:1 and 3.5:1, respectively. To address this imbalance, we set the class weights of 8 and 3.5 to the ADR class for the Sarker's and

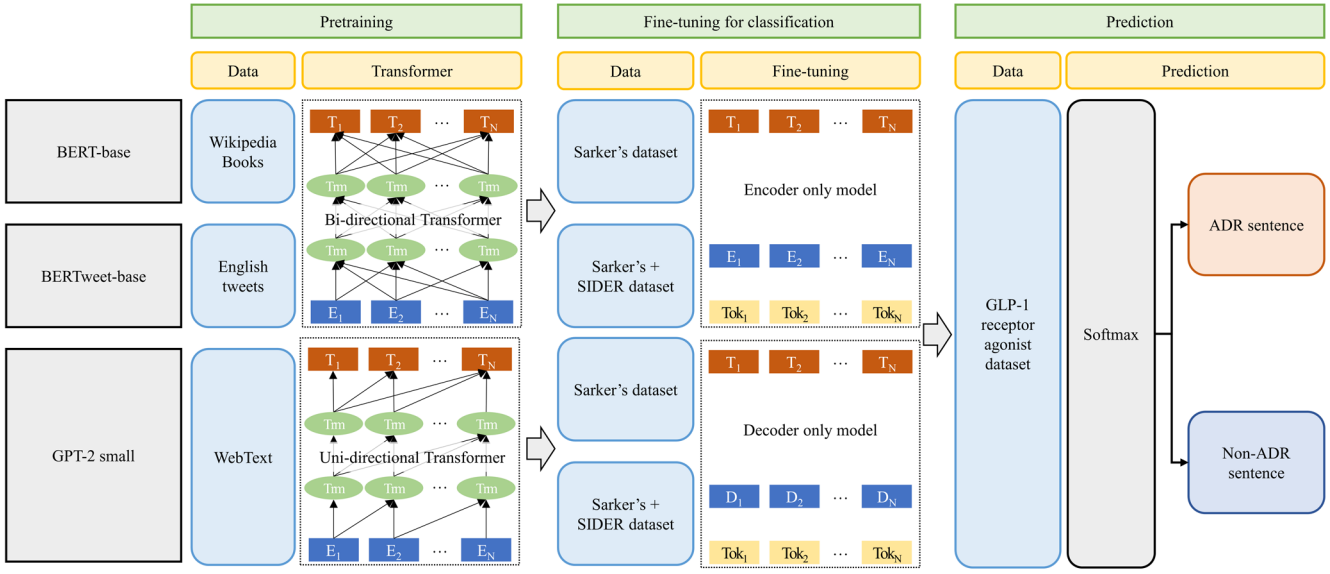


**Figure 2.** Pipeline of BERT-base models and GPT-2 model for classification task.
BERT, bidirectional encoder representations from transformers; GPT-2, generative pre-trained transformer-2.

Sarker's + SIDER datasets, respectively. The fine-tuning datasets were divided into training (80%) and validation (20%) sets. Handles (in the form of "@name"), Uniform Resource Locators, white spaces, and non-ASCII characters were removed from the text. Spaces were inserted between punctuation marks, and words were converted to lowercase.

### *Fine-tuning transformer models*

Figure 2 shows the whole pipeline of the BERT-based models and GPT-2 small model for classification task (pretraining, fine-tuning, and prediction).

For ADR classification, we fine-tuned BERT and GPT-2 models. Table 1 shows the BERT models (BERT-base, BERTweet-base) and GPT-2 small model employed in this study. We used pre-trained BERT-base (uncased), BERTweet-base, GPT-2 small hosted on the Hugging Face Model Hub (accessed on July 8, 2024). All analyses were conducted in the Google Colab environment.

BERT and GPT are among the popular transformer-based neural network models. GPT is an autoregressive model that exclusively utilizes decoder stacks, consisting of masked self-attention and feed-forward layers.[24] Meanwhile, BERT is a bidirectional model that exclusively utilizes encoder stacks, consisting of self-attention and feed-forward layers.[25] When making predictions, GPT considers the left context, while BERT takes both the left and right contexts into account. This enables BERT to excel in sentiment analysis and natural language understanding tasks. In contrast, GPT excels in language modeling for text generation and translation tasks. Therefore, in our study, we utilized BERT and GPT-2.

Two BERT models, BERT-base and BERTweet-base were employed in this study. A classification layer was added on top of BERT models using the "BertForSequenceClassification" and "RobertaForSequenceClassification" interfaces. As input data were fed, the entire pre-trained BERT models and the additional untrained classification layer were trained on our ADR classification task. For GPT-2, the token prediction head needed to be replaced with a classification head. This head, usually consisting of a few additional layers, takes the output of the GPT-2 model and produces logits corresponding to different classes. The "GPT2ForSequenceClassification" interface, which includes a classification head on top of the GPT-2 transformer layers, was used. The hyperparameters were set empirically without formal optimization: learning rate for the Adam optimizer = $2 \times 10^{-5}$; number of epochs = 4–10; batch size = 32. Additionally, early stopping was applied with a patience of 3. The tokenizer

**Table 1.** Transformer models used in the study.

| Model | Parameter count | Architecture | Pre-training data |
|---|---|---|---|
| BERT-base | 110 M | 12 layers, 768 hidden units, 12 attention heads | 2500 M words from English Wikipedia and 800 M words from BookCorpus |
| BERTweet-base | 135 M | Similar architecture to BERT-base | 850 M English tweets |
| GPT-2 small | 117 M | 12 layers, 768 hidden units, 12 attention heads | Text from 45 M websites (WebText) |

BERT, Bidirectional Encoder Representations from Transformers; GPT, Generative Pre-trained Transformer.

evaluated the default tokenizers for each model, and it was found that some instances were classified as false negatives due to over-tokenization. As a result, some ADRs were added as tokens (Supplemental Table 1). All sentences were padded to a single, fixed length of 128 tokens.

### *GLP-1 receptor agonist dataset*

From December 2014 to February 2022, 3963 tweets mentioning either liraglutide or semaglutide were collected using the search query "#saxenda OR #wegovy OR #liraglutide OR #semaglutide." Python version 3.7 (Python Software Foundation, Fredericksburg, VA, USA) and X premium application programming interface were used for data collection.

Tweets include not only experiences of GLP-1 receptor agonist users but also humor, news, study results, or experiences of others.[9] Therefore, to evaluate adverse events using tweets that specifically relate to personal experiences, we implemented a personal experience classifier. In a previous study, personal experience tweets about medication use were classified using BERT.[26] The BERT-base model achieved a classification accuracy of 0.818, which was higher than the BERT-large model's accuracy of 0.813 and the BERTweet model's accuracy of 0.816. (Jiang, Chen and Calix, 2019) Therefore, we also used the BERT-base model in our study. The data used for fine-tuning consisted of tweets from the previous study.[26] A corpus of annotated tweets was available at GitHub.[27] Although the corpus provided a total of 12,331 random Tweet IDs and their annotations, only 6234 tweets were accessible as of July 6, 2022, and these were used for fine-tuning.

Using the fine-tuned BERT model, 515 tweets about GLP-1 receptor agonists were classified as personal experience tweets. Two researchers independently reviewed whether tweets were personal experiences, and in cases of disagreement, a third researcher made the final decision. A total of 396 tweets were identified as personal experience tweets, resulting in a precision of 0.77. Subsequently, two researchers independently annotated these tweets for ADRs. In cases where their opinions did not match, a third researcher made the final decision. The types of ADRs were named according to the Medical Dictionary for Regulatory Activities (MedDRA) Preferred Terms version 25.0.

### *Model evaluation*

Tweets mentioning GLP-1 receptor agonists were classified as ADRs or non-ADRs using the fine-tuned transformer-based models. The test dataset was preprocessed in the same way as the fine-tuning datasets. Model performance was evaluated using a $2 \times 2$ confusion matrix, from which accuracy, precision, recall, and F1 score were calculated as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP}/(\text{FN} + \text{TP})$$

$$\text{F1 score} = (2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$$

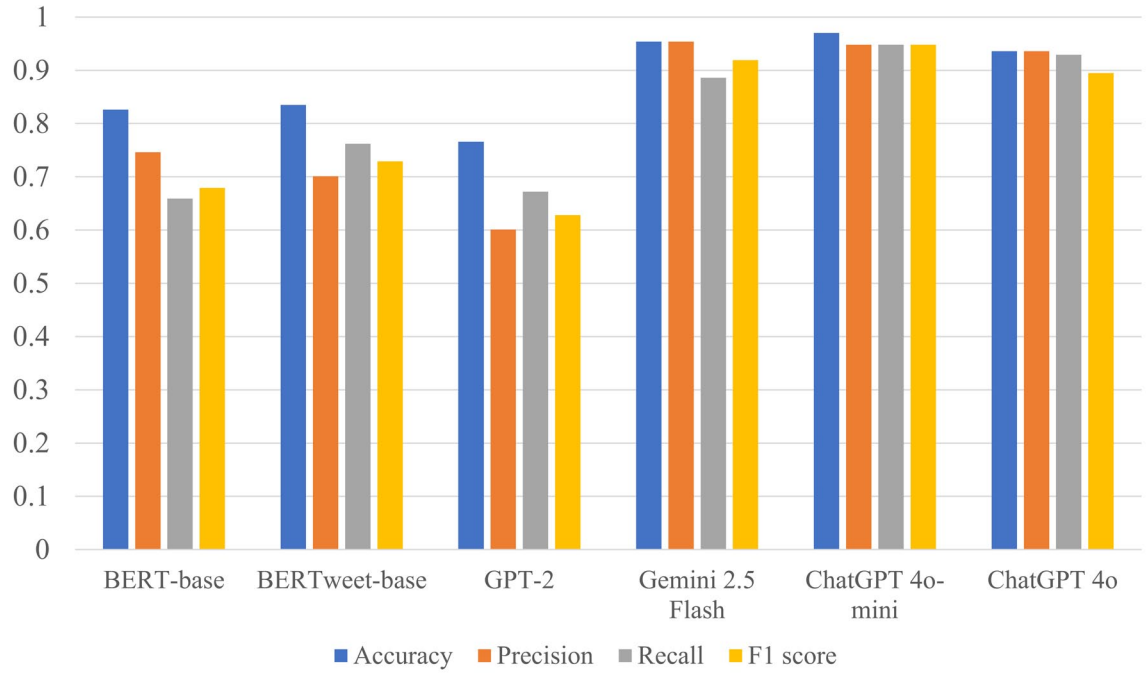


**Figure 3.** Comparison of best-performing results across different model configurations.

where TP, TN, FP, and FN mean true positive, true negative, false positive, and false negative, respectively. The optimal model was selected based on the F1 score, given the imbalance in the dataset, and performance was further assessed using receiver operating characteristic (ROC) and precision–recall (PR) curves.

To compare transformer-based models, each experiment was repeated five times with different random seeds (42, 52, 62, 72, 82). Mean and standard deviation were calculated for each metric, and pairwise *Z*-tests were conducted to assess whether performance differences were statistically significant. The *Z*-score for two models was computed as:

$$Z = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

where $\mu$ and $\sigma$ denote the sample mean and standard deviation, respectively. *p*-Values were derived from a two-tailed standard normal distribution, with a significance threshold of $p < 0.05$. Finally, the performance of fine-tuned transformer models was benchmarked against LLMs. State-of-the-art LLMs, including OpenAI's ChatGPT 4o, ChatGPT 4o-mini, and Google's Gemini 2.5 Flash (all accessed on July 21, 2025), were applied in a zero-shot setting. Tweets were input individually using a standardized binary classification prompt (1 = ADR, 0 = non-ADR) that was refined through pilot testing and then executed five times; run-to-run variability was calculated as the standard deviation. To prevent adaptive behavior, the prompt explicitly instructed the models not to adjust responses between trials. All LLM experiments were conducted directly through the respective online platforms, using the default settings with no manual adjustments to temperature or other parameters. The full prompt is provided in Supplemental Table 2.

### *Error analysis*

To gain further insight into model performance beyond quantitative metrics, we conducted an error analysis of misclassified tweets from the fine-tuned transformer-based models. Using the results from a single test run with a fixed random seed, all instances of false positives and false negatives were systematically reviewed. Each misclassified tweet was systematically reviewed and assigned to conceptually distinct categories based on recurrent linguistic and contextual patterns, to

**Table 2.** Analysis of false positives by category according to frequency.

| Category | Examples |
|---|---|
| Expected effect (no appetite) | • ". . . it has certainly curbed my appetite!" |
| | • "Day 3 liraglutide, no appetite all day . . ." |
| Indication or underlying condition | • "I got put on semaglutide to help with my amphetamine/dextroamphetamine induced binge eating habit" |
| | • ". . . depression made me eat my feelings. I hit 100kgs . . ." |
| | • ". . . 4 years of frequent #crohns flares and the liraglutide treatment have not definitely killed my love of good food." |
| Unknown reason | • "#semaglutide I initially lost 80 lbs on low calorie diet. despite strenuous exercise I quickly gained back 50 lbs. I realized diet meds are an essential part of maintaining weight loss." |
| | • "Nearing the end of my first week on liraglutide . . . I went from 318 lbs to 310 lbs as of this morning. I am quaking, you guys." |

identify common sources of error. This qualitative categorization was used to complement the quantitative evaluation, providing insights into systematic challenges of ADR detection from noisy, user-generated social media text.

## Results

### *GLP-1 receptor agonist dataset*

Out of 396 tweets, 116 (29.3%) were ADR tweets, and 280 (70.7%) were non-ADR tweets. A total of 155 ADRs were mentioned (Supplemental Table 3). Among these, the most common were nausea ($n = 43$), abdominal pain ($n = 24$), headache ($n = 8$), and gastrooesophageal reflux disease ($n = 7$). Five tweets were difficult to categorize under specific adverse reactions: "This shot is really messing me up," ". . . don't love the side effects. . .," and ". . .hope the side effects keep lessening."

### *Model evaluation*

Supplemental Table 4 summarizes all evaluated configurations, with metrics averaged over five runs. Figure 3 presents the best results for each model type, while Supplemental Figure 1 shows the ROC and PR curves of the corresponding fine-tuned transformer-based models.

Among all models, ChatGPT 4o-mini achieved the best performance, recording an F1 score of 0.948, an accuracy of 0.970, a precision of 0.948, and a recall of 0.948. With one random seed, the model resulted in 112 TP, 272 TN, 8 FP, and 4 FN. Other LLMs also outperformed the fine-tuned transformer-based models. Gemini 2.5 Flash achieved an F1 score of 0.919 and an accuracy of 0.954, while ChatGPT 4o recorded an F1 score of 0.895 and an accuracy of 0.936.

Among the fine-tuned transformer-based models, BERTweet-base model showed the best performance, with a mean F1 score of 0.729, an accuracy of 0.835, a precision of 0.701, and a recall of 0.762. With one random seed, the model resulted in 77 TP, 253 TN, 27 FP, and 39 FN. The BERT-base model achieved a mean F1 score of 0.679 and an accuracy of 0.826, whereas GPT-2 achieved a mean F1 score of 0.628 and an accuracy of 0.766.

Supplemental Table 5 summarizes the results of pairwise $Z$-tests across all model combinations. Fine-tuned transformer models (BERT-base, BERTweet-base, GPT-2) showed no statistically significant differences across most metrics, indicating comparable performance within this group. In contrast, all LLMs significantly outperformed the fine-tuned transformer models. Among LLMs, ChatGPT 4o-mini achieved the strongest performance, significantly surpassing both Gemini 2.5 Flash (accuracy: $Z = 2.744$, $p = 0.006$; F1 score: $Z = 3.074$, $p = 0.002$) and ChatGPT 4o

**Table 3.** Analysis of false negatives by category according to frequency.

| Category | Examples |
|---|---|
| Minimal ADRs with positive effects | • ". . . I honestly am so proud and happy with myself! only side effect I'm having a real bad heart burn!" |
| | • ". . . my body is responding well. minimal nausea . . ." |
| | • ". . . I feel so much better already! I think my insomnia is a side effect though, which isn't great but totally worth it" |
| | • ". . . no adverse side effects today apart from I feel really tired." |
| | • ". . . going on liraglutide . . . has helped so much. I'll take the random nausea . . ." |
| | • ". . . although there are subtle side effects such as cramps the injection works really well . . ." |
| Reduction or disappearance of past ADRs | • ". . . day4 still not lost any weight but all nausea has subsided . . ." |
| | • ". . . mild morning nausea that went away pretty quickly after each dose change . . ." |
| Indirect expression | • "Ondansetron (antiemetic drug) saved me from my recent semaglutide" |
| | • ". . . how long did it take to work for you and did the nausea ever go away?" |
| General expression without explicit mention of ADRs | • ". . . don't love the side effects . . ." |
| | • ". . . hope the side effects keep lessening" |
| | • ". . . not feeling well today. only managed a bowl of muesli cereal!" |
| | • ". . . feeling bad on high dose liraglutide . . ." |
| Unknown reason | • ". . . me realizing my thigh is now burning . . ." |
| | • ". . . shot in the thigh instead of my stomach today . . . that shit hurt . . .." |
| | • "The side effect since last night have not been fun. this nausea sucks . . ." |

ADR, adverse drug reaction.

(accuracy: $Z = 6.668$, $p < 0.001$; F1 score: $Z = 6.203$, $p < 0.001$).

### *Error analysis of transformer-based models*

Table 2 provides some examples of false positive, which were categorized into three main groups: (1) "Expected effect (no appetite)" where statements reflected the anticipated outcome of reduced appetite, (2) "Indication or underlying condition" where tweets referenced the use of medication for specific conditions or underlying health issues, and (3) "Unknown reason" where the tweets did not clearly indicate why they were classified as false positive.

Table 3 provides some examples of false negative, which were categorized into five main groups: (1) "Minimal ADRs with positive effects" where ADRs were reported alongside positive experiences with the treatment, (2) "Reduction or disappearance of past ADRs" where previous adverse

effects had diminished or resolved, (3) "Indirect expression" where ADRs were implied rather than directly stated, (4) "General expression without explicit mention of ADRs" where non-specific symptoms or general discomfort were mentioned, and (5) "Unknown reason" where the reasons for the false negative were unclear.

## Discussion

This study shows that transformer-based models fine-tuned with a general ADR dataset can achieve reasonable performance in classifying GLP-1 receptor agonist-related tweets, although state-of-the-art LLMs (ChatGPT, Gemini) significantly outperformed these models.

Among the transformer-based model, BERTweet-base model achieved the best performance. This advantage can be attributed to its pre-training on tweets, which differ substantially from the Wikipedia and BooksCorpus datasets used for BERT-base. Tweets are short and often contain irregular words, abbreviations, and typographical errors, making BERTweet-base better suited for tweet classification tasks, especially when balanced precision and recall are critical.[8] In contrast, GPT-2 performed worse due to its decoder-only, autoregressive architecture designed primarily for text generation, whereas BERT's bidirectional encoder architecture is inherently more effective for classification tasks.[24,25]

Classification performance improved when using the combined Sarker and SIDER datasets. Tweets about medications generally contain more non-ADR than ADR content, leading to class imbalance during fine-tuning. In Sarker's dataset, only 11.1% of tweets were ADR-related, compared to 88.9% non-ADR. Incorporating the ADR-only SIDER dataset helped mitigate this imbalance. Tokenization issues also influenced performance; for example, "constipated" was split into "con," "sti," and "pated," reducing the model's ability to capture medical terms. Updating the tokens improved performance, suggesting that further token adjustments may be necessary for other drug classes.

Our fine-tuned BERTweet-base model showed some limitations. Among the false positives, some included the expected effect of reduced appetite and mentions of indication or underlying condition. This occurred because the model was fine-tuned on general drug-related tweets without specifically incorporating the characteristics of GLP-1 receptor agonists. However, developing individual models that account for the characteristics of every drug would be overly resource-intensive. Therefore, it is necessary for humans to confirm the ADRs detected by the model. The model incorrectly identified false negatives. We considered reports of past ADRs, even if they are no longer present, as ADRs because detecting such cases is important from a pharmacovigilance perspective. However, the model classified these as non-ADR, likely due to the absence of current ADRs. This discrepancy stems from the difference in labeling intent between the fine-tuning data created by other researchers and our classification objectives. Additionally, detecting minimal ADRs mentioned alongside positive effects proved to be challenging. The model also faced limitations in accurately detecting indirect expressions or general statements about side effects, making it difficult to fully capture the intent of the tweet's author.

Our findings on fine-tuned transformer-based models are broadly consistent with previously published studies. Huang et al.[16] applied BERT for ADR classification using Sarker's dataset and reported accuracy between 0.90 and 0.91, with F1 scores ranging from 0.51 to 0.53 under cross-validation. In contrast, our study incorporated external validation and achieved improved F1 performance despite dataset imbalance. Similarly, Dong et al.[12] externally validated a model trained on the ADE-Corpus-V2 using the SMM4H dataset, reporting F1 scores of 0.813 for adverse event terms, 0.807 for words within adverse events, and 0.979 for non-adverse terms.

LLMs substantially outperformed all fine-tuned models, with ChatGPT 4o-mini achieving the highest performance, followed by Gemini 2.5 Flash and ChatGPT 4o. ChatGPT 4o-mini achieved an accuracy of 0.970 and a precision of 0.948, reflecting strong ability to minimize false positives. Its recall and F1 score were both 0.948, indicating a well-balanced capacity to detect true ADR cases without over-prediction. The superior performance of LLMs can be partly explained by the limitations revealed in our error analysis of transformer-based models. Fine-tuned models frequently misclassified tweets that mentioned expected effects or underlying conditions as ADRs, and often failed to detect ADRs expressed indirectly, in colloquial

terms, or without explicit ADR keywords. In contrast, LLMs benefit from extensive pretraining on diverse textual sources, which enhances their ability to interpret implicit meanings, handle non-standard language, and distinguish ADRs from related but non-adverse contexts. These capabilities likely contributed to their lower false-positive rates and improved recall. Interestingly, ChatGPT 4o-mini outperformed ChatGPT 4o across most metrics, potentially due to architectural optimizations or inference-time tuning that prioritize classification tasks such as binary ADR detection. However, this comparison should be viewed with caution as LLMs operate at much larger scale and with far greater computational resources. Such differences also entail higher costs and potential privacy concerns. Within pharmacovigilance frameworks, these issues extend to data protection under regulations such as the General Data Protection Regulation (GDPR) and to the need for clinical validation before LLM-generated safety insights can be integrated into regulatory decision-making.

### *Limitations*

This study has several limitations. First, the test set was relatively small and was inherently noisy due to its social media origin, which may affect the reported performance estimates. Second, as seen in the examples of false positive and false negative, the transformer-based model may misclassify cases where the drug's indication, past experiences, or ADRs are described as minimal. Third, the test dataset was limited to tweets mentioning GLP-1 receptor agonists. The evaluation metrics may vary for other drugs and social media platforms. Also, the selective token additions described in Supplemental Table 1 could introduce subtle class-specific bias that might contribute to better performance observed for LLMs. Fourth, the study results are dependent on the labels of the fine-tuning dataset and the test dataset. These labels can be somewhat subjective, and there may be bias in the process of selecting social media posts. Fifth, the fine-tuning datasets were provided as Tweet IDs, which means that access to the data might vary depending on when it is accessed. Some tweets may have been deleted or become restricted, which could impact the model's performance. Lastly, the models evaluated in this study were designed to classify tweets containing ADRs but do not extract specific ADRs.

Future works could re-run LLMs in a few-shot setting to examine whether limited calibration narrows or widens the performance gap. Also, additional evaluation on the best transformer-based model and the LLM on an independent drug class would help assess cross-drug generalizability beyond GLP-1 receptor agonists.

## Conclusion

This study demonstrated that transformer-based models fine-tuned on a general ADR dataset can generalize to specific drug contexts, effectively classifying GLP-1 receptor agonist-related tweets as ADR or non-ADR. Among these models, BERTweet-base achieved the best performance, reflecting the advantage of pretraining on social media text for ADR detection tasks. However, state-of-the-art LLMs, particularly ChatGPT 4o-mini, substantially outperformed fine-tuned models, underscoring their superior ability to capture nuanced and implicit expressions of ADRs in user-generated text. Despite this performance gap, fine-tuned models remain valuable in scenarios where privacy concerns, domain-specific requirements, or resource constraints limit the deployment of large commercial LLMs.

## Declarations

### *Ethics approval and consent to participate*

The study was exempted from Institutional Review Board review because it used publicly available content (tweets) and posed minimal risk to the research participants as well as to the public (ewha-202203-0027-01). The decision date for the exemption was March 18, 2022.

### *Consent for publication*

Not applicable.

### *Author contributions*

**Minjung Kim:** Formal analysis; Software; Writing – original draft.

**Kyoung Eun Kim:** Data curation; Investigation.

**Jae-Hee Kwon:** Data curation; Investigation.

**Ja-Young Han:** Data curation; Investigation.

**Jae Hyun Kim:** Data curation; Writing – review & editing.

**Myeong Gyu Kim:** Conceptualization; Methodology; Project administration; Validation; Writing – review & editing.

*Acknowledgement*
None.

*Funding*
The authors received no financial support for the research, authorship, and/or publication of this article.

*Competing interests*
The authors declare that there is no conflict of interest.

*Availability of data and materials*
Data supporting the findings of this study are available from the corresponding author upon reasonable request.

**ORCID iDs**
Minjung Kim https://orcid.org/0000-0002-8718-8469
Jae Hyun Kim https://orcid.org/0000-0002-8609-7135
Myeong Gyu Kim https://orcid.org/0000-0002-5593-7672

**Supplemental material**
Supplemental material for this article is available online.

**References**

1. World Health Organization. *International drug monitoring : the role of national centres, report of a WHO meeting* (held in Geneva from 20 to 25 September 1971). Geneva: World Health Organization, 1972.
2. Härmark L and van Grootheest AC. Pharmacovigilance: methods, recent developments and future perspectives. *Eur J Clin Pharmacol* 2008; 64: 743–752.
3. Weaver J, Willy M and Avigan M. Informatic tools and approaches in postmarketing pharmacovigilance used by FDA. *AAPS J* 2008; 10: 35–41.
4. Jha AK, Kuperman GJ, Teich JM, et al. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *J Am Med Inform Assoc* 1998; 5: 305–314.
5. Wolfe D, Yazdi F, Kanji S, et al. Incidence, causes, and consequences of preventable adverse drug reactions occurring in inpatients: a systematic review of systematic reviews. *PLoS One* 2018; 13: e0205426.
6. Shojania K and Thomas E. Trends in adverse events over time: why are we not improving? *BMJ Qual Saf* 2013; 22: 273–277.
7. Sarker A and Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J Biomed Inform* 2015; 53: 196–207.
8. Kim MG, Kim M, Kim JH, et al. Fine-tuning BERT models to classify misinformation on garlic and COVID-19 on Twitter. *Int J Environ Res Public Health* 2022; 19: 5126.
9. Kim MG, Kim J, Kim SC, et al. Twitter analysis of the nonmedical use and side effects of methylphenidate: machine learning study. *J Med Internet Res* 2020; 22: e16466.
10. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *The 31st international conference on neural information processing systems (NIPS 2017)*, Long Beach, CA, USA, 4–9 December 2017. DOI: 10.48550/arXiv.1706.03762.
11. Fan B, Fan W, Smith C, et al. Adverse drug event detection and extraction from open data: a deep learning approach. *Inform Process Manag* 2020; 57: 102131.
12. Huang JY, Lee WP and Lee KD. Predicting adverse drug reactions from social media posts: data balance, feature selection and deep learning. *Healthcare (Basel)* 2022; 10: 618.
13. Dong F, Guo W, Liu J, et al. BERT-based language model for accurate drug adverse event extraction from social media: implementation, evaluation, and contributions to pharmacovigilance practices. *Front Public Health* 2024; 12: 1392180.
14. Nikfarjam A, Sarker A, O'Connor K, et al. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc* 2015; 22: 671–681.
15. Cocos A, Fiks AG and Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *J Am Med Inform Assoc* 2017; 24: 813–821.

16. Klein AZ, Banda JM, Guo Y, et al. Overview of the 8th Social Media Mining for Health Applications (#SMM4H) shared tasks at the AMIA 2023 Annual Symposium. *J Am Med Inform Assoc* 2024; 31: 991–996.

17. Wang Y, Zhao Y, Schutte D, et al. Deep learning models in detection of dietary supplement adverse event signals from Twitter. *JAMIA Open* 2021; 4: ooab081.

18. Hussain S, Afzal H, Saeed R, et al. Pharmacovigilance with transformers: a framework to detect adverse drug reactions using BERT fine-tuned with FARM. *Comput Math Methods Med* 2021; 2021: 5589829.

19. Sakhovskiy A and Tutubalina E. Multimodal model with text and drug embeddings for adverse drug reaction classification. *J Biomed Inform* 2022; 135: 104182.

20. Han SH, Safeek R, Ockerman K, et al. Public interest in the off-label use of glucagon-like peptide 1 agonists (Ozempic) for cosmetic weight loss: a google trends analysis. *Aesthet Surg J* 2023; 44: 60–67.

21. Tselebis A and Ilias I. Further research on internet searches for on- and off-label use of weight-loss medications. *Aesthet Surg J* 2023; 43: NP977–NP978.

22. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 2020; 26: 1320–1324.

23. Himmelstein D. SIDER4, https://github.com/dhimmel/SIDER4/blob/master/SIDER4.ipynb (2016, accessed 27 March 2024).

24. OpenAI Blog. Improving language understanding by generative pre-training, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (2018, accessed 26 November 2024).

25. Devlin J, Chang MW, Lee K, et al. *BERT: pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:181004805, 2018.

26. Jiang K, Chen T, Calix RA, et al. Prediction of personal experience tweets of medication use via contextual word representations. *Annu Int Conf IEEE Eng Med Biol Soc* 2019; 2019: 6093–6096.

27. Tweet_corpora, https://github.com/medeffects/tweet_corpora (2016, accessed 6 July 2022).