Adverse Drug Events Across Data-Production Contexts: Multilingual Detection, Alignment, and Cross-Genre Discourse Analysis

Yujun Ma

Department of Computational Biomedicine, Cedars-Sinai Health Sciences University, Los Angeles, CA, USA
E-mail: Yujun.Ma@cshs.org

Davy Weissenbacher

Department of Computational Biomedicine, Cedars-Sinai Health Sciences University, Los Angeles, CA, USA
E-mail: Davy.Weissenbacher@cshs.org

Jake Patock

Department of Computational Biomedicine, Cedars-Sinai Health Sciences University, Los Angeles, CA, USA
E-mail: Jake.Patock@cshs.org

Graciela Gonzalez Hernandez[†]

Department of Computational Biomedicine, Cedars-Sinai Health Sciences University, Los Angeles, CA, USA
E-mail: Graciela.GonzalezHernandez@csmc.edu

Adverse drug event (ADE) evidence is produced across patient-generated, clinical, and scientific settings that differ in language, documentation purpose, terminology, and degree of standardization. These differences shape both which adverse experiences become visible to pharmacovigilance systems and how readily they can be linked to curated drug-safety knowledge. We examine these relationships across six corpora representing distinct data-production settings: ADE Corpus V2 (medical case reports), SMM4H-2026 Task 1 (multilingual user-generated health content), CADEC V2 (patient-forum narratives), the Dutch ADE Corpus and KronosRX (EHR clinical notes), TwiMed-PubMed (biomedical literature). A shared BERTopic analysis of ADE-positive texts concerning antidepressants and antihypertensives across the four English-language corpora identified nine interpretable topics. CADEC V2 contained a more differentiated distribution of symptom-specific themes, including sexual effects, suicidal or panic-related thoughts, vivid dreams, and memory difficulties, whereas SMM4H-2026, TwiMed-PubMed, and ADE Corpus V2 were dominated by a broader medication, sleep, tiredness, and pain theme. These patterns indicate that data-production context shapes what adverse experiences are expressed and standardized, with patient-generated narratives surfacing subjective, symptom-specific experience largely absent from clinical and scientific sources.

We further examine how data-production context affects alignment with curated pharmacovigilance knowledge using SIDER 4.1. Retrieval coverage varied sharply across sources, from 9.5% in CADEC V2 to 91.0% in TwiMed-PubMed, while KronosRX showed high task-specific lexical coverage (91.7% cardiotoxicity; 81.7% neurotoxicity).

We compare corpus-specific QLoRA fine-tuning of Llama-3.2-3B with SIDER-grounded

inference using Llama-3.1-70B/405B. Among the evaluated configurations, QLoRA obtained the highest observed macro-F1 on ADE Corpus V2 (0.91), CADEC V2 (0.88), and SMM4H-2026 (0.80), whereas Llama-3.1-405B RAG-SIDER obtained the highest observed macro-F1 on Dutch ADE (0.95) and TwiMed-PubMed (0.91). In KronosRX, toxicity-specific RAG improved macro-F1 from 0.54 to 0.75 for cardiotoxicity and from 0.58 to 0.64 for neurotoxicity. Together, these findings suggest that ADE representation, knowledge-base alignment, and detection performance vary substantially across data-production contexts, supporting the use of source-sensitive supervision and knowledge grounding.

Keywords: Adverse drug event detection; pharmacovigilance; large language models; retrieval-augmented generation.

1. Introduction

Adverse drug events (ADEs) are harmful or unintended outcomes associated with medication use, including adverse reactions, medication errors, and other drug-related injuries.[1,2] They remain a major patient-safety concern and are associated with increased mortality, prolonged hospitalization, elevated healthcare costs, disability, and death.[3,4] Although clinical trials and spontaneous reporting systems provide important safety evidence, many ADEs are underreported or documented only indirectly in electronic health records (EHRs), patient-generated posts, medical forums, case reports, and biomedical literature.[5,6] Automated methods can support earlier identification of such signals, though their effectiveness depends on how well they account for the representational differences across these sources.[7–9]

These sources are not interchangeable representations of the same pharmacovigilance evidence. They are produced for different audiences and purposes and vary in terminology, language, documentation practices, and degree of standardization. Clinical and scientific sources may explicitly describe drug–outcome relations using specialized terminology, whereas social-media posts and patient-forum narratives often contain misspellings, slang, multilingual expressions, implicit symptom descriptions, and extended accounts of lived experience.[10] In this study, we use the term *data-production context* to capture these differences. Through cross-source topic modeling, we examine how such contexts shape which adverse experiences are expressed and how they are framed, and how readily the resulting drug mentions can be linked to standardized pharmacovigilance knowledge. We complement this analysis with ADE detection experiments, using them as a lens to further probe how source-specific supervision, pretrained model knowledge, and external evidence interact with these representational differences, rather than as an end in themselves.

Prior discourse-oriented work has largely examined patient narratives within a single platform or population,[11,12] with relatively little attention to how ADE expression and framing compare across patient-generated, clinical, and scientific sources within a shared analytic frame. A related open question is the form of knowledge on which downstream detection systems rely.[13] Fine-tuning learns from corpus-specific labeled examples, whereas retrieval-

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

augmented generation (RAG) introduces external evidence at inference time. In this study, retrieval is grounded in SIDER, a curated resource containing standardized drug–adverse-effect associations derived primarily from drug labels and clinical-trial information.[14] Although such knowledge can support ADE detection, SIDER's predominantly English, standardized vocabulary may align unevenly with brand names, misspellings, colloquial forms, and language-specific drug mentions found in real-world text. Prior work has shown that SIDER-grounded RAG can improve drug–side-effect question answering on a single benchmark;[15] however, its coverage across heterogeneous data-production contexts remains unclear. Recent LLM-based ADE detection work using zero- and few-shot prompting, parameter-efficient fine-tuning, and structured knowledge construction[16–18] has likewise been evaluated largely on a single corpus or a small number of closely related sources,[19,20] with limited attention to how data-production context shapes both ADE expression and model performance.

Taken together, these gaps motivate three research questions: (1) how ADE discourse differs across patient-generated, clinical, and scientific sources; (2) how well drug mentions from these settings can be linked to SIDER; and (3) how corpus-specific fine-tuning, prompting, and SIDER-grounded inference perform across different data-production contexts, as a means of probing the downstream consequences of (1) and (2).

Figure 1 summarizes the study design. As shown in Panel 1, we analyze six corpora representing distinct data-production contexts: ADE Corpus V2 medical case reports,[21] SMM4H-2026 Task 1 multilingual user-generated health content,[22] CADEC V2 patient-forum narratives,[23] the KronosRX and Dutch ADE Corpus EHR clinical notes,[24] and TwiMed-PubMed biomedical literature.[25] Panel 2 presents the shared BERTopic analysis[26] of ADE-positive English-language texts concerning antidepressants and antihypertensives from four corpora, which we use to examine how adverse experiences differ across data-production contexts and how they relate to SIDER coverage. Panel 3 presents the ADE detection approaches, including corpus-specific QLoRA fine-tuning of Llama-3.2-3B[27] and SIDER-grounded inference with Llama-3.1-70B and Llama-3.1-405B, alongside zero- and few-shot prompting baselines, which together let us assess how readily real-world drug mentions from each corpus can be linked to the standardized external knowledge resource and how this alignment relates to detection outcomes.

Our results reveal corpus-dependent differences in ADE discourse, knowledge-base alignment, and detection performance. The topic analysis shows that patient-generated, clinical, and scientific sources differ in the adverse experiences they foreground and in the level of thematic specificity with which those experiences are described. The retrieval analysis reveals uneven alignment between real-world drug mentions and SIDER's standardized vocabulary, with coverage ranging from 9.5% in patient-forum narratives to 91.0% in biomedical literature. Detection performance also varies by corpus, with no single configuration dominating across all six settings (Section 4.1). Together, these findings demonstrate that data-production context shapes what ADE evidence is expressed, how it is framed and standardized, and how readily it can be retrieved and computationally detected.

2. Background and Related Work

This section reviews two areas relevant to this study: variation in pharmacovigilance evidence across data-production contexts, and computational approaches—detection models and knowledge-grounded retrieval—that depend on corpus-specific supervision or structured pharmacovigilance knowledge.

ADE Evidence Across Data-Production Contexts. Pharmacovigilance increasingly draws on evidence from patient-generated content, electronic health records (EHRs), medical case reports, biomedical literature, and curated drug-safety resources. These sources are complementary but not interchangeable. Patient forums and social media platforms provide direct accounts of medication use and adverse experiences, often expressed in colloquial, subjective, and multilingual language. Comparative reviews suggest that patient-generated sources can supplement spontaneous reports, product labels, and clinical literature by contributing timely or unexpected adverse-event evidence and information about how adverse experiences affect patients' daily lives.[9] At the same time, such data exhibit substantial linguistic variability, incomplete context, and platform-specific biases. Clinical notes contain longitudinal observations and documentation of suspected ADEs, but relevant evidence is often expressed indirectly in unstructured text and varies across institutions and documentation practices.[9] Medical case reports and biomedical publications generally use more formal terminology and more explicit drug–outcome relations, but they are also shaped by clinical selection and scientific reporting practices.[21,25] Consequently, each data-production context makes different aspects of drug-related harm visible and available for computational analysis.

Computational content analysis has been used to study patient discussions of health experiences and medication use. Wexler *et al.* identified recurring concerns in online patient forums,[11] while Chen *et al.* showed that combining adverse-effect extraction, signal detection, and topic modeling can reveal contextual information beyond event extraction alone.[12] BERTopic further enables semantically related documents to be organized in a shared topic space.[26] However, prior work has largely focused on single platforms, populations, or medications, with limited cross-source analysis of ADE discourse and its relationship to pharmacovigilance knowledge alignment and detection performance.

Computational Approaches to Detection and Knowledge Grounding. ADE detection has progressed from conventional classifiers such as support vector machines[28,29] to transformer-based models, including BERT,[30] Bio-ClinicalBERT,[31] and RoBERTa,[32] and more recently to LLMs adapted through prompting, fine-tuning, and data augmentation.[10,17,33] Domain-adapted models often perform well when trained and evaluated within a single corpus,[34,35] but Dai *et al.* showed through the MultiADE benchmark that models trained on one ADE domain did not generalize reliably to unseen text types,[23] underscoring the influence of source language, annotation practices, and task formulation. Most LLM-based studies, however, remain limited to one or two source types and rarely examine source-specific supervision and structured external knowledge within the same cross-source analysis—a gap we address by evaluating prompting, QLoRA fine-tuning, and SIDER-grounded inference across six heterogeneous corpora.

Two adaptation strategies dominate this space and rely on different sources of pharma-

covigilance evidence. Parameter-efficient fine-tuning methods such as LoRA and QLoRA freeze most pretrained parameters and update only low-rank weights, enabling corpus-specific adaptation with reduced memory requirements.[36] Prior work shows that QLoRA can approach full fine-tuning performance in clinical summarization and bilingual primary-care notes while using far fewer trainable parameters.[27,36] RAG, by contrast, grounds inference in external evidence without updating model parameters, making performance dependent on the coverage and terminology of the retrieved resource. Examples include MALADE, which uses drug labels, biomedical literature, and FDA resources for drug–outcome extraction;[37] SIDER-based dense retrieval for drug–side-effect queries;[15] and AskFDALabel, which retrieves FDA labeling for adverse-event annotation.[13] These approaches demonstrate the value of external pharmacovigilance knowledge, but primarily target querying or structured association extraction rather than ADE detection across heterogeneous text sources. We therefore compare local supervision and external knowledge grounding across multiple data-production contexts, with particular attention to how SIDER coverage relates to detection performance.

3. Methods

We combine cross-source discourse analysis with ADE detection experiments to examine how data-production context shapes both the expression of adverse drug events and the performance of computational detection methods. Figure 1 summarizes the overall study design. Panel 1 introduces the six ADE corpora, which represent distinct languages and data-production settings, and shows their preprocessing into a common binary ADE detection task. Panel 2 presents a cross-source analysis of how adverse drug experiences are discussed in patient-generated text, medical case reports, clinical notes, and biomedical literature. Panel 3 evaluates corpus-specific QLoRA fine-tuning, zero- and few-shot prompting, and SIDER-grounded retrieval-augmented inference for ADE detection, together with instance-level SIDER retrieval coverage as a measure of how well real-world drug mentions align with standardized pharmacovigilance knowledge.

3.1. *Data Sources and Task Construction*

We use six expert-annotated ADE corpora spanning patient-generated, clinical, and scientific data-production contexts. Because the corpora differ in annotation schemes and units of analysis (sentence-, document-, and relation-level), we map their outputs to a shared binary label space (`ADE`, `Not-ADE`) while retaining corpus-specific operational criteria. Performance is therefore evaluated within each corpus.

ADE Corpus V2 comprises medical case reports with sentence-level ADE annotations (17,637 train, 5,879 test). Following reports of cross-split duplication,[19] we normalize case and whitespace and remove exact train–test duplicates.

SMM4H-2026 Task 1 addresses document-level ADE detection in multilingual user-generated health content (German, French, Russian, English, Mandarin, Japanese), using predefined training, validation, and test splits with normalized formatting.

CADEC V2 extends the AskAPatient corpus with drug and adverse-reaction annotations. We convert `brat` annotations to sentence-level binary labels and construct document-level splits to prevent sentences from the same review appearing across partitions.

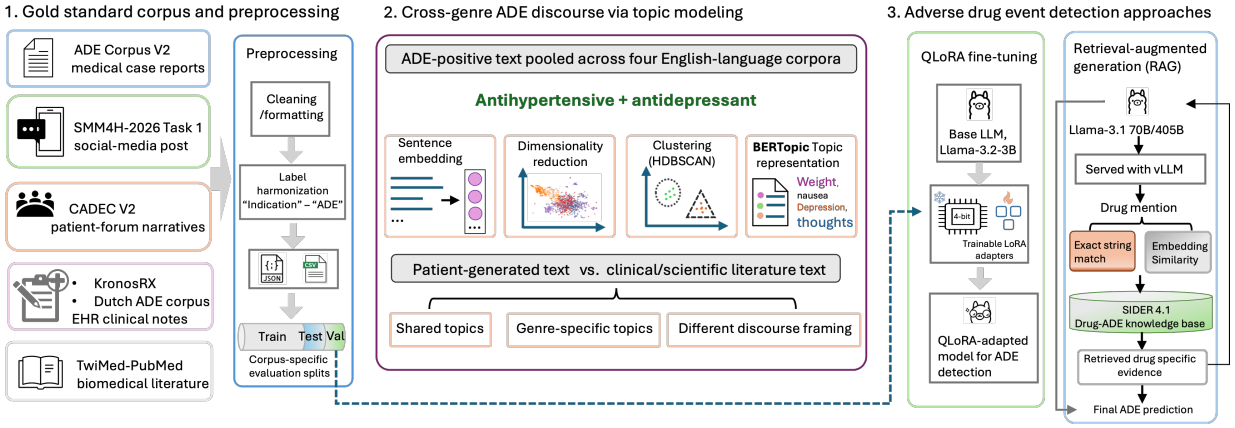


Fig. 1: Overview of the study design. Six ADE corpora spanning distinct data-production settings are standardized for binary ADE detection. We analyze cross-source discourse with BERTopic and compare source-specific fine-tuning with SIDER-grounded retrieval-augmented inference across documentation settings.

Dutch ADE Corpus consists of ICU clinical notes annotated with drug–disorder relations. We map adverse drug events to `ADE` and indications to `Not-ADE`, and split the 102 notes at the note level (70/15/15) to prevent leakage.

TwiMed contains PubMed sentences and Twitter posts annotated for drugs and outcomes; we use the PubMed subset only. Sentences containing at least one outcome-negative relation are labeled `ADE`, and we construct a stratified 80/20 train–test split.

KronosRX is a non-public, physician-adjudicated EHR benchmark derived from Cedars-Sinai clinical data, with cohorts for medication-related cardiotoxicity and seizure-related neurotoxicity. Model predictions are made at the note level and aggregated to encounters using an any-positive rule. The cardiotoxicity cohort contains 583 encounters from 533 patients, split into 404/92/87 train/development/test encounters; the test set contains 22 ADE-positive and 65 Not-ADE encounters. The neurotoxicity cohort contains 88 encounters from 80 patients (12 ADE-positive, 76 Not-ADE) and is evaluated as a single external cohort.

3.2. *Cross-Genre ADE Discourse via Topic Modelling*

To examine how ADE discourse varies across data-production contexts, we focus on antihypertensives and antidepressants, two widely used drug classes associated with diverse adverse effects and represented across multiple corpora. The analysis includes ADE-positive English-language texts from ADE Corpus V2, CADEC V2, TwiMed-PubMed, and the English subset of SMM4H-2026 Task 1.

Drug mentions are obtained from gold-standard entity annotations for CADEC V2 and TwiMed-PubMed. Because ADE Corpus V2 and SMM4H-2026 do not provide complete drug-entity annotations for the classification instances, mentions are extracted using Llama-3.1-405B and verified against the original text. The Dutch ADE Corpus is excluded from this analysis because its Dutch-language texts are not directly comparable within the shared English-language topic space and provide limited representation of the selected drug classes; it remains included in the SIDER-coverage and ADE-detection analyses.

ADE-positive instances mentioning either drug class are pooled across the four corpora. To reduce clustering by medication identity, extracted drug mentions are replaced with a generic placeholder before embedding. A single BERTopic model[26] is fitted to the combined drug-masked texts, using sentence embeddings, UMAP, HDBSCAN, and class-based TF-IDF to derive topics in a shared space. Topics are labeled from their highest-weighted terms and representative sentences. Topic prevalence is calculated within each corpus as the proportion of qualifying instances assigned to each topic, and prevalence, terms, and representative examples are compared across sources to identify shared and source-specific themes. Because corpus sizes differ substantially, prevalence is interpreted comparatively rather than as a population-frequency estimate.

3.3. *Adverse Drug Event Detection Approaches*

We evaluate three ADE detection settings: zero- and few-shot prompting, corpus-specific QLoRA fine-tuning, and SIDER-grounded retrieval-augmented inference. All configurations use corpus-specific instructions reflecting the unit of analysis and operational ADE criteria.

Zero- and Few-Shot Inference. Llama-3.2-3B-Instruct, Llama-3.3-70B-Instruct, and Llama-3.1-405B-Instruct are evaluated under zero-shot prompting. Each prompt includes the corpus-specific task definition, the input (sentence, document, or drug–disorder context), and an instruction to return one of two labels: `ADE` or `Not-ADE`. Llama-3.3-70B-Instruct and Llama-3.1-405B-Instruct are additionally evaluated under few-shot prompting with labeled demonstrations. These baselines do not include retrieved SIDER evidence.

QLoRA Fine-Tuning. We fine-tune a separate Llama-3.2-3B-Instruct model for each corpus using Quantized Low-Rank Adaptation (QLoRA).[38] The base model is loaded with 4-bit NormalFloat quantization (double quantization), with computation in `bfloat16` precision. Following standard LoRA parameterization, low-rank adapters (rank $r = 16$, scaling $\alpha = 32$, dropout 0.05) are applied to all linear layers; base-model parameters remain frozen and only adapter weights are updated. Each training instance is formatted as a corpus-specific instruction-following exchange using the Llama chat template, where the system message defines the task, the user message contains the input, and the assistant message provides the gold label (`ADE` or `Not-ADE`). Models are trained with a causal language-modeling objective using a paged 8-bit AdamW optimizer, cosine learning-rate schedule, and gradient checkpointing.[38] At inference, the assistant response is omitted, greedy decoding is applied, and outputs are mapped to the binary label space.

SIDER-Grounded Retrieval-Augmented Inference . Our retrieval-augmented configuration introduces external pharmacovigilance evidence without task-specific parameter updates. Llama-3.3-70B-Instruct and Llama-3.1-405B-Instruct are served with vLLM and remain frozen during inference. Each input is augmented with evidence from SIDER 4.1, a curated resource containing drug–adverse-effect associations derived primarily from drug labels and clinical-trial information, with adverse effects standardized using MedDRA terminology.[14]

We constructed a knowledgebase mapping canonical drug names from SIDER and synonyms to lists of associated adverse effects (restricted to MedDRA Preferred Terms), retaining up to ten effects per entry to limit prompt length. Very short drug names are excluded from the lexical index to reduce ambiguity.

Algorithm 1 SIDER-grounded retrieval-augmented inference for ADE detection

Require: Input instance x; extracted drug mentions $M(x)$; target drug m and disorder a when explicitly annotated; corpus and language identifier g; SIDER knowledge base $\mathcal{D}$; precomputed SIDER embeddings $\mathcal{V}_g$; retrieval size k; similarity threshold τ_g
Ensure: Predicted label $\hat{y} \in \{\text{ADE}, \text{Not-ADE}\}$

$\mathcal{E} \leftarrow \varnothing$
if $g = \text{DutchADE}$ **then**
  $\mathcal{C} \leftarrow \text{NormalizeDutchDrug}(m)$
  $\mathcal{E} \leftarrow \text{LexicalMatch}(\mathcal{C}, \mathcal{D})$
  if $\mathcal{E} = \varnothing$ **then**
    $\mathbf{q} \leftarrow f_g(m)$
    $\mathcal{E} \leftarrow \text{SemanticRetrieve}(\mathbf{q}, \mathcal{D}, \mathcal{V}_g, k, \tau_g)$
  end if
else
  $\mathcal{E}_{\text{lex}} \leftarrow \text{ExactMatch}(M(x), \mathcal{D})$
  $\mathbf{q} \leftarrow f_g(x)$
  $\mathcal{E}_{\text{sem}} \leftarrow \text{SemanticRetrieve}(\mathbf{q}, \mathcal{D}, \mathcal{V}_g, k, \tau_g)$
  $\mathcal{E} \leftarrow \text{MergeExactFirst}(\mathcal{E}_{\text{lex}}, \mathcal{E}_{\text{sem}})$
end if
$p \leftarrow \text{BuildCorpusSpecificPrompt}(x, M(x), m, a, \mathcal{E}, g)$
$\hat{y}_{\text{raw}} \leftarrow \text{LLM}(p, \text{temperature} = 0)$
$\hat{y} \leftarrow \text{ParseLabel}(\hat{y}_{\text{raw}})$
return $\hat{y}$

Fig. 2: Extracted drug mentions are matched lexically and semantically against a SIDER-based knowledge base, with a sequential lexical-first branch for Dutch ADE notes and a parallel lexical+semantic branch for English and multilingual corpora. Retrieved evidence is inserted into a corpus-specific prompt, and a frozen Llama model predicts a binary `ADE`/`Not-ADE` label under deterministic decoding.

Retrieval combines exact lexical matching with embedding-based semantic retrieval and is adapted to the language and annotation structure of each corpus. For English-language inputs, lexical matching is applied to extracted drug mentions $M(x)$, while semantic retrieval uses the full input text as a query. The two evidence sources are merged, prioritizing exact matches when duplicates occur. Semantic retrieval uses `BAAI/bge-large-en-v1.5` for English inputs and `paraphrase-multilingual-MiniLM-L12-v2` for cross-lingual inputs, retaining the top $k = 3$ entries above corpus-specific similarity thresholds ($\tau_{\text{en}} = 0.80$, $\tau_{\text{nl}} = 0.62$).

For the Dutch ADE Corpus, retrieval follows a sequential strategy: the target drug entity is first normalized using rule-based transformations to approximate English forms (e.g., `-ine`→`-in`, `-ole`→`-ol`), and lexical matching is attempted before falling back to cross-lingual semantic retrieval if needed. The SMM4H corpus uses the English retrieval branch for English-language instances and the cross-lingual branch otherwise.

Retrieved evidence is inserted into a corpus-specific prompt containing the original input and any available drug or disorder. The model is instructed to use the retrieved information as supporting evidence and to generate `ADE` or `Not-ADE`. Inference uses deterministic decoding (temperature 0), and outputs are normalized and mapped to the binary label space. We evaluate both exact-match-only and hybrid lexical–semantic retrieval settings where applicable.

4. Results

SIDER Coverage Across Data-Production Contexts. Table 1 summarizes SIDER retrieval coverage across the corpora. For the original five datasets, exact coverage denotes lexical matching to a SIDER drug name or synonym, while semantic-only coverage denotes embedding-based retrieval when lexical matching fails. KronosRX uses task-specific lexical retrieval only, with coverage measured at the note level.

Coverage varied substantially across data-production settings. TwiMed-PubMed showed the highest coverage among the original corpora (91.0%), consistent with the standardized medication terminology common in biomedical literature. KronosRX also showed high lexical

Table 1: SIDER retrieval coverage across corpora. Semantic-only denotes embedding-based retrieval after lexical matching fails; KronosRX uses task-specific lexical retrieval only.

Corpus	Exact (%)	Semantic-only (%)	Total (%)
ADE Corpus V2	52.0	0.0	52.0
SMM4H-2026	4.7	58.3	63.0
Dutch ADE	28.2	2.5	30.6
CADEC V2	9.5	0.0	9.5
TwiMed-PubMed	91.0	0.0	91.0
KronosRX-Cardio	91.7	N/A	91.7
KronosRX-Neuro	81.7	N/A	81.7

coverage, reaching 91.7% for cardiotoxicity and 81.7% for neurotoxicity. In contrast, SMM4H-2026 had very low exact coverage (4.7%), but semantic retrieval increased total coverage to 63.0%, highlighting the value of embedding-based retrieval for multilingual and linguistically variable user-generated content.

ADE Corpus V2 and Dutch ADE showed intermediate coverage (52.0% and 30.6%, respectively), whereas CADEC V2 remained lowest at 9.5%. The low coverage in CADEC reflects mismatches such as brand names, spelling variation, abbreviations, and colloquial drug references. Overall, these patterns indicate that the usefulness of curated pharmacovigilance resources depends strongly on the terminology and documentation practices of the source, which helps explain why retrieval-based configurations are more competitive in some corpora than others.

Cross-Genre Patterns in ADE Discourse. Figures 3 and 4 summarize the pooled BERTopic analysis of 6,713 drug-masked, ADE-positive instances from the four English-language corpora, comprising 6,001 instances from CADEC V2, 466 from SMM4H-2026, 181 from ADE Corpus V2, and 65 from TwiMed-PubMed. The model identifies nine interpretable topics in addition to an outlier cluster. Sexual side effects, neurological symptoms, and swelling or localized pain appear across multiple corpora, whereas suicidal or panic-related thoughts, night sweats, vivid dreams, and memory difficulties are concentrated in CADEC V2. These patterns suggest that long-form patient narratives provide greater visibility into emotionally salient and highly specific experiences that are rarely represented in shorter or more formally structured sources.

Medication/Sleep/Tired is the largest topic in the pooled corpus and dominates SMM4H-2026, TwiMed-PubMed, and ADE Corpus V2, accounting for 50.4%, 66.2%, and 66.9% of their qualifying instances, respectively. In CADEC V2, *Weight/Nausea/Loss* (18.7%) and *Medication/Sleep/Tired* (18.2%) are similarly prevalent, with the remaining instances distributed across a broader range of symptom-specific topics and outliers. Because CADEC V2 contributes most of the pooled instances, prevalence estimates for ADE Corpus V2 and TwiMed-PubMed should be interpreted descriptively rather than as precise frequency estimates. Consistent with this pattern, Figure 4 shows that topics with more distinctive vocabularies form relatively compact clusters, whereas the two largest and more general topics occupy more diffuse regions of the embedding space.

Cross-Genre ADE Discourse Topics: Antihypertensive + Antidepressant Case Study (CADEC V2, SMM4H-2026, TwiMed, ADE Corpus V2)

Topic Label	Word Cluster	Total Sentences	CADEC V2 (%)	SMM4H-2026 (%)	TwiMed (%)	ADE Corpus V2 (%)
Medication / Sleep / Tired	medication, sleep, tired, pain, day, withdrawal, taking, heart	1490	18.2	50.4	66.2	66.9
Weight / Nausea / Loss	weight, nausea, loss, gain, appetite, constipation, dizziness, mouth	1147	18.7	4.7	1.5	1.7
Zaps / Brain / Headaches	zaps, brain, headaches, dizzy, dizziness, vision, vertigo, headache	300	4.8	1.7	3.1	0.6
Orgasm / Sex / Libido	orgasm, sex, libido, drive, loss, decreased, sexual, delayed	225	3.5	1.3	3.1	2.8
Suicidal / Panic / Thoughts	suicidal, panic, thoughts, depression, anxiety, depressed, attacks, suicide	215	3.6	0.2	0.0	0.0
Sweating / Sweats / Night	sweating, sweats, night, hot, sweat, flashes, drenched, cold	207	3.4	0.9	0.0	0.0
Swelling / Pain / Ankles	swelling, pain, ankles, legs, itching, rash, muscle, skin	198	3.1	1.7	4.6	1.1
Dreams / Vivid / Nightmares	dreams, vivid, nightmares, dream, night, scary, sleep, violent	96	1.6	0.2	0.0	0.0
Memory / Concentrate / Term	memory, concentrate, term, remember, word, short, focus, words	83	1.4	0.0	0.0	0.0



Fig. 3: Cross-genre ADE discourse topics identified by BERTopic (antihypertensive + antidepressant case study, CADEC V2, SMM4H-2026, TwiMed, ADE Corpus V2).

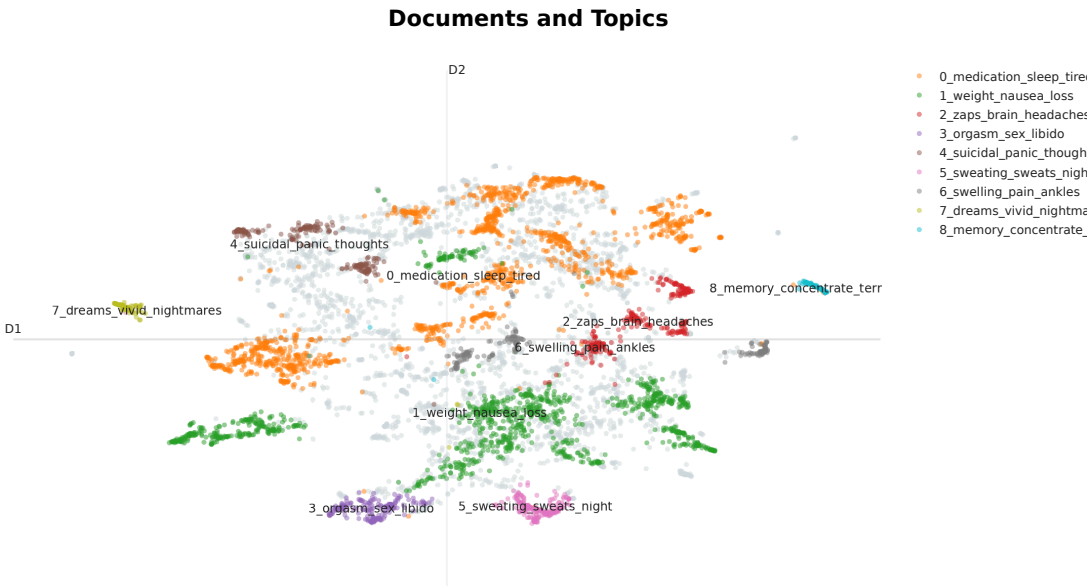


Fig. 4: UMAP projection of drug-masked ADE-positive instances colored by BERTopic topic assignment. Gray points denote outliers. Topics with distinctive vocabularies form tight clusters, whereas the two largest, more generic topics (Medication/Sleep/Tired and Weight/Nausea/Loss) occupy a more diffuse region of the embedding space.

Topic prevalence is broadly similar across antidepressant- and antihypertensive-related instances, supporting their representation within a shared topic space. The most notable differences are clinically plausible: swelling and ankle pain are more prevalent among antihypertensive-related instances, whereas sexual side effects are more prevalent among antidepressant-related instances. Overall, the findings indicate that patient forums provide non-redundant pharmacovigilance evidence, particularly for subjective, psychologically salient,

and highly specific adverse experiences that are less visible in social-media posts, medical case reports, and biomedical literature. From a sociotechnical perspective, these differences reflect how platform structure, documentation purpose, intended audience, and opportunities for extended narration shape which adverse experiences become visible for computational analysis.

Detection Performance Across Corpora. Table 2 compares zero-shot, few-shot, QLoRA, and RAG-SIDER configurations across the five corpora. Because the datasets are generally imbalanced toward `Not-ADE`, we report $F1_{\text{macro}}$ together with ADE-class precision and recall to distinguish overall classification performance from performance on the clinically relevant ADE class. Among the evaluated configurations, QLoRA-Llama-3.2-3B obtained the highest observed $F1_{\text{macro}}$ on ADE Corpus V2, SMM4H-2026, and CADEC V2 (0.91, 0.80, and 0.88, respectively). These three corpora provide more labeled training examples than the Dutch ADE Corpus and TwiMed-PubMed, suggesting that corpus-specific fine-tuning may be particularly effective when sufficient in-domain supervision is available. Under these evaluation settings, the compact QLoRA-adapted model performed comparably to or better than the substantially larger inference-only models evaluated in this study.

RAG-SIDER with Llama-3.1-405B attains the highest $F1_{\text{macro}}$ on the Dutch ADE Corpus and TwiMed-PubMed (0.95 and 0.91), and on the Dutch ADE Corpus achieves an ADE-class recall of 0.95, indicating high sensitivity to positive ADE instances. Few-shot prompting yields inconsistent changes relative to zero-shot prompting across corpora, with clear improvements in some settings and minimal or no benefit in others.

KronosRX external EHR evaluation. Table 3 shows a similar pattern in long-form English EHR data. On the cardiotoxicity test set, zero-shot Llama-3.1-405B achieved perfect ADE recall but produced substantial overprediction ($P_{\text{ADE}} = 0.35$). Few-shot prompting increased $F1_{\text{macro}}$ from 0.54 to 0.68, while the cardiac-specific RAG-SIDER configuration further increased it to 0.75 and improved P_{ADE} to 0.54, with recall remaining 0.86. For neurotoxicity, seizure-specific RAG-SIDER increased $F1_{\text{macro}}$ from 0.58 to 0.64 and precision from 0.27 to 0.31, while recall decreased from 1.00 to 0.92. In KronosRX, the toxicity-specific RAG configurations therefore reduced false-positive predictions while maintaining high sensitivity in encounter-level EHR evaluation.

Table 2: ADE detection performance across five corpora. Each cell reports $F1_{\text{macro}}/P_{\text{ADE}}/R_{\text{ADE}}$. Results are single-run descriptive estimates. Comparisons across model families and configurations are not statistically controlled because model size, supervision, and retrieval evidence differ.

Model	Configuration	ADE Corpus V2	SMM4H-2026	Dutch ADE	CADEC V2	TwiMed-PubMed
Llama-3.2-3B	Zero-shot	0.51/0.40/0.80	0.55/0.15/0.20	0.62/0.72/0.28	0.65/0.58/0.48	0.70/0.64/0.90
Llama-3.2-3B	QLoRA	0.91/0.85/0.82	0.80/0.76/0.53	0.87/0.94/0.75	0.88/0.87/0.83	0.85/0.75/0.77
Llama-3.3-70B	Zero-shot	0.67/0.40/0.95	0.73/0.39/0.79	0.83/0.95/0.61	0.70/0.66/0.82	0.88/0.76/0.90
Llama-3.3-70B	Few-shot	0.73/0.49/0.86	0.74/0.41/0.80	0.88/0.90/0.70	0.71/0.66/0.80	0.87/0.73/0.89
Llama-3.3-70B	RAG-SIDER	0.74/0.50/0.86	0.75/0.42/0.80	0.91/0.87/0.80	0.73/0.66/0.85	0.89/0.78/0.89
Llama-3.1-405B	Zero-shot	0.70/0.34/0.97	0.75/0.40/0.85	0.92/0.90/0.89	0.77/0.69/0.87	0.89/0.78/0.89
Llama-3.1-405B	Few-shot	0.76/0.49/0.85	0.77/0.45/0.76	0.92/0.89/0.85	0.77/0.69/0.80	0.89/0.78/0.90
Llama-3.1-405B	RAG-SIDER	0.78/0.52/0.93	0.78/0.48/0.80	0.95/0.89/0.95	0.80/0.68/0.74	0.91/0.82/0.89

Table 3: Encounter-level KronosRX results ($F1_{\text{macro}}/P_{\text{ADE}}/R_{\text{ADE}}$).

Cohort	Zero-shot	Few-shot	RAG-SIDER
Cardiotoxicity ($n = 87$)	0.54/0.35/1.00	0.68/0.45/0.95	0.75/0.54/0.86
Neurotoxicity ($n = 88$)	0.58/0.27/1.00	–	0.64/0.31/0.92

Across corpora, the strongest configuration is source-dependent: QLoRA performs best in several corpora with stronger task-specific supervision, whereas large-model RAG-SIDER is more competitive when external pharmacovigilance knowledge aligns well with the source and task. These patterns should not be interpreted as a controlled comparison of adaptation strategies, because the configurations differ in model scale, supervision, and retrieval evidence. Direct comparison with prior work is also limited by differences in task formulation and splits (for example, our ADE Corpus V2 experiments use a leakage-corrected split, and the Dutch ADE Corpus and CADEC V2 lack established binary baselines). We therefore emphasize comparisons among configurations evaluated on the same held-out test set within each corpus.

5. Discussion and Limitations

Our results show that data-production context strongly shapes how adverse drug experiences are expressed and how well they align with curated pharmacovigilance resources. Patient forums, social media, case reports, EHR notes, and biomedical literature are therefore not interchangeable sources of ADE evidence. SIDER coverage varies substantially across corpora, from 91.0% exact matches in TwiMed-PubMed to 9.5% in CADEC V2, reflecting its stronger alignment with standardized biomedical terminology than with patient-generated or multilingual language. Topic analysis similarly shows that CADEC V2 foregrounds subjective and psychologically salient experiences, whereas other sources emphasize broader medication-, sleep-, tiredness-, and pain-related themes. Model performance reflects these differences. QLoRA-Llama-3.2-3B performs well when sufficient in-domain supervision is available, whereas RAG-SIDER with Llama-3.1-405B is more competitive when external drug knowledge aligns with the source. The KronosRX EHR evaluation shows a related pattern: toxicity-specific RAG reduced false-positive encounter predictions while maintaining high recall. Together, these findings suggest that no single adaptation strategy is optimal across data sources; model design should account for both available supervision and knowledge-base alignment.

Several limitations remain. Comparisons confound model scale and training-set size, use binary ADE detection and single-run evaluations, and rely on LLM-assisted drug-mention extraction for some corpora. Repeated runs and confidence intervals were not available; therefore, the reported performance differences should be interpreted as descriptive observations rather than statistically significant differences. Topic estimates are also influenced by the larger CADEC V2 corpus, while KronosRX was evaluated only with 405B inference configurations rather than a matched 3B fine-tuned model. Future work should use matched model scales, repeated runs, joint extraction, and long-EHR training strategies such as evidence-preserving chunking or encounter-aware objectives. More broadly, combining source-sensitive models with knowledge-grounded retrieval and coverage auditing may help pharmacovigilance systems capture both documented harms and patient-expressed adverse experiences.

Ethics Statement

The KronosRX study was reviewed and approved by the Cedars-Sinai Institutional Review Board under protocol STUDY00004421 (KronosRx: Sending Patient Avatars to the Future to Transform Drug Safety Prediction). The study received expedited IRB approval effective August 24, 2025.

Acknowledgments

This work was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM014731 and by the Advanced Research Projects Agency for Health (ARPA-H) CATALYST program through the KronosRx project, 'KronosRx: Sending Patient Avatars to the Future to Transform Drug Safety Prediction' (ARPA-H Award 3586 - https://arpa-h.gov/explore-funding/awards/3586) The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or ARPA-H.

References

1. M. S. Donaldson, J. M. Corrigan and L. T. Kohn, To err is human: building a safer health system (2000).
2. X. Qiu, S. Shao, H. Wang and X. Tan, Bio-k-transformer: A pre-trained transformer-based sequence-to-sequence model for adverse drug reactions prediction, *Computer Methods and Programs in Biomedicine* **260**, p. 108524 (2025).
3. D. C. Classen, S. L. Pestotnik, R. S. Evans, J. F. Lloyd and J. P. Burke, Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality, *Jama* **277**, 301 (1997).
4. M. De Rosa, G. Fenza, A. Gallo, M. Gallo and V. Loia, Pharmacovigilance in the era of social media: discovering adverse drug events cross-relating twitter and pubmed, *Future Generation Computer Systems* **114**, 394 (2021).
5. E. Kopacheva, A. Henriksson, H. Dalianis, T. Hammar and A. Lincke, Identifying adverse drug events in clinical text using fine-tuned clinical language models: Machine learning study, *JMIR Formative Research* **9**, p. e71949 (2025).
6. S. Golder, K. O'Connor, G. Lopez-Garcia, N. P. Tatonetti and G. Gonzalez-Hernandez, Leveraging unstructured data in electronic health records to detect adverse events from pediatric drug use: A scoping review, *Annual review of biomedical data science* **8**, 227 (2025).
7. L. Chen, Y. Gu, X. Ji, Z. Sun, H. Li, Y. Gao and Y. Huang, Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning, *Journal of the American Medical Informatics Association* **27**, 56 (2020).
8. I. Guellil, Y. Berrachedi, N. E. Chenni, M.-N. Abboud, J. Wu, H. Wu and B. Alex, Detecting adverse drug events in social media: A brief literature review, *SN Computer Science* **7**, p. 199 (2026).
9. S. Golder, K. O'Connor, Y. Wang, A. Klein and G. Gonzalez Hernandez, The value of social media analysis for adverse events detection and pharmacovigilance: scoping review, *JMIR public health and surveillance* **10**, p. e59167 (2024).
10. O. Elbiach, H. Grissette *et al.*, Benchmarking large language models for adverse drug reaction extraction in social media and clinical texts, *Results in Engineering* , p. 107362 (2025).
11. A. Wexler, A. Davoudi, D. Weissenbacher, R. Choi, K. O'Connor, H. Cummings and G. Gonzalez-

Hernandez, Pregnancy and health in the age of the internet: A content analysis of online "birth club" forums, *PloS one* **15**, p. e0230947 (2020).
12. X. Chen, C. Faviez, S. Schuck, A. Lillo-Le-Louët, N. Texier, B. Dahamna, C. Huot, P. Foulquié, S. Pereira, V. Leroux *et al.*, Mining patients' narratives in social media for pharmacovigilance: adverse effects and misuse of methylphenidate, *Frontiers in pharmacology* **9**, p. 541 (2018).
13. L. Wu, H. Fang, Y. Qu, J. Xu and W. Tong, Leveraging fda labeling documents and large language model to enhance annotation, profiling, and classification of drug adverse events with askfdalabel: L. wu et al., *Drug Safety* **48**, 655 (2025).
14. M. Kuhn, I. Letunic, L. J. Jensen and P. Bork, The sider database of drugs and side effects, *Nucleic acids research* **44**, D1075 (2016).
15. S. Nygren, O. Erdogan, P. Avci, A. Daniels, R. Rassool, A. Beheshti and D. Galeano, Rag-based architectures for drug side effect retrieval using compact llms, *Scientific Reports* **16**, p. 12754 (2026).
16. H. Prioleau, S. K. Aryal and J. Blackstone, Leveraging large language models for adverse drug event detection: A comparative study of token and span-based named entity recognition, 205 (2025).
17. F. Zeidi, R. Christof, R. König and L. Childs, 'pei at# smm4h-heard 2025: Enhancing adverse event detection via error-driven data augmentation, p. 56 (2025).
18. Z. Duan, K. Wei, Z. Xue, J. Zhou, S. Yang, S. Ma, J. Jin *et al.*, Crowdsourcing-based knowledge graph construction for drug side effects using large language models with an application on semaglutide, *arXiv preprint arXiv:2504.04346* (2025).
19. Z. Huang, J. Lu, Q. Sun and H. Lu, Multi-axis evaluation of adverse drug event classification: Clinical error cost, linguistic failure taxonomy, and leakage-corrected benchmarks on ade corpus v2, 223 (2026).
20. F. Dong, W. Guo, J. Liu, T. A. Patterson and H. Hong, Bert-based language model for accurate drug adverse event extraction from social media: implementation, evaluation, and contributions to pharmacovigilance practices, *Frontiers in Public Health* **12**, p. 1392180 (2024).
21. 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* **45**, 885 (2012).
22. G. Lopez-Garcia, J. Cortina, J. Berkowitz, J. Chan, S. K. Dey, I. F. Amaro, F. Gallego, L. Gryboski, A. Z. Klein, F. Z. Kolehparcheh *et al.*, Overview of the 11th social media mining for health (# smm4h) and health real-world data (heard) shared tasks at acl 2026, 353 (2026).
23. X. Dai, S. Karimi, A. Sarker, B. Hachey and C. Paris, Multiade: A multi-domain benchmark for adverse drug event extraction, *Journal of Biomedical Informatics* **160**, p. 104744 (2024).
24. R. M. Murphy, D. A. Dongelmans, N. F. de Keizer, R. J. Jongeneel, C. H. Koster, K. J. Jager, A. Abu-Hanna, I. Calixto and J. E. Klopotowska, Creation of a gold standard dutch corpus of clinical notes for adverse drug event detection: the dutch ade corpus, *Language Resources and Evaluation* **59**, 2763 (2025).
25. N. Alvaro, Y. Miyao and N. Collier, Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations, *JMIR public health and surveillance* **3**, p. e24 (2017).
26. M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, *arXiv preprint arXiv:2203.05794* (2022).
27. P. R. Shakya, A. Khaneja and K. B. Wagholikar, For clinical data extraction, qlora attains accuracy close to lora while requiring lower compute resources, *medRxiv* , 2025 (2025).
28. J. Liu, G. Wang and G. Chen, Identifying adverse drug events from social media using an improved semisupervised method, *IEEE Intelligent Systems* **34**, 66 (2019).
29. J. Wang, L. Zhao, Y. Ye and Y. Zhang, Adverse event detection by integrating twitter data and vaers, *Journal of biomedical semantics* **9**, p. 19 (2018).

30. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, **1**, 4171 (2019).
31. E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann and M. McDermott, Publicly available clinical bert embeddings, 72 (2019).
32. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
33. H. Li and H. Lin, Adverse drug reaction detection on social media based on large language models, *Information* **17**, p. 352 (2026).
34. C. Hiba, E. H. Nfaoui and C. Loqman, Fine-tuning transformer models for adverse drug event identification and extraction in biomedical corpora: A comparative study, 957 (2023).
35. K. Huang, J. Altosaar and R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, *arXiv preprint arXiv:1904.05342* (2019).
36. C. Estupiñán-Ojeda, R. J. Sandomingo-Freire, L. Padró and J. Turmo, High-fidelity parameter-efficient fine-tuning for joint recognition and linking of diagnoses to icd-10 in non-standard primary care notes, *JAMIA open* **8**, p. ooaf120 (2025).
37. J. Choi, N. Palumbo, P. Chalasani, M. M. Engelhard, S. Jha, A. Kumar and D. Page, Malade: Orchestration of llm-powered agents with retrieval augmented generation for pharmacovigilance, *arXiv preprint arXiv:2408.01869* (2024).
38. P. N. Srinivasu, A. Samudrala, P. G. Devesh, S. A. Reddy and A. A. Nuaim, Quantized low-rank adaptation in large language models for clinical text simplification, *International Journal of Computational Intelligence Systems* **19**, p. 267 (2026).