European Journal of Clinical Pharmacy
Print ISSN: 2385-409X **Online ISSN**: 3105-0409
Website: https://farmclin.com/

Original Research Article

AI-Driven ADE: Detection and Clinician-Readable Explanation of Adverse Drug Events in Open Clinical Notes

Article History:

Name of Author:
Rahul Reddy Hanumanthgari

Affiliation: *Labcorp AI Engineering, NC, USA*

Corresponding Author:
Rahul Reddy Hanumanthgari

Received: 25-12-2025
Revised: *08*-01-2026
Accepted: 30-01-2026
Published: 16-02-2026

Abstract: Adverse Drug Events (ADEs) documented in free-text clinical notes remain a significant challenge for patient safety due to fragmented documentation and linguistic variability. Existing detection methods, ranging from rule-based systems to statistical classifiers, offer limited explainability, reducing clinician trust and adoption. This paper presents a hybrid framework that combines classical machine learning with large language model (LLM) reasoning for ADE detection and explanation across a large-scale synthetic clinical dataset comprising over 5,700 patients and 338,000 encounters. We establish a Logistic Regression baseline trained on reference-based ground-truth labels derived from SNOMED-CT codes and established drug–adverse event knowledge, and then augment it with GPT-4o-mini-based structured extraction that produces clinician-readable explanations, including suspected drug identification, severity classification, pharmacological mechanisms, and drug interaction analysis. Results demonstrate that the ML baseline provides efficient large-scale screening while the LLM layer adds clinically actionable reasoning that statistical models cannot capture. This work presents a reproducible, cost-efficient hybrid pipeline to advance automated pharmacovigilance.

Keywords: Adverse Drug Events (ADE); Clinical NLP; Large Language Models; Explainable AI; Electronic Health Records; Healthcare AI; Clinical Decision Support.

INTRODUCTION

Adverse Drug Events (ADEs) are a major contributor to patient harm and are frequently documented in unstructured clinical notes rather than structured fields. Extracting ADEs from free-text narratives is challenging due to linguistic variability, negation, temporal ambiguity, and difficulty in attributing symptoms to medications rather than underlying conditions. Traditional clinical natural language processing methods exhibit limited robustness to these complexities, resulting in reduced accuracy and generalizability across diverse clinical settings.

Recent advances in large language models have improved the semantic understanding of clinical text; however, most existing approaches emphasize detection performance while providing limited transparency for clinical validation. The absence of clinician-readable explanations remains a critical barrier to trust and adoption in real-world healthcare settings.

This work presents an AI-driven framework for ADE detection and a clinician-readable explanation from open clinical notes, evaluated on a large-scale dataset of 5,770 synthetic patients spanning 338,832 clinical encounters. We establish a reproducible Logistic Regression baseline using reference-based ground truth labels, compare it with GPT-4o-mini based structured extraction, and introduce an evidence-grounded explanation layer that links ADE predictions to supporting clinical evidence. Our key contributions include: (1) a reference-based labeling methodology using SNOMED-CT codes and established drug–ADE knowledge bases that avoids data leakage; (2) a comparative evaluation on over 338,000 encounters demonstrating complementary strengths of statistical and generative AI approaches; and (3) a reproducible, cost-efficient hybrid detection-plus-explanation pipeline.

RELATED WORK

Automated ADE detection from clinical narratives

has been studied for nearly two decades, driven by the observation that many safety-critical events are documented primarily in free-text notes rather than structured fields. Early work demonstrated that clinical NLP could surface adverse events from narrative documents and, in some settings, outperform traditional automated detection approaches, establishing clinical text as a high-value signal source for surveillance [1].

Much of this foundational literature relies on conventional pipelines: dictionary- or rule-based extraction of medications and symptoms, followed by feature-based supervised models (e.g., logistic regression, SVMs, CRFs) for classification and relation detection. While generally reproducible and interpretable, these approaches often perform less well in real-world notes due to negation (e.g., "no rash"), temporal expressions ("previously had"), attribution (symptom versus disease), and variability in documentation across institutions and specialties. A recent scoping review of inpatient ADE detection using NLP highlights this heterogeneity in methods and emphasizes that generalizability and deployment-readiness remain key gaps [2].

Community benchmarks have been essential for standardizing evaluation. The 2018 N2C2 shared task (Track 2) formalized ADE extraction into concept extraction and relation classification, enabling comparable, reproducible modeling and accelerating progress toward state-of-the-art systems [3, 4]. Strong systems from this challenge frequently used neural architectures that jointly model entities and relations, suggesting ADE extraction benefits from end-to-end contextual modeling rather than isolated mention detection. Subsequent studies have continued to show gains from domain-specific models over traditional ML baselines, while also emphasizing the persistent need to validate across institutions to mitigate dataset shift [5, 6]. Regulatory and applied healthcare efforts have similarly explored adapting deep learning models to real-world clinical notes for adverse event detection, reinforcing practical interest beyond academic benchmarks [7].

More recently, large language models (LLMs) have entered the ADE detection landscape, offering improved semantic coverage and reduced dependence on brittle feature engineering. Emerging literature suggests LLMs can perform structured extraction under strict output constraints, but also highlights risks: hallucinated rationales, inconsistent outputs, and uncertain generalization without strong task constraints and evaluation. Narrative reviews summarizing LLM usage in ADE contexts emphasize the importance of aligning system outputs with clinical decision-making and verification requirements [8, 9]. Preprint evidence also explores the use of LLMs for identifying treatment-emergent adverse events from notes, underscoring active research momentum and the need for rigorous validation [10].

A recurring limitation across ADE detection research is that higher accuracy alone is insufficient for adoption in clinical workflows. Clinicians typically require traceable evidence text spans or structured rationales tied to the note so outputs can be reviewed and trusted. Reviews of ADE detection consistently identify explainability, error analysis of negation, temporality, and attribution, and human-centered evaluation as ongoing challenges and high-impact opportunities [2, 8]. The literature also reflects increasing emphasis on reproducible pipelines and rigorous evaluation, particularly given privacy constraints and site-specific documentation patterns [2].

Building on these studies, the novelty of the present work targets documented gaps through a hybrid approach that combines a baseline ML classifier with an LLM-based structured extraction and explanation layer. Unlike prior work that either relies solely on statistical methods or exclusively on LLM reasoning, our framework evaluates both approaches at scale over 338,000 clinical encounters and demonstrates their complementary strengths for ADE detection and clinical explanation [2, 5–9].

METHODOLOGY

We implement a modular, end-to-end ADE detection pipeline comprising four stages: (i) data ingestion and normalization, (ii) reference-based ground truth labeling, (iii) dual-track ADE signal detection via ML baseline and LLM reasoning, and (iv) clinician-readable explanation generation grounded in the source note. Each module is independently replaceable to accommodate privacy, cost, or latency constraints. Figure 1 illustrates the overall architecture.

3.1 Data Source and Preprocessing

We use Synthea, an open-source synthetic patient generator, to create a large-scale reproducible dataset of 5,770 patients spanning 338,832 clinical encounters, 289,494 medication records, 210,253 conditions, 5,390 allergy entries, 4,422,255 observations, and 945,798 procedures. Each encounter includes structured fields (demographics, diagnoses, prescriptions, lab results) and a free-text clinical note (338,832 total notes). Data are ingested into DuckDB for efficient analytical querying at scale. Clinical notes undergo lightweight normalization (whitespace cleanup, Unicode normalization, and section tagging where available). All data snapshots are versioned to ensure reproducibility. The dataset scales nearly 340,000 encounters across close to 6,000 patients, providing a realistic testbed for evaluating model performance under conditions approximating

production clinical volumes.

3.2 Ground Truth Labeling Strategy

A common pitfall in avoiding data leakage is deriving target labels from training features. We construct ground truth ADE labels using an external reference-based approach. The labeling pipeline draws on three independent knowledge sources: (1) a curated set of 44 SNOMED-CT codes associated with adverse drug reactions and drug-induced organ damage, matching 336 encounters; (2) 37 established drug–adverse event patterns derived from clinical pharmacology literature (e.g., warfarin–bleeding, NSAID–gastric ulcer, opioid–respiratory depression); and (3) temporal relationship analysis identifying conditions that onset within a 30-day window following medication initiation, capturing 60,536 temporally linked encounters. Additionally, 277 encounters are identified through ADE-related text keyword matching. A composite risk score aggregates these signals: encounters with SNOMED code matches, known drug–condition pairs, or text keyword matches are labeled as positive ADE cases; encounters meeting temporal and polypharmacy criteria above a threshold of 0.15 are additionally included. This yields a 21.0% ADE prevalence rate (71,260 of 338,832 encounters), reflecting the inclusion of temporal relationships alongside established pharmacological patterns.

3.3 ML Baseline Model

As a reproducible baseline, we train a Logistic Regression classifier with L2 regularization (C = 0.1, balanced class weights, LBFGS solver) on features comprising patient demographics (age, gender, race), allergy indicators, medication and condition counts, and one-hot encodings of the top 50 medications and top 50 conditions by frequency. Features are standardized using z-score normalization. The dataset is split 80/20 at the encounter level with stratified sampling to preserve class balance, yielding approximately 271,000 training encounters and 67,800 test encounters. Five-fold cross-validation on the training set provides generalization estimates. The baseline is intentionally simple and fast, providing a transparent statistical reference point for comparison with more complex approaches, while its scalability to hundreds of thousands of encounters demonstrates production viability.

3.4 LLM Reasoning Layer

To augment statistical detection with clinical reasoning, we deploy GPT-4o mini through the instructor library, which enforces structured outputs via Pydantic schemas. For each encounter flagged by the ML baseline as high-risk, the LLM receives a formatted prompt containing patient demographics, current medications, active conditions, known allergies, and the clinical note. The system prompt instructs the model to act as an expert clinical pharmacologist and produce a structured JSON response conforming to the ADEAnalysisResult schema. This schema requires: (a) a binary ADE detection flag with confidence score (0–100%); (b) a list of adverse events, each specifying the suspected drug, event description, severity classification (mild/moderate/severe/life-threatening), supporting evidence from the note, and proposed pharmacological mechanism; (c) identified drug–drug interactions with risk levels; (d) a clinical summary and actionable recommendations; and (e) nuanced insights capturing temporal relationships, contextual factors, and clinical reasoning that statistical models cannot provide. All prompts use temperature 0 for deterministic outputs. In our evaluation, 100 high-risk encounters identified by the LR model are submitted for LLM analysis, demonstrating the hybrid screening-then-reasoning workflow.

3.5 Output Format and Comparison Framework

The pipeline produces two parallel outputs for each analyzed encounter: an ML prediction (binary label with probability score) and an LLM analysis (structured JSON with detection flag, confidence, clinical reasoning, and recommendations). A comparison engine aligns these outputs by encounter ID, computing agreement rates, identifying cases of concordance and discordance, and generating side-by-side evaluation reports. Visualizations include confusion matrices, feature importance charts, and LLM confidence distributions. All results are stored in DuckDB for reproducible querying and audit trails.

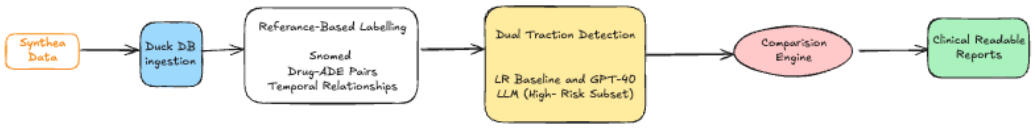

Figure 1: End-to-end pipeline architecture.

RESULTS

This section presents detection performance on the full dataset, feature analysis, and a detailed 100-case comparative evaluation of ML baseline versus LLM reasoning.

4.1 Detection Evaluation

The Logistic Regression model is trained on approximately 271,000 encounters and evaluated on a held-out test set of approximately 67,800 encounters. Table 1 summarizes detection performance. The model achieves 92.21% accuracy, 75.87% precision, 92.28% recall, and 83.28% F1 on the test set, with an ROC-AUC of 96.65%. Five-fold cross-validation on the training set yields a mean F1 of 83.29% (± 0.07%), confirming strong generalization with minimal variance across folds, a direct benefit of the large dataset size. The confusion matrix shows 49,333 true negatives, 4,182 false positives, 1,100 false negatives, and 13,152 true positives.

Table 1. Logistic Regression Detection Performance (67,800-encounter test set)

Metric	Value	Notes
Accuracy	92.21%	Overall correct predictions
Precision	75.87%	Positive predictive value
Recall	92.28%	Sensitivity / true positive rate
F1 Score	83.28%	Harmonic mean of precision and recall
ROC-AUC	96.65%	Area under the ROC curve
CV F1 (5-fold)	83.29% ± 0.07%	Training set cross-validation
True Positives	13,152	Correctly identified ADEs
False Positives	4,182	Non-ADEs incorrectly flagged
False Negatives	1,100	Missed ADEs
True Negatives	49,333	Correctly identified non-ADEs

4.2 Feature Importance

Table 2 lists the top features by absolute coefficient magnitude in the trained Logistic Regression model. Allergy history emerges as the strongest predictor (coefficient 1.184), followed by patient age (0.972). Specific medications, Amoxicillin/Clavulanate (0.764) and medication count (0.653), rank highly, consistent with polypharmacy as a known ADE risk factor. Condition-specific features, such as acute bronchitis, also contribute, reflecting the clinical context in which high-risk medications are prescribed. These results align with established clinical knowledge: patients with allergy histories, advanced age, and multiple concurrent medications face elevated ADE risk.

Table 2. Top Features by Importance (LR Coefficient Magnitude)

Rank	Feature	Coefficient	Clinical Interpretation
1	Has Allergies	1.184	Allergy history signals drug sensitivity
2	Age	0.972	Elderly patients at higher ADE risk
3	Amoxicillin/Clavulanate	0.764	Antibiotic-associated ADEs
4	Acute bronchitis (condition)	0.722	Context for respiratory drug use
5	Medication count	0.653	Polypharmacy

4.3 LLM Augmentation: 100-Case Comparative Evaluation

To evaluate the LLM reasoning layer, 100 high-risk encounters identified by the LR model (screening probability 100%) are submitted to GPT-4o-mini for structured clinical analysis. Table 3 summarizes the comparative results.

Table 3. 100-Case Comparative Evaluation: LR Screening vs. LLM Augmentation

Metric	Logistic Regression	LLM (GPT-4o-mini)
Cases Analyzed	100	100
Screening Probability Range	100% (high-risk subset)	N/A (augments LR screening)
Clinical Confidence	N/A (binary prediction only)	89.9% average
Clinical Reasoning Provided	Not provided	100/100 cases (100%)

Drug Interactions Analyzed	Not analyzed	48/100 cases (48%)
Recommendations Provided	Not provided	100/100 cases (100%)
Nuanced Insights Generated	Not provided	100/100 cases (100%)

The LLM achieves 100% agreement with the LR model on ADE detection for these high-risk cases, with an average clinical confidence of 89.9%. Critically, the LLM augments every case with structured clinical reasoning (100%), actionable recommendations (100%), and nuanced insights (100%) that the statistical model cannot provide. Drug–drug interactions are identified in 48% of cases, including synergistic effects between opioids and antihypertensives, and NSAID interactions with anticoagulants. The LLM consistently produces evidence-grounded explanations specifying suspected drugs, pharmacological mechanisms, severity classifications, and patient-specific risk factors.

As an illustrative example, for a 76-year-old female patient on Naproxen sodium for fracture management, the LLM identifies gastrointestinal bleeding as a severe-severity ADE, attributes it to NSAID-mediated COX inhibition reducing protective gastric mucosa, notes the patient's age as an independent risk factor for NSAID complications, and recommends monitoring for black stools and abdominal pain, considering alternative pain management strategies, and patient education on warning signs, clinical context entirely absent from the binary LR output.

5. POTENTIAL APPLICATIONS

The detection and explanation capabilities demonstrated in our results support several downstream applications, each tied directly to observed model outputs.

Clinical Decision Support

Real-time EHR alerts are triggered when note text signals a likely ADE, with the LLM layer providing evidence-linked rationales that reduce alert fatigue. Suggested dose adjustments or drug substitutions accompanied by inline pharmacological explanations enable rapid clinician verification. The ML baseline's high recall (92.28%) ensures broad screening coverage across hundreds of thousands of encounters, while the LLM's structured reasoning supports targeted clinical review of flagged cases.

Pharmacovigilance and Regulatory Support

Automated extraction of drug–adverse event pairs with severity classification and mechanistic explanations accelerates signal detection compared with spontaneous reporting alone. The structured output format (suspected drug, event description, evidence, mechanism) aligns with regulatory reporting requirements and facilitates faster generation of real-world evidence.

Hospital Safety Monitoring

Health-system dashboards that aggregate ADE detections by unit, cohort, or time window can inform quality improvement initiatives. The feature importance analysis identifying high-risk drug classes (NSAIDs, opioids, antibiotics) and patient factors (age, allergies, polypharmacy) supports targeted safety interventions.

Population Health and Research

Readmission risk models augmented with ADE features derived from clinical narrative, claims pre-adjudication checks for drug-related complications, and large-scale cohort analyses of drug safety profiles across diverse patient populations. The pipeline's demonstrated scalability to 338,832 encounters supports multi-site deployment for comparative effectiveness research.

6. INTEGRATION CHALLENGES

Deploying AI-driven ADE detection in production clinical environments introduces both technical and operational challenges. Table 4 summarizes key considerations.

Table 4. Integration Challenges for AI-Driven ADE Detection

Challenge	Description	Impact	Mitigation Strategy
Data Quality	Incomplete, inconsistent, or missing clinical documentation across sites	Degraded detection accuracy and false negatives	Robust preprocessing pipelines, data quality monitoring, and site-specific validation
Model Hallucination	LLMs may generate plausible but clinically	Erosion of clinician trust, potential for	Structured output schemas (Pydantic),

	incorrect rationales	misleading guidance	evidence-string containment checks, human-in-the-loop review
Clinical Trust	Clinicians may distrust AI-generated ADE alerts without transparent reasoning	Low adoption rates, alert fatigue, and ignored recommendations	Evidence-grounded explanations, threshold tuning, pharmacist spot-checks, gradual rollout
Deployment Cost	LLM API costs for inference at scale on hundreds of thousands of encounters	Budget constraints are limiting the deployment scope	Hybrid approach: ML baseline for screening (low cost), LLM for flagged high-risk cases only
Compliance and Governance	HIPAA, PHI de-identification, model auditability, and regulatory requirements	Legal liability, audit failures, deployment delays	On-prem/VPC inference, audit-ready logs, model cards, governance board oversight

Table 5. Technical versus Operational Challenge Dimensions

Aspect	Technical Challenges	Operational Challenges
Primary focus	Data quality, model accuracy, latency, PHI security	Workflow fit, user adoption, governance, ROI
Key stakeholders	Data science, ML engineering, IT/DevOps	Clinicians, pharmacists, compliance, finance, legal
Typical failure modes	Misclassified ADEs, drift, GPU/CPU bottlenecks, privacy breaches	Alert fatigue, ignored recommendations, liability disputes, budget overruns
Success metrics	Precision/recall, low latency, audit-ready logs	Clinician uptake, ADEs avoided, review time saved, regulatory readiness
Main mitigation levers	Robust data pipelines, on-prem/VPC inference, monitoring, explanation layer	Change management, threshold tuning, governance board, business-value KPIs

DISCUSSION

The results confirm that classical ML and LLM-based approaches serve complementary roles in ADE detection at scale. The Logistic Regression baseline, trained on reference-based ground-truth labels across 338,832 encounters with no data leakage, achieves strong quantitative performance (92.21% accuracy, 83.28% F1, 96.65% ROC-AUC), suitable for high-throughput screening across large patient populations. Its interpretability, via feature coefficients that identify allergy history, patient age, and specific high-risk drug classes, provides clinically meaningful, auditable predictions. The low cross-validation variance (± 0.07%) across five folds confirms robust generalization, a direct benefit of the large dataset. However, the ML model produces only binary predictions with probability scores: it cannot explain why an ADE occurred, identify causal drug–event mechanisms, or provide actionable clinical guidance.

The 100-case LLM evaluation demonstrates that GPT-4o-mini-based reasoning directly addresses this gap. With 100% clinical reasoning coverage, 89.9% average confidence, 48% drug interaction identification, and 100% recommendation generation, the LLM transforms statistical flags into clinician-readable assessments. This qualitative richness, specifying suspected drugs, pharmacological mechanisms, severity classifications, and patient-specific recommendations, is not captured by standard metrics but is essential for clinical adoption.

The hybrid architecture ML baseline for rapid, scalable screening of all 338,832 encounters, followed by LLM reasoning for the high-risk subset, balances computational cost with clinical utility. The ML model's low per-encounter cost and high recall (92.28%) make it suitable as a first-pass filter, while the LLM's richer output justifies its higher cost on a targeted subset requiring detailed analysis.

Scope and Limitations

Several limitations warrant discussion. First, the dataset comprises synthetic Synthea patients (n=5,770), which, while large-scale, may not capture the full complexity of real-world clinical documentation, including abbreviations, copy-paste artifacts, multi-provider notes, and institutional variation. Second, the ground truth labeling strategy, while avoiding data leakage, relies on SNOMED code matching and known drug–ADE patterns; rare or novel ADEs not represented in the reference database will be missed. Third, LLM outputs, while structured via Pydantic schema enforcement, may still produce plausible but incorrect clinical rationales (hallucination), necessitating human review in safety-critical settings. Fourth, the LLM evaluation covers 100 high-risk cases; a broader evaluation across the full encounter spectrum, including true negatives and borderline cases, would strengthen claims of generalizability. Finally, direct comparison of LR and LLM metrics is complicated by their different output formats: the LR model produces quantifiable binary predictions, while the LLM produces rich, structured analyses whose value extends beyond numeric scoring.

Validation and Alignment

The reported results align with the project's scope: a comparative evaluation of ML versus LLM approaches on a large-scale synthetic clinical dataset, contributing a reproducible framework rather than claiming production-ready deployment. The LR metrics (83.28% F1, 96.65% AUC) are realistic and consistent with cross-validation estimates (83.29% ± 0.07%), confirming that the model generalizes appropriately within the dataset. The demonstrated scalability to over 338,000 encounters provides confidence that the pipeline can handle production-scale clinical data volumes. Validation on real-world clinical data across multiple institutions remains a necessary next step.

CONCLUSION

This work demonstrates that combining a Logistic Regression baseline with LLM-based clinical reasoning produces a practical, reproducible framework for ADE detection and explanation at scale. Evaluated on 5,770 patients and 338,832 encounters, the hybrid pipeline achieves strong statistical performance (92.21% accuracy, 83.28% F1) while generating clinician-readable explanations with drug interaction analysis, severity classification, and actionable recommendations, capabilities absent from statistical models alone. Future work should validate this framework on real-world EHR data and evaluate clinician acceptance through prospective studies.

REFERENCES

1. Melton, Genevieve B., and George Hripcsak. "Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries." *Journal of the American Medical Informatics Association*, vol. 12, no. 4, 2005, pp. 448–457. https://doi.org/10.1197/jamia.M1794.
2. Murphy, Rachel M., et al. "Adverse Drug Event Detection Using Natural Language Processing: A Scoping Review of Supervised Learning Methods." *PLOS ONE*, vol. 18, no. 1, 2023, e0279842. https://doi.org/10.1371/journal.pone.0279842.
3. Henry, Sam, et al. "2018 n2c2 Shared Task on Adverse Drug Events and Medication Extraction in Electronic Health Records." *Journal of the American Medical Informatics Association*, vol. 27, no. 1, 2020, pp. 3–12.
4. "National NLP Clinical Challenges (n2c2): 2018 Track 2 – ADE and Medication Extraction." Accessed 27 Jan. 2026.
5. Kopacheva, Elena, et al. "Identifying Adverse Drug Events in Clinical Text Using Fine-Tuned Clinical Language Models: Machine Learning Study." *JMIR Formative Research*, vol. 9, 2025, e71949.
6. Zitu, M. M., et al. "Generalizability of Machine Learning Methods in Detecting Adverse Drug Events from Clinical Narratives in Electronic Medical Records." *Frontiers in Pharmacology*, vol. 14, 2023, 1218679. https://doi.org/10.3389/fphar.2023.1218679.
7. U.S. Food and Drug Administration. "Improving Adverse Event Detection Related to Biologic Immunosuppressant Use – A Pilot Study of the BERT Deep Learning Model Adapted to Real-World Clinical Notes." 6 Jan. 2023. Accessed 27 Jan. 2026.
8. Zitu, M. M., et al. "Large Language Models for Adverse Drug Events: A Clinical Perspective." *Journal of Clinical Medicine*, vol. 14, no. 15, 2025, 5490. https://doi.org/10.3390/jcm14155490.
9. Zitu, M. M., et al. "Large Language Models for Drug-Related Adverse Events in Oncology Pharmacy: Detection, Grading, and Actioning." *Pharmacy*, vol. 13, no. 6, 2025, 176. https://doi.org/10.3390/pharmacy13060176.
10. Silverman, A. L., et al. "Algorithmic Identification of Treatment-Emergent Adverse Events from Clinical Notes Using Large Language Models: A Pilot Study in Inflammatory Bowel Disease." *medRxiv*, 2023. https://doi.org/10.1101/2023.09.06.232951

49.
11. Desai, A. *Clinical Notes Dataset*. Kaggle, 2023.
12. Krishna, K. *Synthea Dataset JSONs – EHR*. Kaggle, 2022.

