Adverse Drug Event (ADE) Detection: Integrating LLMs with SIDER Knowledge

Soluchi Fidel-Ibeabuchi, Howard Prioleau, Saurav K. Aryal, PhD

AI4PC Lab, Institute for Human Centered Artificial Intelligence

INTRODUCTION

Detecting Adverse Drug Events (ADEs) from unstructured clinical text is a critical challenge in pharmacovigilance. While rule-based and classical NLP methods can identify known associations, they often struggle with the variability of clinical language and the identification of novel events. This project explores a hybrid approach that combines clinical reasoning from Large Language Models (LLMs) with structured knowledge from the SIDER (Side Effect Resource) database to improve extraction accuracy from clinical discharge notes.

TASK

The primary objective is to accurately identify (Drug, ADE) pairs within medical notes. The system must:

- Extract symptoms and abnormal patient experiences mentioned in text
- Distinguish between pre-existing conditions and events causally linked to specific medications
- Evaluate whether providing the LLM with structured knowledge (SIDER context) improves detection performance compared to zero-shot inference

DATA PIPELINE

The project utilizes a multi-stage pipeline to process heterogeneous datasets:

- **Data Sources:** The n2c2 (2018 shared task) dataset (500+ training files and 202 test files) for clinical notes and the SIDER database (157,854 pairs) for structured knowledge.
- **Preprocessing:** Parsing Brat-format standoff annotations into tabular data and implementing a centralized configuration for reproducibility.
- **Normalization:** Reducing clinical variability through lowercasing, punctuation removal, and British-to-American spelling conversion (e.g., "anaemia" to "anemia").
- **Knowledge Linking:** A cascaded matching strategy, including exact, alias expansion, and fuzzy matching (≥ 88 token-set ratio), to enrich clinical pairs with SIDER frequency data

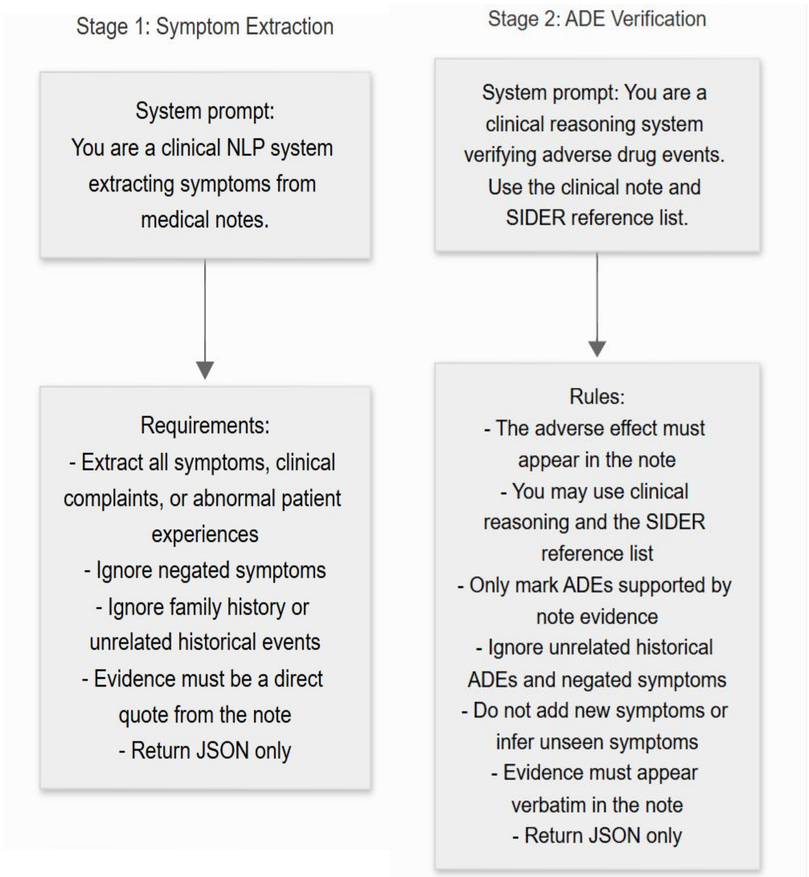
METHODS

The core detection logic employs a two-stage LLM inference workflow using the **meta-llama/llama-3.3-70b-instruct model** via the OpenRouter API.

Evaluation Framework

- **Overall:** All predicted pairs
- **SIDER-matched:** Pairs already present in the knowledge base
- **Novel:** Pairs appearing in clinical notes but absent from SIDER

Figure 1: Two-stage ADE pipeline prompt design: Stage 1 extracts symptoms from clinical notes, and Stage 2 verifies drug-caused ADEs using note evidence plus SIDER context.



RESULTS

Evaluation on the full 202-file n2c2 test set revealed several key insights:

- **Optimal Performance:** The best operating region was found at a 0.75–0.80 confidence threshold using SIDER context, achieving an overall F1 score of 0.5140.
- **SIDER Impact:** Without thresholding, the "no-SIDER" run performed better. However, SIDER context became superior at confidence levels ≥ 0.55, indicating that structured knowledge helps prune lower-confidence false positives.
- **Subgroup Gains:** The refined prompt refactor (2pred family) showed the strongest improvement in the "Novel" subgroup (pairs not in SIDER), with an F1 gain of +0.0219.
- **Metric plateaus:** Due to quantized confidence values (e.g., 0.7, 0.8), metrics remained identical across certain threshold bands

Run (Best-Performing)	Group	Definition	Precision	Recall	F1-score
With SIDER at threshold 0.75-0.80	Overall	All predicted pairs	0.4700	0.5670	0.5140
With SIDER at threshold 0.75-0.80	SIDER-matched	Pairs already present in the knowledge base	0.5217	0.8000	0.6316
With SIDER at threshold 0.75-0.80	Novel	Pairs appearing in clinical notes but absent from SIDER	0.4589	0.5295	0.4917

Table 1: Performance summary for the best-performing configuration, reported by evaluation stratum: Overall, SIDER-matched, and Novel, using Precision, Recall, and F1-score.

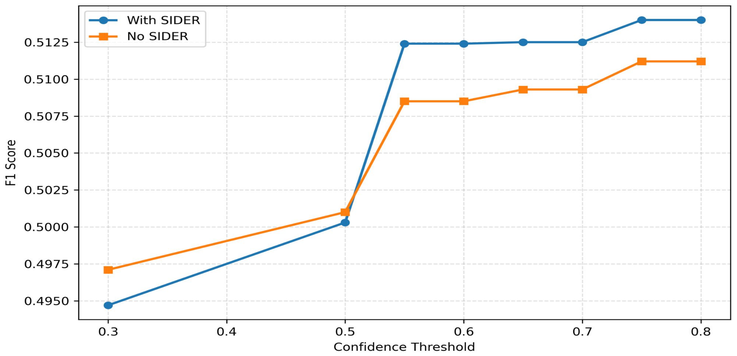


Figure 2: F1 score vs. confidence threshold for With SIDER and No SIDER runs, showing a small but consistent performance advantage for SIDER at higher thresholds.

CONCLUSION

Integrating SIDER context is a conditional benefit; it enhances LLM performance primarily when low-confidence predictions are filtered out. While the prompt refactor significantly improved both recall and precision, the system still faces challenges with "over-anchoring" on known SIDER effects. Future iterations will focus on temporal reasoning (ordering drug-start vs. symptom-onset) and explicit causality filtering in Stage 2.

REFERENCES

1. Meta AI, "Llama 3.3 70B Instruct Model Card," Hugging Face, Dec. 6, 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Accessed: Feb. 17, 2026
2. M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," Nucleic Acids Research, vol. 44, no. D1, pp. D1075–D1079, Jan. 2016, doi: 10.1093/nar/gkv1075.
3. S. Henry, K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner, "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, pp. 3–12, Jan. 2020, doi: 10.1093/jamia/ocz166.

ACKNOWLEDGEMENTS

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714 and an Amazon Research Award. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.