TruPharma: A GenAI-Powered Verification Engine for Real-World Drug Safety

Nithin Songala[1] Salman Mirza[2] Amy Ngo[3]
[1]University of Missouri - Kansas City

Abstract. TruPharma is a framework designed to quantify the disparity between structured regulatory labels (NDC) and unstructured post-market adverse event data (FAERS). By implementing a Hybrid Retrieval-Augmented Generation (RAG) architecture powered by Snowflake Cortex AI, the platform utilizes agentic orchestration and vector search to harmonize official clinical "ground truth" with live patient narratives in real-time. Functioning as a "Check Engine" light for personal health, TruPharma distills millions of noisy reports into clear, data-backed assessments that empower patients to distinguish between established side effects and emerging safety risks, while simultaneously providing a scalable, automated pipeline for proactive institutional signal detection.

Key words: Pharmacovigilance, Hybrid RAG, Snowflake Cortex, Vector Search, agentic orchestration, Governance, OpenFDA

1. Introduction

There is a critical latency gap between the "official" safety profiles of drugs (derived from limited clinical trials) and "real-world" safety (what actually happens to millions of patients post-market). The FDA FAERS dataset contains millions of patient narratives describing adverse events, but this data is unstructured, noisy, and disconnected from the official National Drug Code (NDC) labeling. Currently, no unified tool exists that allows a user to cross-reference their personal symptoms against both the official label and live population data instantly.

We propose TruPharma, a consumer-grade web interface hosted directly within Snowflake. Users can input a drug name and a symptom. The GenAI responds with a verified assessment citing both the official label and anonymous real-world case counts. For analysts, a visual dashboard will show drugs with the highest disparity between Official label and User Reports over the last 90 days. In this proposal, we will discuss the real world relevance, impact, system architecture, methodology, and evaluation plan.

2. Team Members & GitHub Repo

Team Name:

Team Members:

- Nithin Songala - `https://github.com/reddy-nithin`
- Salman Mirza - `https://github.com/SalmanM1`
- Amy Ngo - `https://github.com/ango3636`

Project can be found at the GitHub Repo: `https://github.com/ango3636/TruPharma`

3. Related Work

3.1. RAG-based Architectures for Drug-Side Effect Retrieval in LLMs

The paper finds that GraphRAG (using Neo4j) achieves near-perfect accuracy in side-effect retrieval compared to standard vector search. For TruPharma, this supports the use of RxNorm and DrugBank to create a structured relationship between drugs and events rather than relying on simple keyword matching. It explicitly addresses the problem of using LLMs to identify drug side effects and compares standard Vector RAG against GraphRAG.

GitHub: `https://github.com/diegogalpy/RAG-based-models-for-drug-side-effect-retrieval`

3.2. HM-RAG: Hierarchical Multi-Agent Multimodal Retrieval Augmented Generation

This paper informs the agentic orchestration layer. The project deals with heterogeneous data ecosystems (structured NDC labels vs. unstructured FAERS narratives), a single-agent RAG may struggle with the reasoning required to map these two different worlds. HM-RAG uses a Decomposition Agent to break complex queries into sub-tasks (e.g., "Find the NDC code" then "Search for adverse events"). This is what the Snowflake Cortex agents need to do when a user asks a colloquial question like "Why is this blue pill making me dizzy?"

GitHub: `https://github.com/ocean-luna/HMRAG`

3.3. MegaRAG: MultiModal Knowledge Graph Based RAG

This paper informs the Data Ingestion and Knowledge Layer of the project. TruPharma is designed as a hybrid framework that must reconcile diverse data modalities, specifically structured regulatory metadata (NDC) and unstructured patient narratives (FAERS). MegaRAG provides a specialized blueprint for building a Multimodal Knowledge Graph that allows an LLM to understand medical concepts by simultaneously accessing their structured relationships and their natural language descriptions. This methodology directly supports our colloquial-to-medical mapping, ensuring that user symptoms are grounded in a unified medical knowledge space rather than simple vector similarity.

GitHub: `https://github.com/pfnet-research/megarag`

3.4. Tree-of-Reasoning (ToR): Towards Complex Medical Diagnosis via Multi-Agent Reasoning with Evidence Tree

This informs the Evaluation Framework, specifically the answer groundedness and citation precision metrics. ToR introduces an "Evidence Tree" that records the reasoning path of an LLM alongside clinical evidence. For TruPharma, this provides a model for how to display the disparity analysis to a user—showing them the exact path from their symptom to the FAERS data and finally to the FDA label to prove it is a verified risk.

GitHub: `https://github.com/tsukiiiiiiiii/TOR`

4. Data Sources

The TruPharma framework relies on a curated pipeline of pharmaceutical and clinical datasets to bridge the information gap between official regulatory guidelines and post-market patient experiences. By integrating structured drug labeling with unstructured adverse event narratives, the system creates a multi-modal knowledge base capable of performing complex

colloquial-to-medical mapping and disparity analysis. The following sources provide the necessary scale and clinical depth to support the system's hybrid retrieval-augmented generation (RAG) architecture and signal detection workflows.

4.1. OpenFDA: Adverse Event & Product Labeling APIs

Direct Datalink: `https://open.fda.gov/apis/drug/`

Description

- Size: Millions of records; the Adverse Event endpoint alone contains over 15 million reports.
- Modality: Semi-structured JSON.
- Content: Includes Adverse Event Reports (patient demographics, drug details, reactions) and Drug Labeling (indications, warnings, and precautions).

Relevance

- Retrieval & Grounding: Used to populate the Snowflake Cortex vector search, providing the "ground truth" (labels) and "real-world evidence" (Adverse Events).
- Benchmarking: Acts as the primary source for identifying emerging safety signals.

Limitation: Data is noisy and unstructured; patient-reported narratives often lack clinical verification and contain colloquial language that requires normalization.

4.2. RxNorm

Direct Datalink: `https://www.nlm.nih.gov/research/umls/rxnorm/`

Description

- Size: 100k+ concepts with millions of relationships.
- Modality: Structured text/graph.
- Content: A normalized naming system for generic and branded drugs.

Relevance

- Fine-tuning/Orchestration: Critical for the agentic orchestration layer to map colloquial drug names mentioned in patient narratives to official National Drug Codes (NDC).

Limitation: Generally excludes radiopharmaceuticals, bulk powders, and dietary supplements.

4.3. DrugBank (v5.1.14)

Direct Datalink: `https://go.drugbank.com/`

Description

- Size: Comprehensive database covering 14,000+ drug entries.
- Modality: Structured documents (XML/JSON) and text.
- Content: Detailed drug chemistry, pharmacology, and drug-drug interactions (DDI).

Relevance

- Grounding: Provides deep clinical context for the RAG system to explain why an adverse event might be occurring based on biochemical pathways.

Limitation: Full access requires a commercial license; public versions may have slight latency in updates.

4.4. BLUE (Biomedical Language Understanding Evaluation) Benchmark

Direct Datalink: `https://github.com/ncbi-nlp/BLUE_Benchmark`

Description

- Size: 10 datasets across 5 biomedical tasks.
- Modality: Specialized biomedical and clinical text.

Relevance

- Benchmarking: Specifically used for evaluating the performance of NLP models on Named Entity Recognition (NER) and Relation Extraction within the clinical domain.

Limitation: Evaluation is focused on text similarity and inference, which may not fully capture the reasoning required for complex pharmacovigilance signals.

5. Problem & Impact

5.1. Real World Relevance

Clinical trials are conducted under pristine conditions. They are controlled environments with limited, homogenous patient groups. Once a drug hits the mass market, it interacts with diverse genetics, pre-existing conditions, and polypharmacy (drug-to-drug interactions) that trials cannot predict. Currently, the FDA Adverse Event Reporting System (FAERS) is a data graveyard for the average consumer. It is too complex for non-experts to navigate, leaving patients to rely on outdated paper inserts or unreliable internet forums to understand their side effects.

5.2. Stakeholders

- Primary Users (Patients): Individuals experiencing unexpected symptoms who need to know if their experience is a known side effect or a signal in the broader population.
- Healthcare Providers: Doctors and pharmacists who require a quick, data-backed way to validate patient concerns against real-world trends without manual database querying.
- Life Science Analysts: Drug safety (Pharmacovigilance) teams who need to monitor labeling disparities—where real-world incidents significantly outpace official clinical warnings.

5.3. System Workflow

TruPharma transforms the current fragmented research process—where users must choose between dense regulatory PDFs or unverified internet forums—into a unified, automated intelligence loop. The GenAI system improves two primary workflows:

1. The Consumer "Check Engine" Workflow: Currently, a patient feeling a side effect must manually search the FDA's Structured Product Label (SPL) or rely on anecdotal "Dr. Google" results. TruPharma automates this via the Safety Chat:
   (a) Natural Language Entry: Users input symptoms in plain English.
   (b) Intelligent Reconciliation: Utilizing Cortex Vector Search, the system maps natural language to standardized medical terms.
   (c) Dual-Source Validation: The AI performs a "Joins-over-Data" query to compare official clinical trial incidence rates against live FAERS population counts, delivering a verified safety assessment in seconds.
2. The Pharmacovigilance Signal Detection Workflow: Safety analysts currently spend weeks manually aggregating data to find signals (emerging trends in side effects). TruPharma automates this via the Signal Heatmap:
   (a) Automated Pipeline: A Snowpark Container Services pipeline fetches daily JSON updates from the OpenFDA API, ensuring data never becomes stale.
   (b) Automated Disparity Analysis: The system calculates the Latency Gap between official labels and real-world reports.
   (c) Proactive Alerting: Analysts use the dashboard to instantly identify which drugs have the highest disparity over the last 90 days, allowing for rapid regulatory intervention.

5.4. Real World Impact

A low-latency tool that could redefine public health safety. If developed, TruPharma would reduce the latency gap of drug safety updates from years to days, potentially saving lives by identifying dangerous drug interactions or side effects long before they trigger a formal FDA recall.

6. GenAI System Architecture & Pipeline

6.1. Data Ingestion & Knowledge Layer

- Document Ingestion & Pipeline: An automated Python-based pipeline running on Snowpark Container Services (SPCS) fetches daily JSON updates from the OpenFDA API.
- Chunking & Indexing: Unstructured patient narratives from FAERS are chunked and converted into embeddings. Structured NDC labeling data is ingested into relational tables, preserving the hierarchical relationship between Brand Names, Active Ingredients, and official contraindications.
- Storage Layer: Snowflake serves as the single source of truth, utilizing Cortex Search for vector storage and standard relational tables for metadata.
- Versioning & Provenance: The system maintains a data timestamp for every record fetched from OpenFDA, ensuring the Signal Heatmap reflects only the most recent 90-day window for accurate disparity analysis.

6.2. Retrieval, Generation & Fine-Tuning

- Embeddings and Vector Search: The system uses Cortex Search to perform semantic queries on patient narratives. This enables colloquial-to-medical mapping, allowing the system to recognize that a user's "brain fog" matches a clinician's report of "cognitive dysfunction."
- Hybrid RAG Design: A dual-track retriever pulls official ground truth via SQL (structured) and real-world signals via Vector Search (unstructured).
  - Grounding & Citation: The generation phase is grounded in these two specific pools of data. Every AI response is required to cite the NDC label and the specific FAERS Case Count.
- agentic orchestration: Instead of a single prompt, the system uses a Chain-of-Thought reasoning chain powered by Cortex.Complete:
  - Agent A (Regulator): Extracts official side effects.
  - Agent B (Observer): Identifies statistical clusters in real-world data.
  - Agent C (Verifier): Reconciles the two to identify emerging signals (high B, low A).
- Domain Adaptation: Rather than heavy fine-tuning, the system utilizes Instruction Tuning within the agentic prompt to ensure the LLM adheres to medical reporting standards and avoids providing medical advice.

6.3. Interface & Delivery

- Dashboards & Chat Interfaces: The primary delivery mechanism is a Streamlit in Snowflake web application featuring:
  - The Safety Chat: A natural language interface for consumer-facing drug/symptom verification. Delivers a safety assessment based on genetic information, drug, dosage, frequency, and symptom(s).
  - The Signal Heatmap: A high-level visualization for analysts to track labeling disparities.

7. Methods, Technologies & Tools

7.1. Vector Databases & Search Systems

- Snowflake Cortex Search: Cortex Search is used for semantic retrieval of the patient narrative fields in FAERS. This allows the system to bridge the gap between colloquialisms (e.g., "tummy ache") and clinical terminology (e.g., "abdominal pain") via vector embeddings, while simultaneously applying SQL filters for NDC-specific drug labeling.

7.2. Domain Adaptation

- Entity Resolution Mapping: Rather than traditional fine-tuning, the project focuses on a lightweight mapping layer for Entity Resolution. This research-driven approach ensures that searches for brand names (e.g., "Advil") are accurately resolved to active ingredients (e.g., "Ibuprofen") across disparate data sources, ensuring comprehensive coverage without the high compute cost of model retraining.

7.3. Deployment & Visualization

- Streamlit, Tableau, or Power BI: The "Safety Chat" will be deployed via Streamlit in Snowflake for a seamless consumer experience. For the "Signal Heatmap" intended for analysts, the team will leverage Tableau or Power BI to visualize complex data disparities and historical trends.

7.4. Collaboration & Reproducibility tools

- Snowpark: The team will use Snowpark Container Services to build a Python-based automated ingestion pipeline for daily OpenFDA updates.
- GitHub & Python Notebooks: Collaboration is facilitated through shared Python Notebooks on the GitHub Repo.

8. Evaluation & Impact Plan

To ensure the TruPharma engine provides clinically reliable insights rather than just fluent text, we will implement a multi-layered comparative evaluation framework. This framework focuses on the system's ability to resolve the disparity between patient colloquialisms and official medical documentation without introducing hallucinations.

- Grounding & Hallucination Analysis (Primary): We will utilize a citation precision audit to verify that every adverse event claim generated by the system is mapped to a specific record in the FAERS or openFDA labeling datasets. This follows the "Evidence Tree" methodology, where the system must visualize the reasoning path from a user's symptom to a specific regulatory document.
- A/B Testing (Comparative): We will conduct a comparative analysis between a "Vanilla" LLM (GPT-4o/Llama-3 without RAG) and our Snowflake-native Hybrid RAG system. This will specifically measure the difference in accuracy for drug-side effect retrieval, a metric proven to be significantly higher in structured RAG architectures.

Metrics

- answer groundedness / citation precision: Percentage of generated responses where the cited adverse event matches the official FDA label or FAERS report frequency.
- Hallucination Rate: Frequency of false positive side effects—claims made by the LLM that do not exist in the retrieved medical context.
- Semantic Mapping Accuracy: Using the BLUE benchmark tasks, we will measure the system's success rate in mapping colloquial patient terms (e.g., "shaky hands") to formal medical entities (e.g., "Tremors").
- Latency and Cost per Query: Monitoring the efficiency of the agentic orchestration layer to ensure real-time performance within the Snowflake environment.

To quantify the "RAG-Lift," we will compare TruPharma against two primary baselines:

1. Baseline 1 (Standard LLM): A zero-shot LLM (Snowflake Cortex llama3-70b) without access to our specialized vector store or RxNorm mapping. This measures the baseline hallucination risk of general-purpose models on pharmaceutical queries.

2. Baseline 2 (Vector-Only RAG): A standard vector-search implementation without the agentic orchestration or Knowledge Graph layers. This will demonstrate the necessity of our hybrid approach in handling complex drug-side effect relationships, which typically causes standard vector systems to fail.

9. Conclusion

TruPharma addresses a fundamental failure in the current pharmaceutical lifecycle: the information asymmetry between regulatory "ground truth" and the lived reality of the patient population. By leveraging a Snowflake-native Hybrid RAG architecture, this project moves beyond simple data retrieval, introducing an intelligent reconciliation layer that maps colloquial patient experiences to structured clinical warnings in real-time.

The implementation of the "Check Engine" interface via Streamlit and Cortex AI provides a dual-purpose solution. For the consumer, it offers a verifiable, data-backed assessment that demystifies side-effect profiles. For the analyst, it provides a high-velocity Snowpark pipeline that identifies emerging safety signals months before traditional manual reviews might trigger a label change. TruPharma demonstrates that when GenAI is grounded in authoritative datasets and deployed within a secure data cloud, it can transform passive public health records into a proactive tool for drug safety and patient advocacy.

References

[1] Nygren, S., Avci, P., Daniels, A., Rassol, R., Beheshti, A., & Galeano, D. (2025). *RAG-based architectures for drug side effect retrieval in LLMs*. arXiv preprint arXiv:2507.13822. https://github.com/diegogalpy/RAG-based-models-for-drug-side-effect-retrieval

[2] Liu, P., Liu, X., Yao, R., Liu, J., Meng, S., Wang, D., & Ma, J. (2025). *HM-RAG: Hierarchical multi-agent multimodal retrieval augmented generation*. arXiv preprint arXiv:2504.12330. https://github.com/ocean-luna/HMRAG

[3] Hsiao, C.-H., Wang, Y.-C., Lin, T.-S., Yeh, Y.-R., & Chen, C.-S. (2025). *MegaRAG: Multimodal knowledge graph-based retrieval augmented generation*. arXiv preprint arXiv:2512.20626. https://github.com/pfnet-research/megarag

[4] Peng, Q., Cui, J., Xie, J., Cai, Y., & Li, Q. (2025). *Tree-of-Reasoning: Towards complex medical diagnosis via multi-agent reasoning with evidence tree*. arXiv preprint arXiv:2508.03038. https://github.com/tsukiiiiiiiii/TOR

[5] Peng, Y., Yan, S., & Lu, Z. (2019). *Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets*. arXiv preprint arXiv:1906.05474. https://github.com/ncbi-nlp/BLUE_Benchmark

[6] U.S. Food and Drug Administration. (2026). *openFDA drug product labeling and adverse event APIs*. https://open.fda.gov/apis/drug/

[7] National Library of Medicine. (2024). *RxNorm: Normalized names for clinical drugs*. Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/rxnorm/index.html

[8] DrugBank Online. (2024). *DrugBank release version 5.1.14*. https://go.drugbank.com/releases/latest