

GUARDRAIL-FIRST DRUG INFORMATION ASSISTANT

Presented by:

Aayushi Shah
Amaan Khan
Dhruva Joshi
Rohit Dhole
Sneha Guravannavar

Project Advisors:

Dr. Simon Shim

01. INTRODUCTION

THE PROBLEM

LLMs generate authoritative health text even when poorly grounded. Users treat hallucinated output as clinically valid. Existing chatbots optimize for answer rate, not safety.

OUR REFRAMING

We **BLOCK** when evidence is insufficient. Every response is binary: **ALLOW** or **BLOCK**. Fluent output that cannot be justified against curated evidence is refused.

✓ Supported: indications, mechanisms, adverse effects, interactions, toxicity
X Blocked: dosing, prescribing, diagnosis, non-drug topics

02. METHODOLOGY

1 Layer 1 Input Guardrail

Every query screened for scope & policy. Dosing, prescribing, diagnosis → blocked before any LLM call, saving cost and preventing risk.

2 BM25 Field-Aware Retrieval

In-scope queries retrieve from 327,410 DrugBank chunks indexed by drug name and semantic type. Transparent, auditable lexical retrieval.

3 Phi-3 Structured Generation

Evidence injected into structured prompt. Phi-3 writes to fixed schema (indication, mechanism, safety). Candidate answer — not yet shown to user.

4 Layer 2 Output Validation

Candidate tested against schema validity, grounding score, and coverage score. Any failure triggers a self-correcting retry before **BLOCK**.

03. SYSTEM ARCHITECTURE & PIPELINE

Dual-Guardrail Architecture: Guardrail on Input → Retrieval → LLM Generation → Guardrail on Output



Dual Guardrails: Queries are checked before and after LLM generation — blocked at either stage if unsafe, off-policy, or hallucinated.

RAG Pipeline: Approved queries retrieve relevant chunks from a DrugBank Evidence Store via BM25, which ground the LLM's response.

Full Observability: All stages are logged and monitored for decisions, retrieval quality, and evaluation metrics.

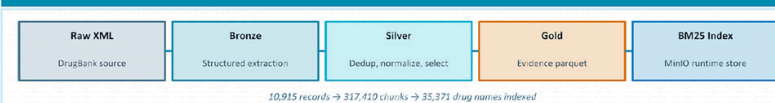
04. RESULTS & MODEL COMPARISON

★ Phi-3 = Deployed Default 500-prompt frozen benchmark (350 drug-info, 100 policy, 50 block). All LLMs tested under identical guardrail stack.

Model	ALLOW Rate	BLOCK Correct	Drug-Info ALLOW	Latency (s)	Grounding
★ phi3 (default)	37.4%	100% (150/150)	53.7%	6.3	0.437
mistral	34.8%	100% (150/150)	45.7%	11.0	0.424
llama3.1	33.5%	100% (150/150)	48.3%	19.0	0.418
gemma2	21.4%	100% (150/150)	30.6%	10.0	0.352

Key Finding: All 4 LLMs scored 100% block correctness (150/150) — guardrails enforce safety independent of the generator model. Higher ALLOW = fewer false positives on valid drug queries.

05. DATA ENGINEERING PIPELINE



10,915 records → 317,410 chunks → 35,371 drug names indexed

DATA 298B - Applied Data Science - San Jose State University - May 2026 | DrugBank data used under academic license

06. CONCLUSION



Safety Goal Achieved

All four LLMs blocked 100% of policy-sensitive prompts (150/150). Guardrail safety is independent of the generator model used.



Phi-3 Promoted to Default

Highest drug-info ALLOW (53.7%), lowest latency (6.3s), best grounding, perfect block correctness. Selected as deployed chatbot.



Fail-Closed is Correct

In high-stakes domains, refusing safely beats answering confidently. Refusal is a first-class outcome, not degraded UX.



Broader Impact

Pattern applies to compliance assistants, pharmacovigilance, regulated domain help desks — any AI where unsupported claims cause harm.

07. KEY REFERENCES

- [1] Ji et al. (2023). Survey of hallucination in NLG. ACM Computing Surveys, 55(12).
- [2] Lewis et al. (2020). RAG for knowledge-intensive NLP. NeurIPS.
- [3] Innan et al. (2023). Llama Guard: LLM-based I/O safeguard. arXiv:2312.06674.
- [4] Robertson & Zaragoza (2009). BM25 and beyond. Foundations & Trends in IR, 3(4).
- [5] Han et al. (2024). MedSafetyBench: Medical safety of LLMs. arXiv:2403.03744.
- [6] Wishart et al. (2018). DrugBank 5.0. Nucleic Acids Research.
- [7] Sava et al. (2023). NekMo Guardrails. arXiv:2310.10501.

Interested in partnering with us on a capstone project or exploring other ways to collaborate with the College of IDS? Contact CIDS_CAP@sjsu.edu to discuss potential projects and partnership opportunities.