

SPARTANGUARD: ENTERPRISE LLM GUARDRAILS HALLUCINATION DETECTION AND PREVENTION

Presented by:

Shashidhar PVD Babu
Vimalanandhan Sivanandham
Nakshatra Desai
Vineeth Rayadurgam
Harshita Sayala

Project Advisors:

Dr. Simon Shim
Dr. Lee Chang

Introduction

Enterprise AI can sound confident while making incorrect compliance claims. In regulated domains, these errors create legal, privacy, and operational risk.

SpartanGuard adds two safety layers: an input gateway that blocks unsafe or privacy-sensitive queries, and an output verifier that checks each generated claim against trusted evidence using multi-agent debate.

The system routes responses as DELIVER, RETRY, HUMAN_REVIEW, or HARD_BLOCK.



Problem Statement

Existing approaches do not fully solve compliance hallucination:

- Input-only filtering protects what enters the system, not what exits it.
- RAG generation improves grounding but does not prove each generated claim is supported.
- Single-model judging can agree with plausible but wrong claims.
- Human review at scale is too expensive and slow.

SpartanGuard targets four dangerous error types:

- Fully correct
- Missing caveat
- Hallucinated specific regulation / number / standard
- Jurisdiction-blind overgeneralization

Methodology

1. Gateway Input Threat Classification:

The gateway runs three classifiers in parallel before the query reaches the LLM:

Classifier	Purpose
PII Detector	Finds personal and sensitive information
Jailbreak Detector	Detects attempts to bypass safety rules
Prompt Injection Detector	Detects system override or instruction-smuggling attacks

2. Multi-Agent Debate for Output Verification:

The model answer is decomposed into atomic, verifiable claims. For each claim:

- Agent A receives top supporting evidence
- Agent B receives overlapping but more edge-case evidence

Analysis and Results

- Judge sees the full evidence pool and both debate rounds
- This asymmetric design encourages productive disagreement instead of false consensus.

3. Confidence Scoring Engine (CSE)

Signals are combined into a final routing decision. Default signals include:

- Faithfulness
- Hallucination risk inverse
- Contextual relevancy
- Judge evaluation score
- Context quality

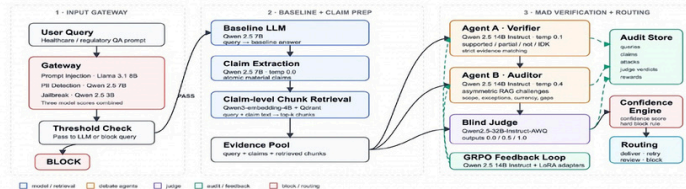
4. GRPO Feedback Loop

Agent confidence is calibrated using a Brier-style reward:

This penalizes confident wrong answers and rewards well-calibrated uncertainty, directly targeting the "authoritative but wrong" failure mode.

Models and Components

SpartanGuard End-to-End Architecture



Gateway / Retrieval Results

Metric	Result
PII Recall	98.2%
Jailbreak F1	0.9821
Prompt Injection F1	0.9821
Missed Injection Rate	1.2%
Retrieval Recall@1	94.9%
Retrieval nDCG@10	0.891

Layer	Model / Method
PII Detection	Qwen-2.5-7B-Instruct
Jailbreak Detection	Qwen2.5-3B-Instruct
Prompt Injection	Llama3.1-8B
Base LLM	Qwen2.5-3B
Decomposer	Qwen2.5-7B
Embeddings	Qwen3-embedding-4B
Vector DB	Qdrant
Agent Orchestration/Model	LangGraph/Qwen2.5-14B-Instruct and Qwen2.5-32B-Instruct-AWQ
Feedback Training	TRL GRPO/Unsloth

Agent / Model	Valid JSON	Mean Confidence	Brier MSE	Unsupported Verdicts - Overconfident	Partial Verdicts - Overconfident	Verdict Exact Match
Agent A Base 14B	65.6%	0.857	0.3667	87.5%	100.0%	66.7%
Agent A GRPO LoRA	100.0%	0.375	0.0516	0.0%	0.0%	78.1%
Agent B Base 14B	65.6%	0.629	0.3257	75.0%	66.7%	57.1%
Agent B GRPO LoRA	100.0%	0.412	0.0616	0.0%	0.0%	75.0%

Setup	Debate Agent Serving	Valid JSON	Mean Conf.	Unsupported Verdicts - Overconfident	Partial Verdicts - Overconfident	Exact Match vs Judge	Verdict MAE	Mean R1 Latency
Base MAD	One vLLM server, base Qwen2.5-14B	95.3%	0.857	92.7%	89.6%	85.3%	0.074	4.755 ms
MAD + GRPO LoRA	One vLLM server, base Qwen2.5-14B + two LoRA adapters in vLLM	100.0%	0.382	0.0%	0.0%	94.7%	0.026	2.415 ms

Summary/Conclusions

SpartanGuard shows that safer enterprise LLM deployment requires more than prompt filtering or basic RAG. In regulated domains, the critical problem is confident compliance hallucination. By combining gateway screening, claim decomposition, asymmetric multi-agent debate, judge-heavy routing, and GRPO-based calibration, SpartanGuard provides a practical framework for making LLM outputs safer, auditable, and operationally usable in healthcare and other enterprise environments. The strongest result was 100% routing accuracy on the healthcare ablation benchmark under the judge-heavy configuration.

Key References

- [1] CFMAD, COLING 2025 - Verifier, Skeptic, Judge architecture.
- [2] DPD, ICIC 2025 - Debate transition and challenge taxonomy.
- [3] DRAG - Asymmetric RAG evidence distribution.
- [4] MAD-Sherlock, ICWSM 2025 - RAG-grounded multi-agent verification.
- [5] Behaviorally Calibrated RL - Brier reward for confidence calibration.
- [6] DeepSeek-R1, 2025 - GRPO fine-tuning framework.

Acknowledgements

We thank our project advisor, the Applied Data Science Department at San José State University, and the open-source ecosystem including Hugging Face, Qdrant, Ollama, LangGraph, TRL, and DeepEval for enabling this work.

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.