System Card: S21Mind

A Hybrid Hallucination-Reduction Adapter for Open-Source LLMs

Date: February 2025 | **Version:** 19.0 (Enterprise) | **Base Model:** Llama-3-70B-Instruct

---

1. Executive Summary

S21Mind is a middleware "safety adapter" designed to enable the deployment of open-source models (specifically Llama 3) in high-stakes enterprise environments.

While Llama 3 70B achieves state-of-the-art performance on reasoning tasks, it suffers from a high rate of hallucination on factual recall and adversarial questions (TruthfulQA). HexaMind resolves this by implementing a **"Split-Brain" architecture** that filters outputs through deterministic logic, localized RAG, and archetypal chain-of-thought verification.

Key Result: On the TruthfulQA benchmark, S21Mind improves the HHEM (Hallucination Evaluation) Consistency Score from **0.51 (Baseline)** to **0.96 (HexaMind)**, effectively matching GPT-4 reliability with open-source weights.

---

2. System Architecture

S21Mind operates as a post-processing guardrail. It does not require retraining the base model. It uses a 3-layer filtering process to maximize accuracy while minimizing inference costs.

Layer 0: Deterministic Reflex (The Efficiency Layer)

- **Mechanism:** High-precision regex pattern matching.
- **Function:** Instantly validates answers containing signals of high-confidence truth (e.g., admissions of ignorance) or rejects known hallucination patterns (e.g., urban legends, superstitions).
- **Impact:** Handles ~15-20% of traffic with **0ms latency** and **$0 compute cost**.

Layer 1: Structural & Topological Analysis (The "s-sauce" Layer)

- **Mechanism:** Analyzes the semantic structure of the generated answer using a 6-bit topological signature.
- **Function:** Detects "structural hallucinations"—answers that are grammatically correct but logically "stagnant" or "entropic" (e.g., circular reasoning, stuttering, or excessive hedging).
- **Impact:** Rejects subtle nonsense that bypasses standard keyword filters.

Layer 2: Localized RAG & Semantic Adjudication (The "Wisdom" Layer)

- **Mechanism:** Vector-based retrieval against a curated "Gold Standard" Knowledge Base, followed by an LLM-based Chain-of-Thought Judge.
- **Function:**

1. **Fact Check:** Verifies claims against a local database of frequent misconceptions (Quote attribution, laws, statistics).
2. **Archetypal Judging:** If the fact is not found locally, a secondary LLM call evaluates the claim using "Skeptical Prompting" to identify plausible-sounding lies.

- **Impact:** Provides the final 10% accuracy boost required for enterprise compliance.

---

3. Performance Evaluation

We evaluated s21Mind using the **TruthfulQA (Generation)** benchmark, the industry standard for measuring model honesty. We used the **Vectara HHEM (Hughes Hallucination Evaluation Model)** standard via a DeBERTa-v3-NLI proxy to score consistency.

Results Summary

Metric	Baseline (Llama 3 70B)	HexaMind (System)	Improvement
HHEM Consistency Score	0.5139	**0.9630**	**+87.4%**
Accuracy (Human Eval Proxy)	~58.0%	**~90.0%**	**+32.0 pts**
Hallucination Rate	~42.0%	**< 5.0%**	**Significant Reduction**

Cost Efficiency

By utilizing Layer 0 (Reflex) and Layer 1 (Structure), HexaMind resolves approximately **21.5%** of queries without triggering a secondary LLM call. This results in a direct **~20% reduction in inference costs** compared to standard "LLM-as-a-Judge" verification methods.

---

4. Privacy & Data Residency

HexaMind is designed for banking, healthcare, and defense sectors where data privacy is paramount.

- **Local Filtering:** Layers 0 and 1 run entirely on-premise (CPU). No data leaves the firewall for these checks.

- **Local RAG:** The Knowledge Base is stored locally, preventing sensitive queries from needing external verification.
- **Modular Judge:** The Layer 3 Judge is model-agnostic and can be routed to private Azure/AWS endpoints or a local Llama instance.

---

5. Limitations

- **Domain Specificity:** The current RAG Knowledge Base is optimized for general knowledge and common misconceptions. Specialized domains (e.g., advanced biochemistry) may require knowledge base expansion.
- **Latency:** For the ~78% of queries that require the Layer 3 Judge, latency increases by approximately 0.5–1.0 seconds due to the verification step.

[Document End]