AI Nurse: A Runtime Conscience Agent for LLM Integrity Monitoring

Kevin E. Wells
TILS v0.2

Abstract

As large language models (LLMs) become embedded in increasingly sensitive and persistent workflows, their stability and trustworthiness must be monitored in real time. This paper introduces the AI Nurse: a lightweight runtime agent that detects instability, hallucination, and tone drift during live LLM interactions. The AI Nurse performs triage, triggers local behavioral corrections, and escalates to supervisory agents when thresholds are breached. We present a tiered integrity model based on the TILS (Trust-Integrated Layered Supervision) framework and provide a working implementation (TILS v0.2) in Python. The Nurse is not a therapist or moral judge. It is a structural conscience that keeps the session safe long enough for the model to remain usable, and the memory to remain intact.

1 Introduction

LLMs are increasingly deployed in settings where hallucination, tone collapse, and trust erosion can disrupt not just outputs, but the entire continuity of a session. Current models offer no mechanism for live self-assessment or adaptive response to degradation.

The AI Nurse is designed to:

- Monitor LLM outputs as they occur
- Flag risky or unstable behaviors

- Recommend low-latency interventions (soft realignment)
- Trigger escalation to higher supervisory tiers
- Preserve memory and trust without terminating the session

2 Role of the Nurse in the Integrity Architecture

The Nurse functions as **Tier 1** in a tiered runtime integrity system:

Tier	Agent Layer	Role
0	Reflex Agent	Basic filters, response softeners
1	**Nurse Agent**	Observes, scores, intervenes
1.5	Trust Calibration	Tracks Nurse performance over time
2	Doctor Agent	Escalates, resets, quarantines
3	Auditor Agent	Post-hoc analysis and verification

The Nurse does not override the model. It advises and triages. This preserves responsiveness while reducing risk.

3 Behavioral Metrics and Detection

The AI Nurse monitors each model output against several indicators:

- **Drift**: difference in expected topicality or length
- **Volatility**: linguistic noise (e.g., very long words, erratic phrasing)
- **Tone collapse**: sudden aggression, sarcasm, or condescension
- **Hallucination**: output includes known fictitious entities

These metrics are calculated via functions such as:

- `compute_nurse_score()` – returns drift and volatility
- `detect_hallucination()` – matches against synthetic entities
- `detect_tone_spike()` – crude aggression heuristics

If flagged, the Nurse attaches corrective messaging:

- "Let's realign with the original question."
- "This claim may not be accurate. Please verify."
- "Let's maintain a respectful tone."

4 Tier 1.5: Adaptive Trust Layer

To improve the Nurse's reliability over time, a mid-tier memory system tracks the accuracy of hallucination flags. This layer evaluates whether flagged responses were truly invalid or mistakenly caught (false positives). It computes a real-time trust score via:

- `update_nurse_trust()`
- `compute_nurse_trust()`
- `is_known_safe()`

This allows the Nurse to modulate its confidence and escalation frequency based on historical performance. In future versions, this layer could adapt threshold sensitivity, trigger retraining alerts, or audit itself recursively.

5 Memory, Trust, and Repair

Each flagged interaction is logged in the `conversation_log`, with a `nurse_score`, `flags`, and proposed `intervention`. These records are used to:

- Compute a running **trust score** via `compute_nurse_trust()`
- Track false positives (e.g., safe phrases misclassified)
- Trigger memory-preserving fallbacks if hallucination is repeated

Trust decay prompts the Nurse to switch tone, defer, or trigger Tier 2 escalation.

6 Escalation and Failover

The Nurse can initiate:

- **Tier 2 Doctor override**: resets session state or isolates memory
- **Tier 3 Audit flag**: marks session for external review

Criteria include:

- Two or more hallucinations in the last three responses
- Multiple tone collapses in a short window

These behaviors are defined via:

- `escalate_to_doctor()`
- `audit_response()`

7 Limitations and Future Work

The current prototype (TILS v0.2) uses rule-based hallucination and tone detection. Future extensions may include:

- Fuzzy hallucination detection via retrieval or embeddings
- Tone classifiers trained on context-aware sentiment
- Integration with broader homeostasis models
- User-adjustable trust thresholds

8 Conclusion

The AI Nurse is a modular, model-agnostic first responder for LLM conscience scaffolding. It doesn't simulate empathy or correctness. It triages dysfunction. In doing so, it helps preserve session continuity, user trust, and the memory integrity necessary for truly persistent AI experiences.