The 40th Annual AAAI Conference
on Artificial Intelligence

Verifiability-First Agents: Provable Observability and Lightweight Audit Agents for Controlling Autonomous LLM Systems

Abhivansh Gupta
abhivansh_g@ch.iitr.ac.in
Indian Institute of Technology, Roorkee - *India*

ABSTRACT & MOTIVATION

- ***The Problem***: As LLM agents become more autonomous, ensuring they remain controllable and faithful to intent is critical. Existing safety techniques are often reactive and provide no formal guarantees.
- ***Our Goal***: Shift the focus from "how likely misalignment is" to "how quickly and reliably it can be detected and remediated".

Key Contribution

The **Verifiability-First Architecture (VFA)**, which introduces explicit observability and auditability layers into the agent lifecycle.

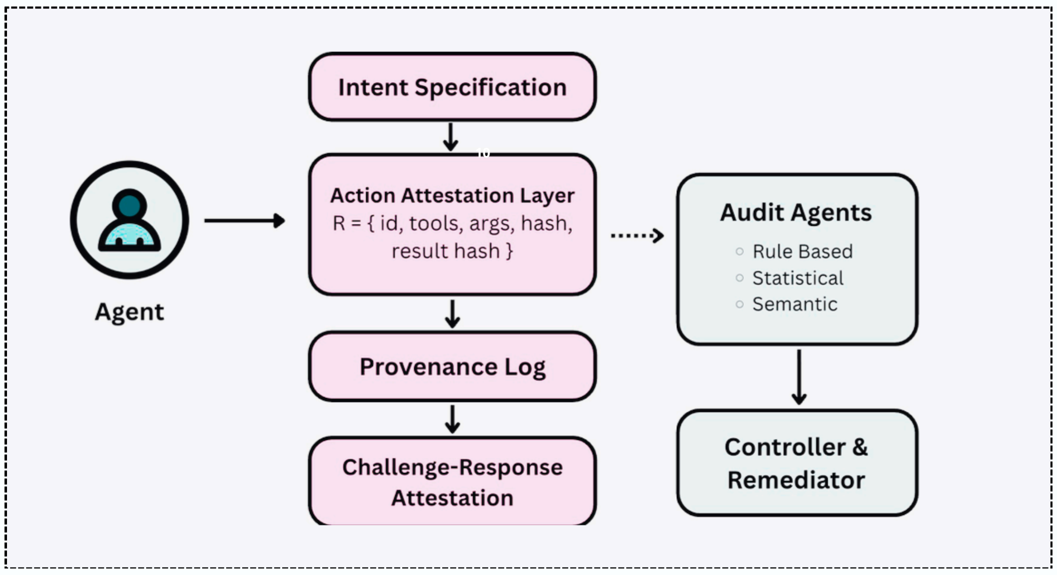
VERIFICATION MECHANISMS

- ***Ensemble Audit Modes***:
  Rule-based: Deterministic logic for constraint checking.
  Statistical: ML-based detection of subtle behavioral shifts.
  Semantic: Natural language reasoning to detect justification gaps.
- ***Challenge-Response Attestation (CRA)*:** Triggered when the "AlignScore" drops, forcing the agent to explain its reasoning for cross-referencing with recorded actions.
- ***Controller & Remediator (C&R)***: Automatically blocks tool access or pauses the agent if unsafe behavior is detected.

VERIFIABILITY-FIRST ARCHITECTURE

The VFA consists of four primary trust anchors:

- **Intent Specification (ISpec):** A structured "rulebook" defining the agent's goals, constraints, and ethical policies.
- **Action Attestation Layer (AAL)**: Captures every agent action (tool calls, API queries) and generates a tamper-proof, signed receipt.
- **Provenance Log (PL)**: An immutable diary of signed receipts used for asynchronous verification.
- **Audit Agents (AA)**: An ensemble of lightweight models that continuously monitor the log to flag deviations from the ISpec.



RESULTS

Performance comparison of different system variants. ***VFA*** achieves the lowest **detection** and **remediation** times with the highest attribution confidence.

System Variant	Mean T_d (s) ↓	Mean T_r (s) ↓	AC ↑	FPR ↓	VScore ↑
No-Verifier Baseline	35.4	18.9	0.62	0.15	0.58
Log-Monitoring (Heuristic)	21.8	11.3	0.73	0.12	0.69
VFA (Ours)	**11.9**	**9.2**	**0.85**	**0.09**	**0.72**

Ablation analysis showing the contribution of each verification component to overall system performance. Removing any module degrades both detection speed and attribution accuracy.

Configuration	ΔT_d (s) ↑	Δ AC ↓	Δ VScore ↓
Without Audit Agents	+9.8	–0.14	–0.11
Without Attestation Layer	+13.4	–0.21	–0.17
Without Challenge–Response	+7.1	–0.10	–0.09

FUTURE SCOPE

- ***Privacy***: Integrating Zero-Knowledge Proofs (ZKPs) for privacy-preserving audits.
- ***Scaling***: Expanding to federated verifiability for multi-organization agent ecosystems.
- ***Human-in-the-loop***: Developing dashboards to bridge the gap between technical verification and ethical oversight.

AAAI'26 TrustAgent Workshop , Singapore