Constitutional AI: Eliminating Hallucination Drift in Production

PUBLISHED	CATEGORY	PREPARED BY
June 2026	Security Brief	Principal Architect

ABSTRACT

Quantifying factual error risks in multi-agent LLM systems and deploying deterministic consensus mechanisms to prevent brand liability in B2B deployments.

Executive Summary

The transition to multi-agent LLM orchestration unlocks capabilities — parallel exploration, role-specialized reasoning, iterative self-critique — that no single model can replicate. It simultaneously introduces a failure mode that no single-model quality improvement can address: compounding factual error propagation across sequential context handoffs. This paper formalizes the Hallucination Drift Coefficient (HDC), presents empirical HDC measurements across pipeline depths of 2–7 agents, and introduces the Deterministic Consensus Layer (DCL) as a production-validated architectural mitigation. Findings carry direct implications for agentic AI deployment in regulated enterprise contexts.

Architectural Methodology

HDC is defined as the ratio of terminal-agent hallucination rate to single-model baseline hallucination rate on equivalent tasks at constant temperature and model family. Measurement was conducted across a benchmark suite of 4,800 factual verification tasks from legal, financial, and biomedical domains, with a single seeded factual error introduced at Agent 1.

Empirical HDC results by pipeline depth:

- **2-Agent pipeline:** HDC 1.9× — modest amplification, within acceptable tolerance for low-stakes workflows
- **3-Agent pipeline:** HDC 2.9× — crosses enterprise reliability threshold for regulated domains
- **5-Agent pipeline:** HDC 4.7× — critical amplification; source hallucination exits the pipeline at 4.7× input confidence
- **7-Agent pipeline:** HDC 6.1× — catastrophic amplification; unsuitable for any production deployment without mitigation

The Deterministic Consensus Layer addresses drift at its architectural root. Rather than passing a single agent output downstream as authoritative context, the DCL instantiates 3–5 independent agent instances with divergent random seeds and system prompt perturbations per factual claim-generating step. Outputs are resolved through three sequential verification gates:

- **Gate 01 — Majority Vote:** Claim must appear in $\geq \lceil n/2 \rceil + 1$ independent instances; seeding divergence prevents correlated failure
- **Gate 02 — Contradiction Detection:** A single dissenting instance triggers a re-query cycle with elevated temperature and an explicit adversarial prompt, stress-testing the majority position before propagation

- **Gate 03 — Source Grounding:** Claim must map to a retrievable passage in the source corpus; ungrounded claims are routed to human review regardless of majority agreement

> **Key Metric:** DCL deployment reduces HDC from 4.7× to 1.1× in legal/regulatory pipelines, from 4.2× to 0.9× in financial analysis pipelines, and from 3.9× to 1.4× in biomedical workflows — at a latency overhead of 29–44% per pipeline execution, representing a structurally acceptable reliability-latency trade-off for high-consequence enterprise AI deployments.

© 2026 VALENCE INTELLIGENCE GROUP