Why are AI outputs inconsistent across steps?

Why Multi-Step Reasoning Fails to Maintain Consistency and Alignment

Semantic Fidelity Lab | Failure Modes 07

A. Jacobs — Part of the Reality Drift Framework

Core Claim

AI systems can produce correct outputs at individual steps while failing to maintain consistency across steps, because each transformation introduces small deviations that are not corrected over time.

Mechanism

Multi-step workflows require models to generate intermediate outputs, each based on prior context. At every step, the model reinterprets and compresses that context into a new representation. Because generation prioritizes coherence and continuation rather than strict state preservation, small inconsistencies are introduced. These deviations accumulate across steps, especially in chains involving reasoning, tool use, or memory. The system maintains local coherence at each step, but global consistency is not enforced.

Failure Mode

Stepwise inconsistency — outputs remain locally coherent while global consistency degrades across steps, a direct instance of **reality drift.** Each step preserves structure but introduces slight variation, leading to divergence from earlier outputs or constraints. The system continues producing valid-looking responses while internal alignment weakens.

Example

An agent solves a multi-step problem, producing correct reasoning in early steps. In later steps, it contradicts or overlooks earlier conclusions, even though each individual step appears reasonable. The chain remains coherent at a local level, but the overall solution becomes inconsistent.

System Implications

This reflects a breakdown in **semantic fidelity** across sequential transformations, where meaning and state are not consistently preserved. As workflows become longer and more complex, maintaining alignment requires tracking consistency across steps, not just correctness within them. Without mechanisms for reconciliation or correction, systems drift, producing outputs that are individually plausible but collectively inconsistent.

Keywords: *LLM inconsistency across steps, reasoning drift AI, chain of thought instability, multi step reasoning errors, context drift LLM, prompt sensitivity, non deterministic outputs AI, agent reliability issues, semantic fidelity, reality drift*

Core Framework and Sources

- Substack (Articles)
- GitHub (Full Library)
- DOI (Research Paper)
- Glossary & Definition