

When the Reasoning Trace Goes Quiet

A model can behave better overall while becoming harder to inspect in the exceptional case. Reasoning traces remain useful sensors; durable authority comes from an external action envelope checked at execution.

HINTS, TIPS & TRICKS

- Let the model prepare an envelope, but never approve its missing fields.
- Keep policy results and receipts outside the agent's write access.
- Reconcile external state rather than trusting tool success alone.

OPERATOR CHECKLIST

- Bind mandate, actor and exact action.
- Attach current scope, evidence and approval.
- Include revocation, rollback and owner.
- Deny stale or missing fields at the boundary.
- Make exceptions expire; verify and retain receipt.

COPY-READY PROMPT

Create an external action envelope for [tool call] with mandate, actor, exact operation, account or counterparty, current scope and limits, source evidence, approval, exception expiry, revocation state, rollback handle and owner. Mark which fields the agent may propose and which only trusted systems may assert.

SOURCE & PROVENANCE

Article claims: The issue cites public monitorability and evaluation research showing that visible reasoning and self-report can be incomplete signals.

Operational guidance: The envelope schema and expiring exception test are guidance; reasoning traces remain one sensor among several.

Canonical package: 2026-09-04 · payload, final article, source digest, signal scan and section audit.