Agent Failure Modes in Production

Common breakdowns, what they look like, and the first fix to try.

Hallucination Loop

- Symptom: confident wrong answers repeating across retries.
- First fix: add eval checks + cap retries; force citations for claims.

Tool Overreach

- Symptom: agent calls tools too broadly or in the wrong order.
- First fix: tool allowlists + parameter validation; require approvals for actions.

Silent Drift

- Symptom: quality degrades slowly; users complain before alerts fire.
- First fix: canary tests + periodic re-evals; set drift thresholds.

Cost Spike

- Symptom: sudden jump in tokens/tool calls; caching misses; runaway loops.
- First fix: budgets, rate limits, caching, and loop detectors.

Compliance Blind Spot

- Symptom: missing logs, unclear provenance, no traceability.
- First fix: structured logging + audit checks; enforce logging in pipeline gates.