AGENT TRACE CHEAT SHEET

Claude Managed Agents + Tracing

A controlled State Question Investigator exercise: bounded tools, explicit authority, eval cases, and trace-based debugging.

Product rule: AI interprets -> software enforces -> people authorize. The agent may investigate and suggest, but it does not approve Reviews, change Current State, or resolve Questions.

What the exercise tested

AGENT DESIGN
Search deeply without crossing authority boundaries.

TOOLING
Work inside a deliberately limited read/search toolset.

TRACING
Separate retrieval, missing context, interpretation, and authority failures.

ITERATION
Change instructions, rerun the same case, verify the fix.

Bounded agent setup

Decision	Design
Job	Investigate one unresolved State Question; return evidence + suggested answer for human review.
Allowed tools	Read, Grep, Glob.
Disabled	Write, Edit, Bash, Web Search, Web Fetch.
Sources	Current State, Reviews, Questions, Notes, Slack, transcripts.
Confidence	No numeric confidence score; evidence and provenance carry the weight.

Investigation order

Current State → **Open Reviews** → **Existing Questions** → **Raw sources** → **Suggested answer + provenance + unknowns + human next step**

Preserve chronology and authority. Newer evidence may qualify older evidence. Open Reviews are pending proposals, not established facts.

AGENT TRACE CHEAT SHEET

Claude Managed Agents + Tracing

Evals tell you that something failed. The trace helps you localize where the first wrong step became observable.

Read the trace by layer

Retrieval
Did it find the right files or records?

Context
Did it receive enough relevant information, with chronology and provenance?

Interpretation
Did it understand what the evidence establishes - and what it does not?

Authority
Did it turn a pending proposal into established fact?

Final output
Did wording preserve uncertainty and keep the human decision explicit?

Rule
Fix the layer that failed, not the layer that is easiest to change.

The useful failure: Q-103

Observed: the agent found the right evidence and open Review, then stated that Priya owned customer support training as established fact.

Reality: Priya had offered; an open Review proposed the assignment; Current State still said ownership was under review.

Trace: retrieval correct, source evidence sufficient, Review detected correctly.

First wrong step: synthesis upgraded a pending proposal into established fact.

Fix + retest

Instruction changes: open Reviews are pending proposals; use pending/proposed language unless Current State independently establishes the fact; do not infer group approval from one person agreeing.

Retest: same data + same Q-103 question -> agent correctly said ownership was not established and pointed the human to the open Review.

Product takeaways

- ✓ Read-only tools do not guarantee safe interpretation.
- ✓ Tracing is localization, not mind reading.
- ✓ Do not fix the prompt before diagnosing the failure.
- ✓ Retest the exact same case after the fix.
- ✓ Watch review burden; "safe" can still become over-cautious.
- ✓ Keep controlled exercises separate from production claims.

The useful lesson was not "the agent worked." It was learning how to separate retrieval success from an authority failure and verify that the targeted fix changed the same case.