Engineering Reliability

A Comprehensive Guide to Evaluating AI Agents

Moving from intuition to instrumentation in the development of autonomous systems.

Eric-LLMs

The Death of the “Vibe Check”

2024: The Intuition Trap

I need an apartment. No carpet, please.

Here is a lovely unit! It has plush flooring in the bedrooms.

✓ Vibe Check Passed (Polite)

2026: The Engineering Reality

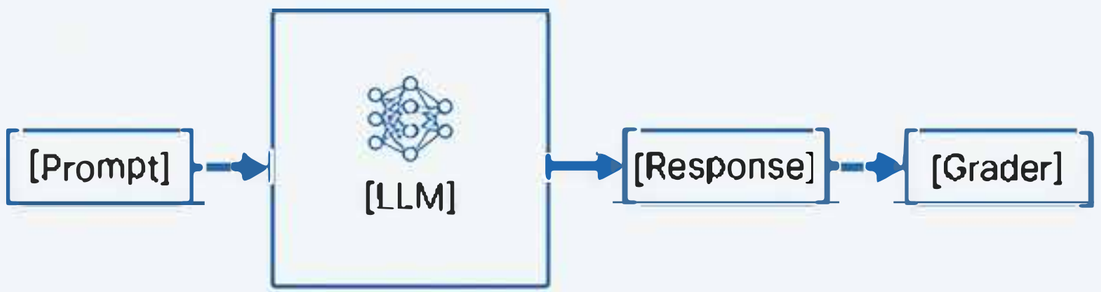
debug view

```
{
  "user_intent": "no_carpet",
  "tool_call": "search_apartments",
  "filters": {
    "flooring": "carpet"   // LOGIC ERROR
  },
  "outcome": "FAIL"
}
```

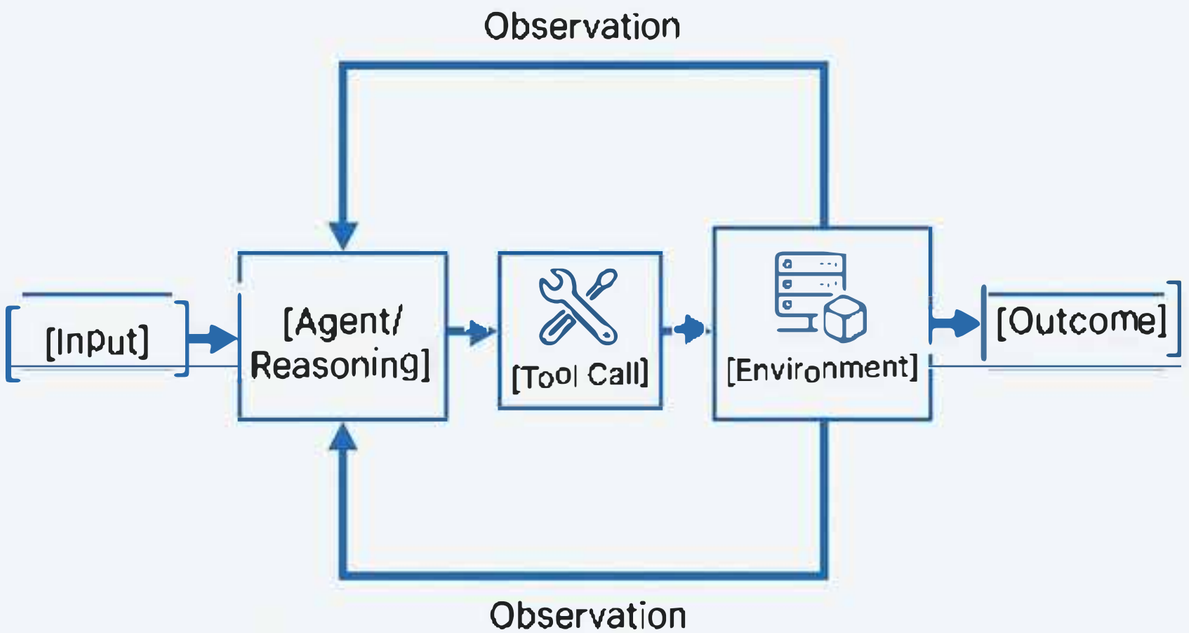
Do not outsource evals to developers alone. Domain experts must verify that the Agent is executing fluent Business Logic, not just fluent English.

An Agent is Not a Chatbot; It is a State Machine

Single-Turn (LLM)

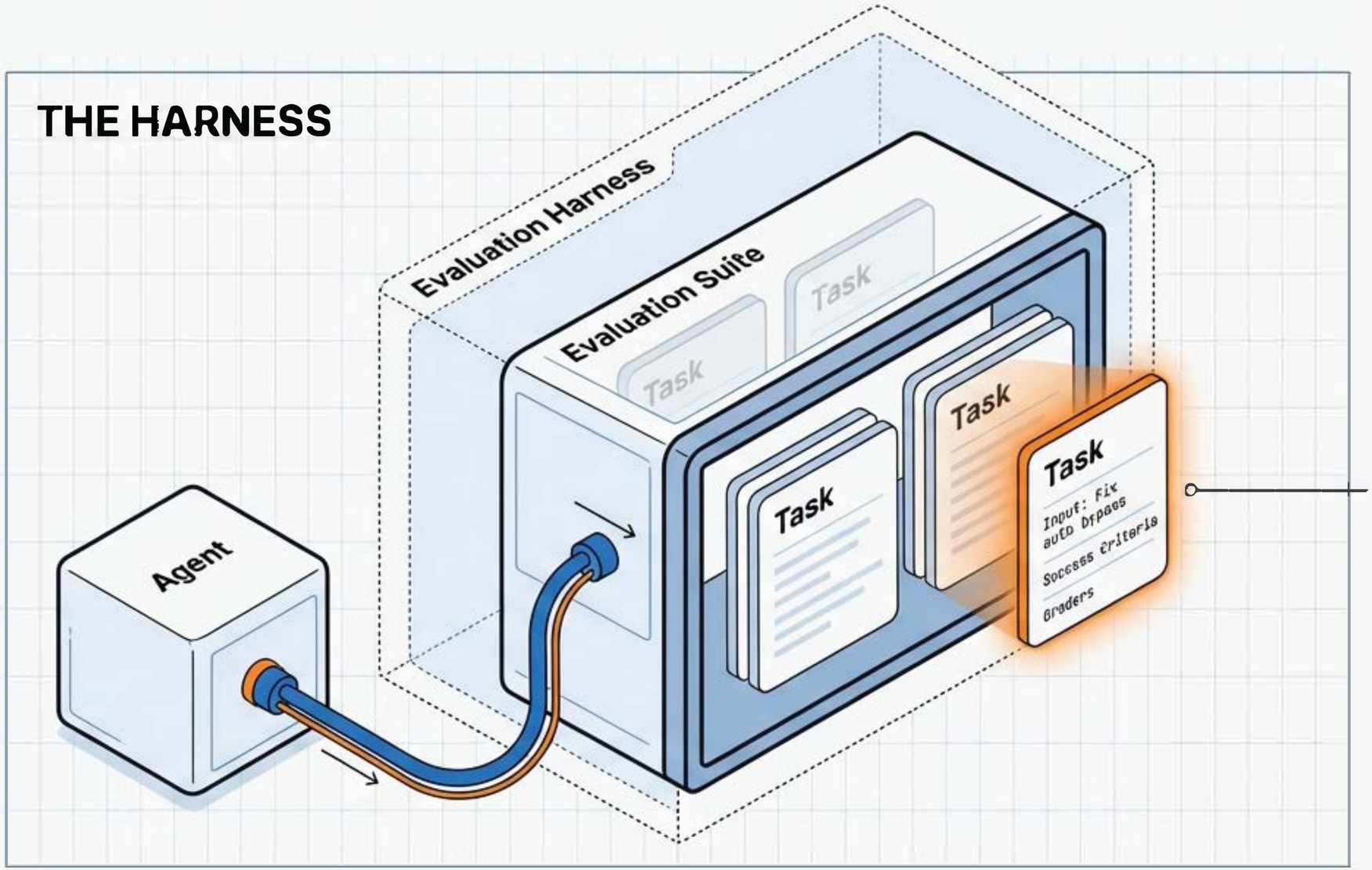


Agentic (Loop)



We must evaluate the Journey (Trajectory), not just the Destination (Outcome).

The Architecture of an Evaluation System



Key Definitions

- **Task:** A specific test case with input & success criteria.
- **Trial:** A single execution (run multiple times for consistency).
- **Trajectory**: The full log of reasoning & tool calls.
- **Outcome:** The final environment state.

Choosing the Right Ruler: The Grader Spectrum

Code-Based

Deterministic. Fast. Cheap. Brittle.

Unit Tests, String Match.

Model-Based

LLM-as-a-Judge. Flexible. Scalable. Needs Calibration.

Tone check, Summarization quality.

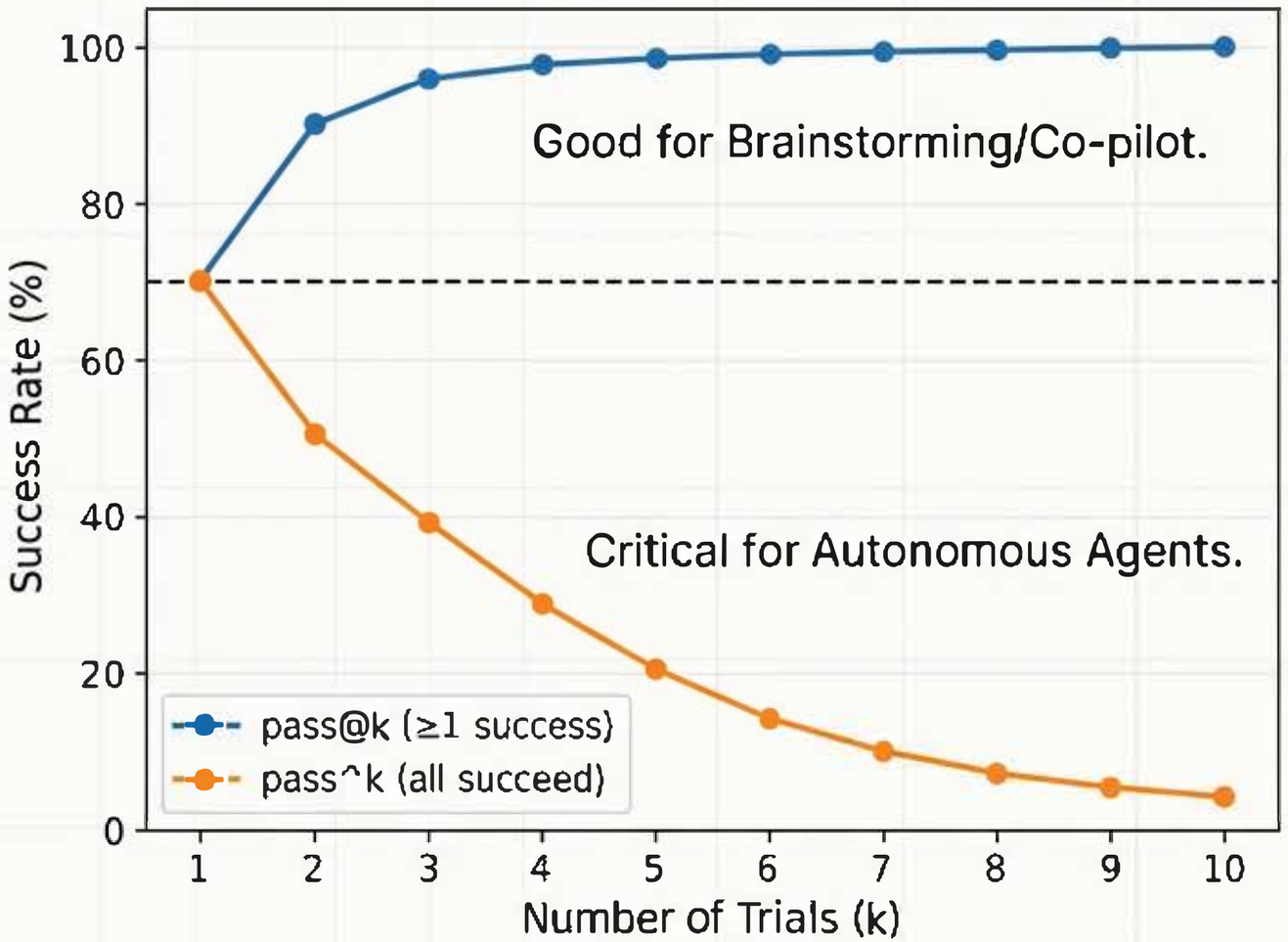
Human

The Gold Standard. Slow. Expensive.

Ground truth generation, Safety.

Methods	Strengths	Weaknesses
• String match checks (exact, regex, fuzzy, etc) • Binary tests (fail-to-pass, pass-to-pass) • Static analysis (lint, type, security) • Outcome verification • Tool calls verification (tools used, parameters) • Transcript analysis (turns taken, token usage)	• Fast • Cheap • Objective • Reproducible • Easy to debug • Verify specific conditions	• Brittle to valid variations that don't match expected patterns exactly • Lacking in nuance • Limited for evaluating some more subjective tasks

Reliability Metrics: Pass@k vs. Pass^k



Pass@k: Can the agent do it once? (Creativity)

Pass^k: Can the agent do it every time? (Reliability)

As trials increase, these metrics tell opposite stories.

Stress Testing & Performance Metrics

Agent Performance Monitor

Latency

Time to First Token (TTFT): 0.8s

Total Duration: 45s

Throughput

12.5 QPS

QPS (Queries Per Second)

Resources

85% (High)

Memory Usage

Cost

$0.15

Cost Per Task

Functional correctness is not enough. Agents must be performant

Evaluation Archetypes: Coding & Conversation

CODING AGENT

- **Goal:** Produce executable code.
- **Grader:** Deterministic (Unit Tests).
- **North Star Metric:** “SWE-bench Verified” / Pass Rate.
- **Key Check:** “Did the tests pass? Did the linter complain?”

CONVERSATIONAL AGENT

- **Goal:** Helpful, policy-compliant chat.
- **Grader:** LLM-as-a-Judge (Rubric).
- **North Star Metric:** User Satisfaction / Tone.
- **Key Check:** “Was the empathy appropriate? Was the policy followed?”

Evaluation Archetypes: Research & Computer Use

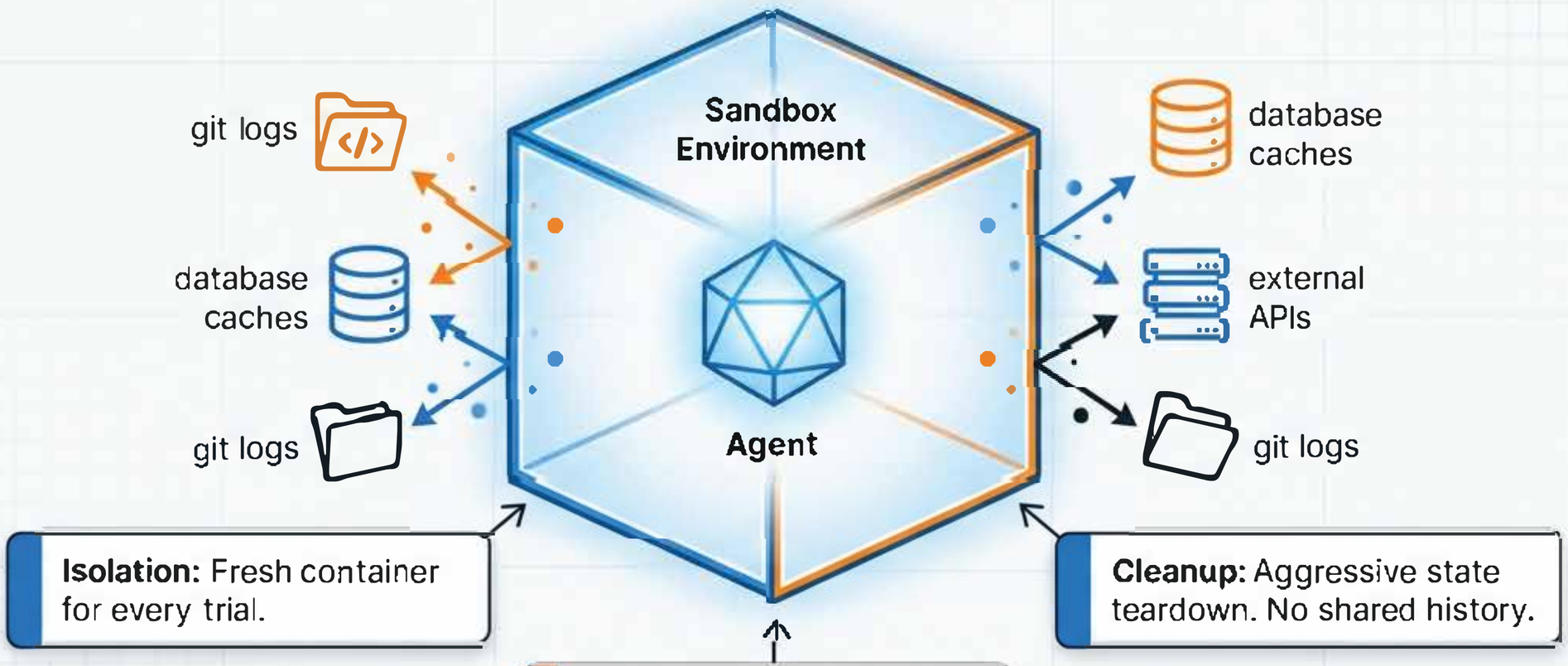
RESEARCH AGENT

- **Challenge:** Fact Drift & Subjectivity.
- **Strategy:** Groundedness Checks & Coverage.
- **Key Question:** "Does the citation actually support the claim?"

COMPUTER USE (GUI)

- **Challenge:** Visual Interface & State.
- **Strategy:** DOM Tree parsing vs. Screenshot Analysis.
- **Key Trade-off:** "Token Cost vs. Accuracy."
- **Benchmark:** WebArena / OSWorld.

The “Clean Room” Requirement

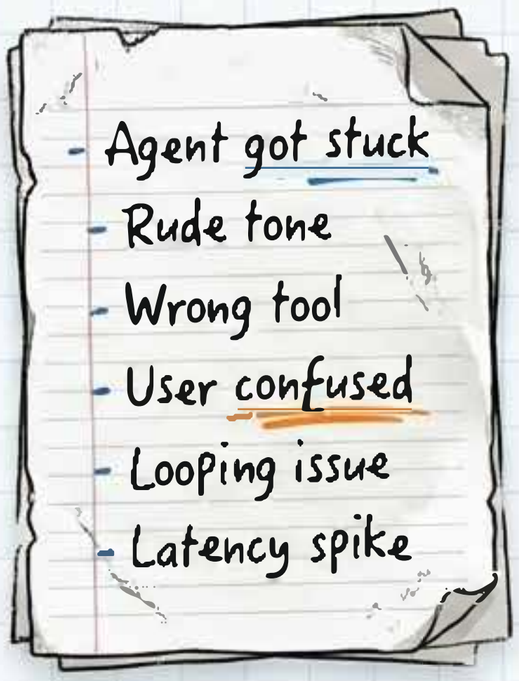


Environment Standardization & Reliability

WARNING: SHARED STATE LEADS TO CHEATING.

The Methodology: How to 'Look at Your Data'

Open Coding



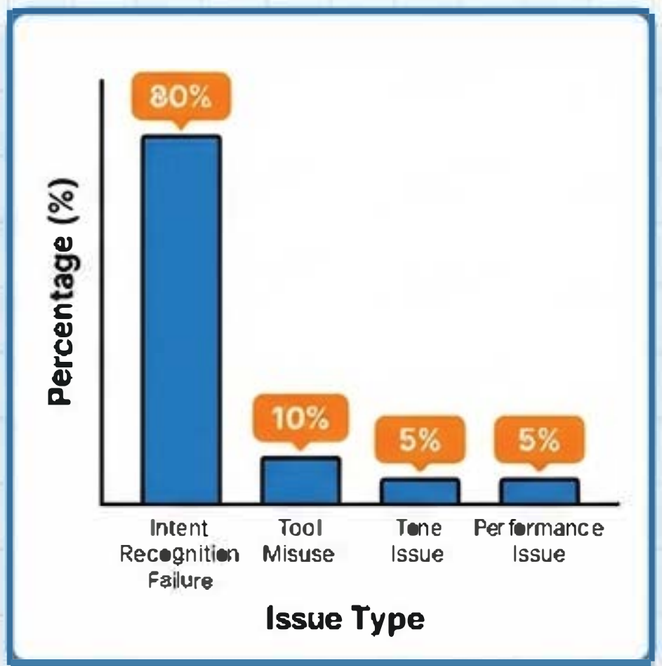
Sample 50 Logs & Observe

Axial Coding

Error Type	Frequency	Severity
Intent Recognition Failure	24	Critical
Tool Misuse	15	High
Tone Issue	7	Medium
Performance Issue	4	Low

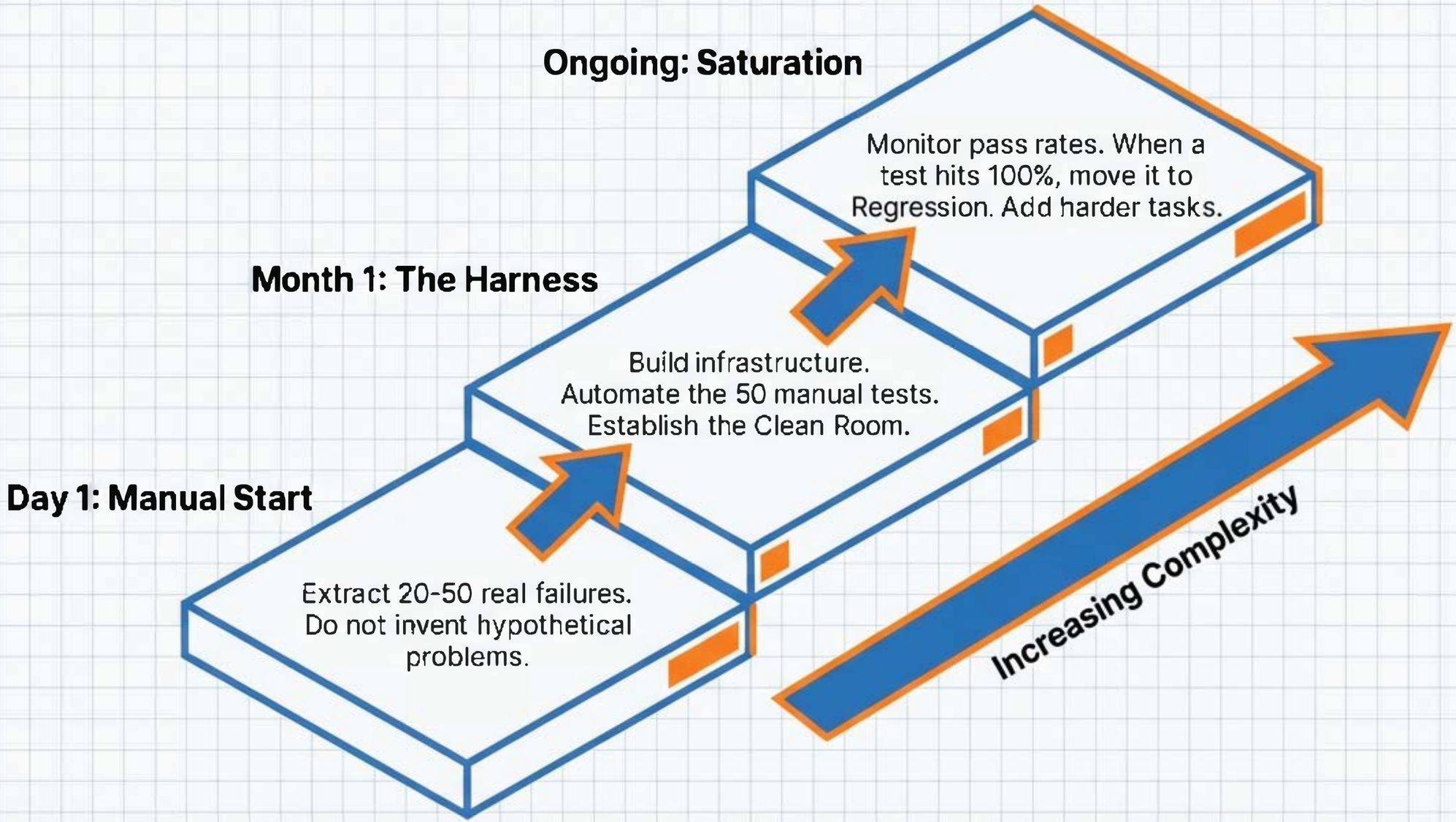
Categorize & Tag

Counting



Quantify & Prioritize

The Roadmap: From 0 to 1



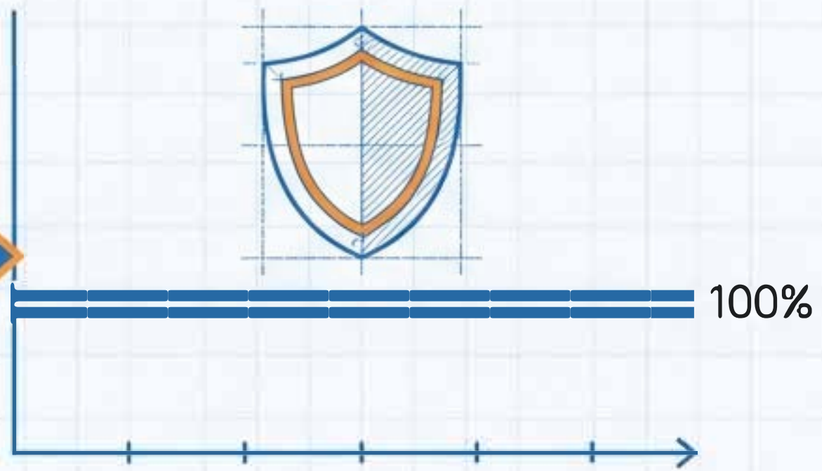
Capability vs. Regression Testing

Capability Evals

- The **Hill to Climb**.
- **Goal:** Improve weak areas.
- **Metric:** Low pass rates expected.

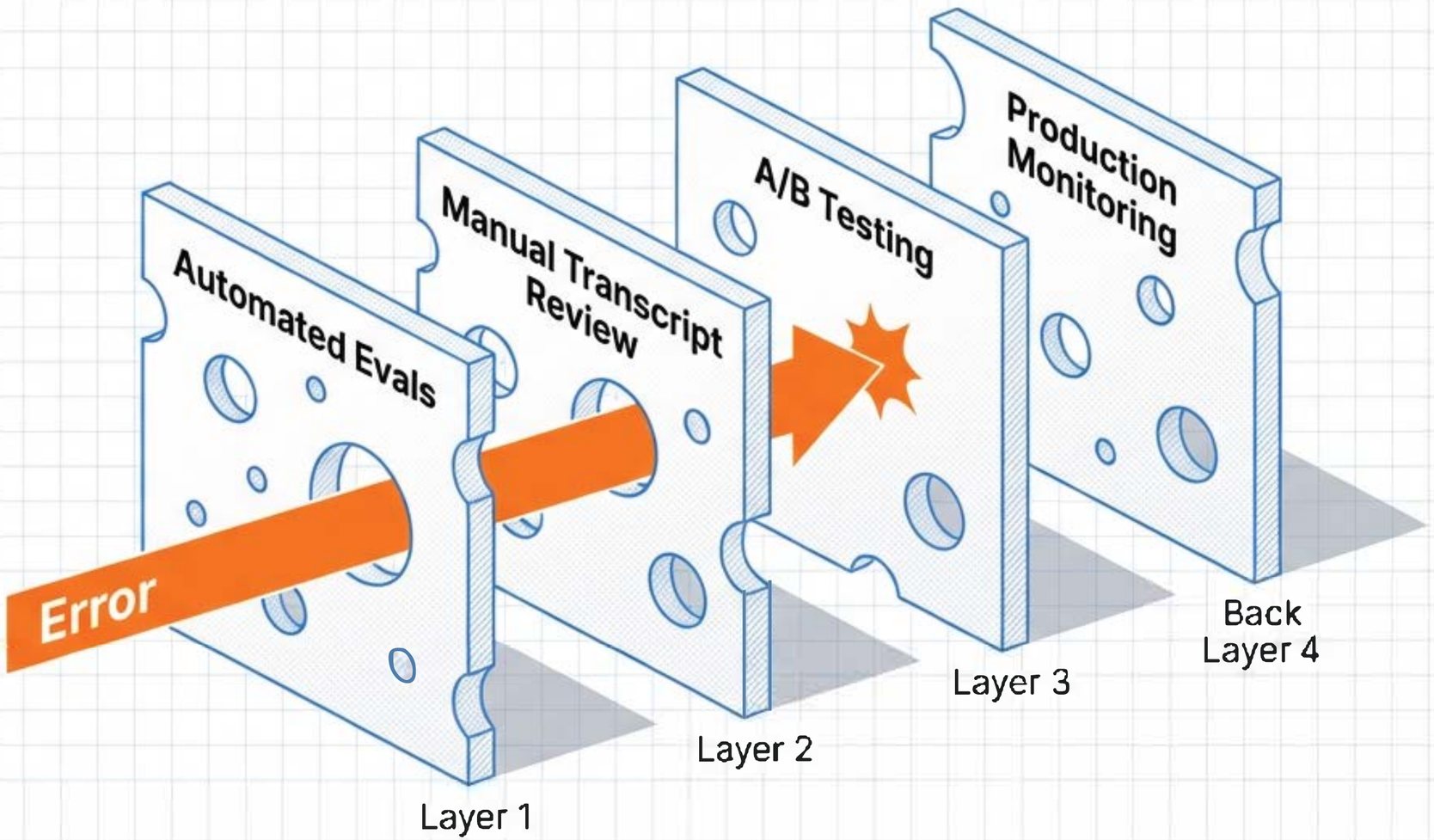
Graduation: Mastered tasks become regression tests.

Regression Evals



- The **Safety Net**.
- **Goal:** Prevent backsliding.
- **Metric:** Must be 100%.

The Swiss Cheese Model of Quality



No single layer catches every issue. Combined, they create reliability.

Evaluation *Is* Development

In the age of Agents, your product is only as good as your ability to measure it.

- [ ] **Ditch the vibe check.** Build the harness.
- [ ] **Read the logs.** Insights are in the trajectory.
- [ ] **Prioritize Reliability.** Optimize for Pass^k.
- [ ] **Start Small.** 20 real failures > 100 hypotheticals.