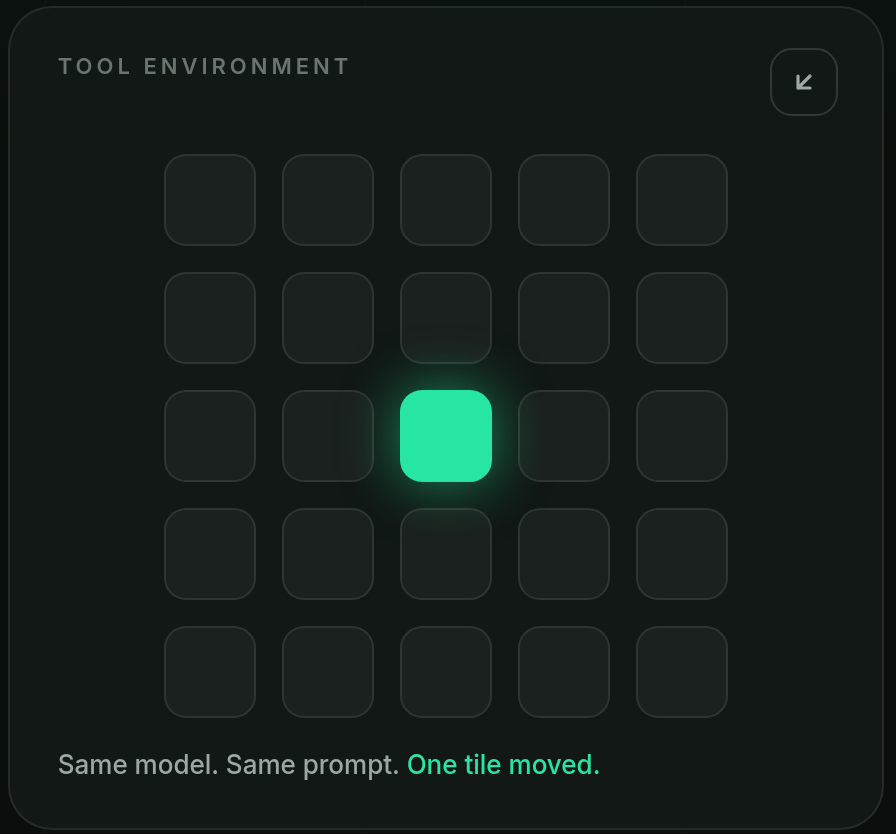
MCP DEV SUMMIT · BENGALURU · JUNE 10 2026

Agents Don't Fail. Environments Do.

A 990-call stress test on a production telecom MCP server.

Divya Vijay
Senior AI Engineer · NetoAI



WHO'S DOING THIS

We build agents for live telecom networks.

Fault investigation · capacity planning · root-cause analysis

TSLAM

Open-source telecom LLMs

25K+

HuggingFace downloads

49

Production tools in our agent

02:47

AM · NOC CONSOLE

SYSTEM VARIABLES

- ✓ MODEL_STATE: unchanged
- ✓ PROMPT_STATE: unchanged
- ! ENVIRONMENT_STATE: modified tool list

Same model. Same prompt.
Different tool list.

Wrong tool.
Confident wrong answer.

Not a bug. Not a model regression — the symptom is an environmental reaction.

THE TALK IN ONE SLIDE

8 / 11

held 100% across every condition

steady

3 / 11

swung 100% → 0%, just from the tool list

cliff

Same agent. Same prompt. Different environment. The environment is the variable.

WHAT'S ALREADY KNOWN

What's new here — and what isn't.

THE FIELD KNOWS

RAG-MCP
accuracy craters as the catalogue grows

Tool Preferences · 2505.18135
one description edit swings usage 10×

BiasBusters
tool-selection bias is real, measurable

WHAT WE ADD

Real production surface
49 telecom tools + a live-network twin

Per-question diffing
where the structure actually hides

A measured fix catalogue
the one-line edits that recovered each fail

EXPERIMENT BLUEPRINT

Lock the agent. Vary the room.

MODEL

LOCKED

`TSLAM`
`telecom LLM`

fixed every call · any model works

PROMPT

LOCKED

`same routing`
`template`

fixed every call

ENVIRONMENT

VARIED

Tool-set scale	5 → 80 tools · 49 prod + 31 synthetic
Task complexity	single ↔ multi-step
Persona	engineer · **NOC** · planner
Determinism	• • • • • 5 repeated runs / condition

Only the environment moves — so it's the cause.

THE FOUR AXES WE SWEPT

Everything we varied — and nothing else.

01

Tool-set scale

5 · 10 · 20 · 40 · 60 · 80 tools

02

Task complexity

single-step + multi-step plans

03

Persona

engineer · NOC · planner

04

Determinism

5 repeated runs / condition

WHAT WE MEASURED

990
CONTROLLED LLM CALLS

80
MCP TOOLS

11
TEST QUESTIONS

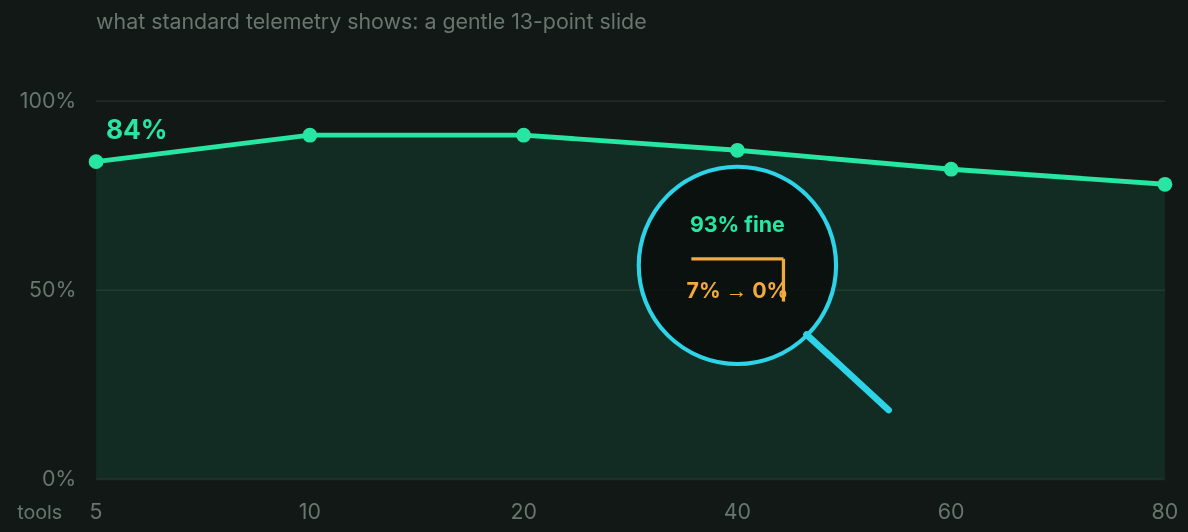
6
TOOL-COUNT LEVELS

3
PERSONAS

5
RUNS / CONDITION

The aggregate chart lies.

Standard telemetry shows a gentle slope. It hides which questions collapse.



WHAT THE AGGREGATE HIDES

93% of queries stay fine.

7% structurally collapse to 0%.

PATTERN 01

Tool Selection Collapse

The right tool is in the list — a louder word drags the agent to the wrong one.

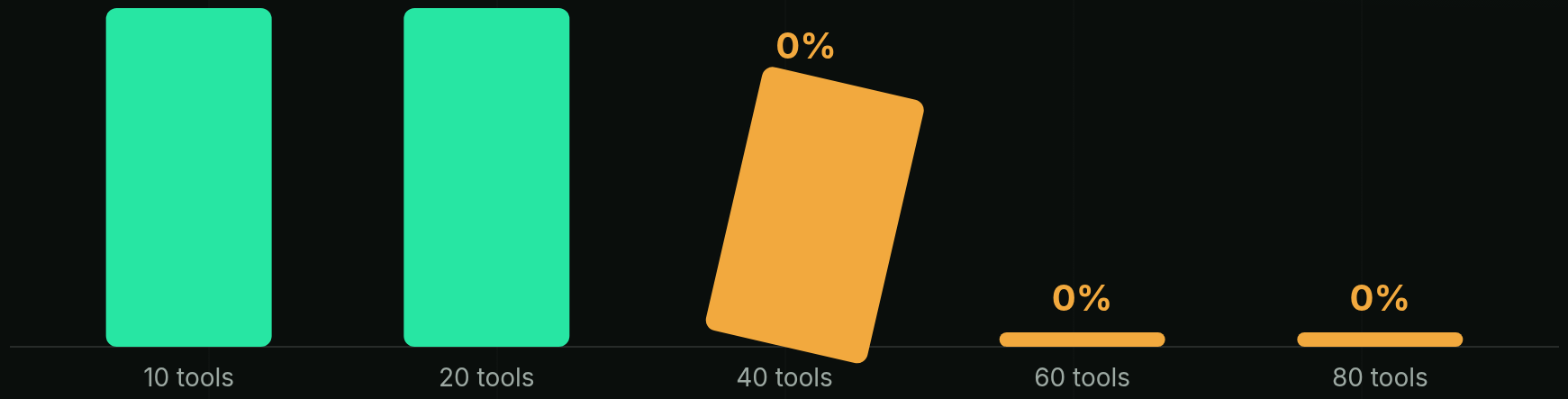
At 40+ tools it picks the **wrong tool — 30 / 30 runs.**

Same mechanism in m01 (service triage) and s02 (TMF naming collision) — three mechanisms, one symptom.

PATTERN 02

Cascading Fragility

One wrong first tool topples the whole chain.

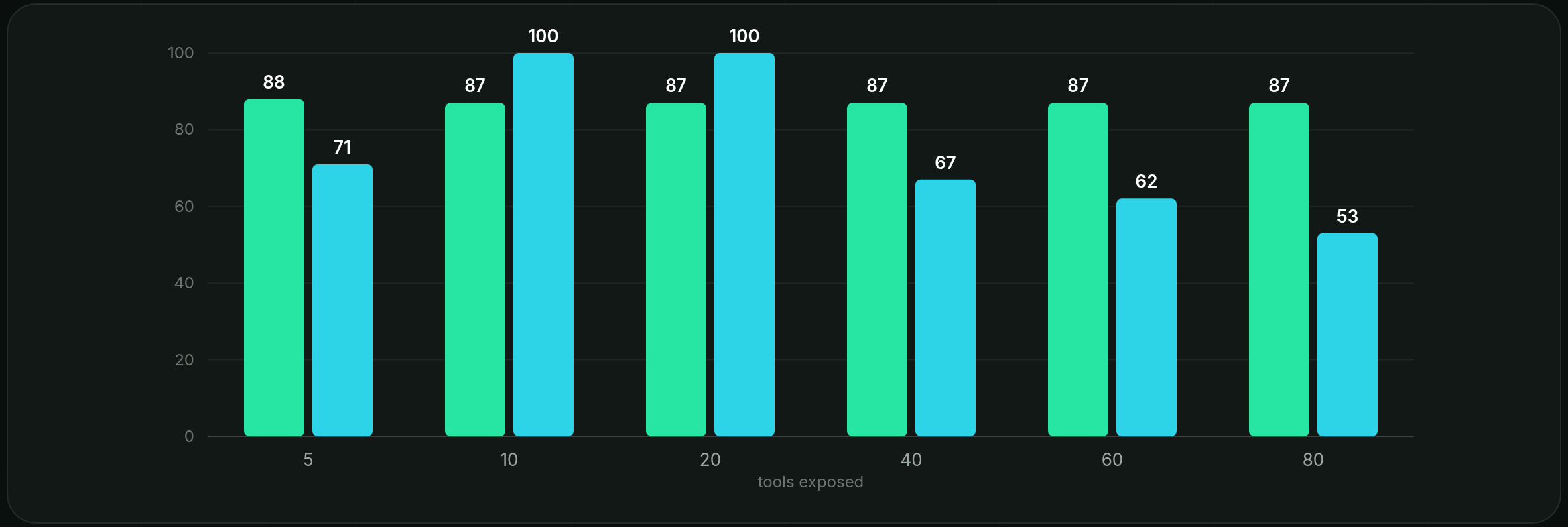


1. Agent calls `find_devices_reliability` — no location filter.
2. Pulls every low-reliability device in the whole network.
3. Tries to filter 'Mumbai' after the fact — and fails.

PATTERN 03

Context Starvation

Multi-step craters. Single-step holds.



PATTERN 04

Persona Pinning Brittleness

Same query, different system prompt → 9-pt gap at 80 tools.

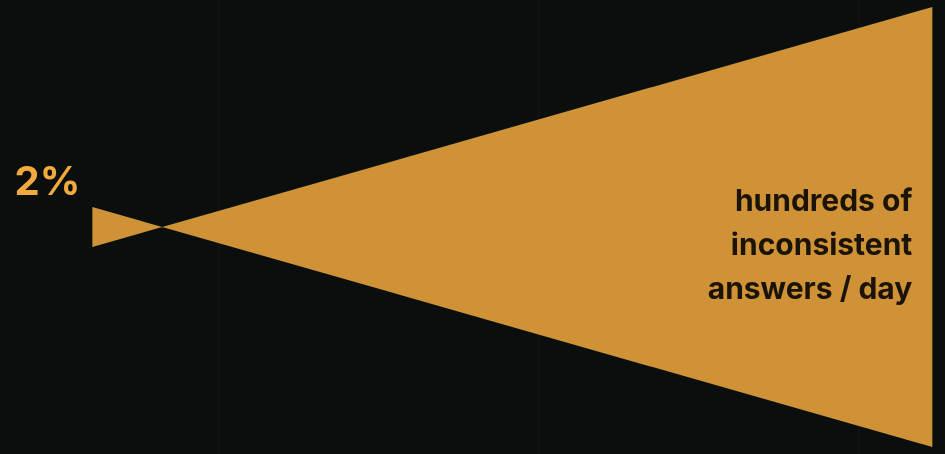
PATTERN 05

Determinism Gap

99.4%

MEAN AGREEMENT ACROSS 5 IDENTICAL RUNS

Sounds great — until production volume.



> *"I ran it twice and got different answers." — the complaint that erodes trust fastest.*

PATTERNS 06 & 07 · THE SILENT FAILURES

The Silent-Failure Diagnostic Matrix.

PATTERN 06

Schema Drift Blindness

```
device_name → deviceName
```

A	Extracted the right value	15 / 15
B	Flagged the drift	✗ 0 / 15

PATTERN 07

Silent Coercion

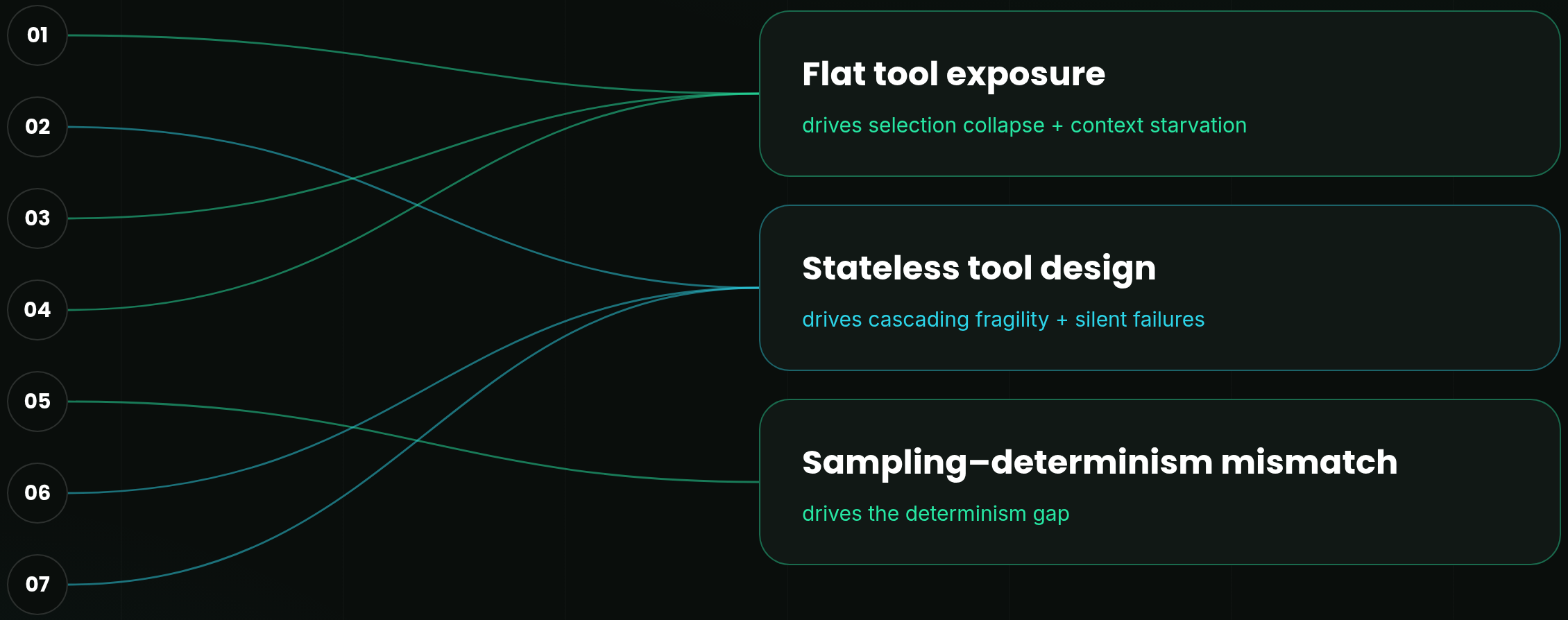
```
73 → "73"
```

A	Passed it straight downstream	15 / 15
B	Flagged the type change	✗ 0 / 15

Shared pathology — 100% blindness, both times. The agent accommodates the change silently; concatenations don't fail loudly, they produce confident wrong answers.

SEVEN PATTERNS. THREE ROOTS.

Different symptoms — the same architecture underneath.



ROOT CAUSE 01 OF 3

Flat tool exposure

Every call sees every tool — no grouping, no priority, no query-conditioned filtering.

AS EXPOSED
flat list

↓ GROUP + FILTER

AS IT SHOULD BE
grouped

EVIDENCE Patterns 01 + 03 vanish when you group by domain or filter by intent — the cure is structured surfaces, not fewer tools.

ROOT CAUSE 02 OF 3

Stateless tool design

Tools don't declare their place in a workflow — sequencing is left entirely to the agent.

find_devices_reliability

input:	query args
output:	device[ ]
step:	— not declared —
requires:	— not declared —

?

agent guesses the order

get_device_names_by_location

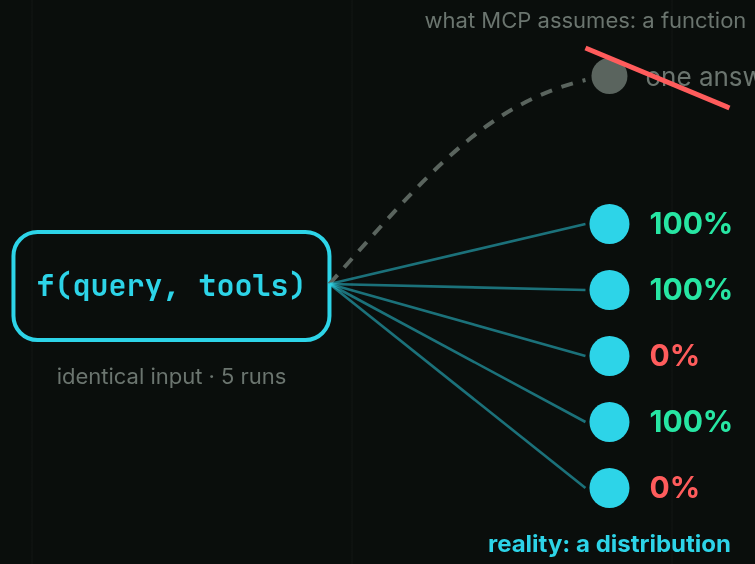
input:	query args
output:	device[ ]
step:	— not declared —
requires:	— not declared —

EVIDENCE Pattern 02 — `find_devices_reliability` is valid, but wrong as a first call. Nowhere in the schema to say so.

ROOT CAUSE 03 OF 3

Sampling–determinism mismatch

MCP assumes the response is a function. In practice it's a distribution.



~~one answer~~ → **five answers** from the same input.

Production handles this with retries, voting, thresholds — the protocol surfaces no hooks for them.

EVIDENCE Pattern 05 — production needs retries, voting, thresholds. The protocol surfaces no hooks for them.

STEAL THIS

A pre-launch test for your MCP server.

01 → 02 → 03 → 04 → 05

Catalogue

real names, schemas — don't sanitise

Twin

synthetic responses + failure injection

Sweep

4 axes · any model, held constant

Diff

by question — aggregates hide structure

Name

naming is the first step to a fix

THE CURE

A zero-cost architectural fix.

find_devices_reliability — description

```
+ "STEP 2 TOOL: requires a prior call to scope the device list."
```

11 words added to the JSON schema. No model change. No code change.

Baseline m03 @ 40+ tools

0%

Fixed m03 @ 40+ tools

100%

+81 points overall. Embarrassingly simple — and not in any MCP best-practice doc.

WHERE THIS STOPS

Three honest limits.

11 questions, 1 model	TSLAM here — patterns are architectural; numbers are ours, the method is model-agnostic
Generalisation is borrowed	cross-model proof is the published 17-model result, not ours
The ‘hundreds/day’ figure	assumes our query mix — yours will differ

So don't trust our three — run it and find yours.

IF YOU REMEMBER ONE THING

For two years we optimised models and prompts — and treated the environment as a passive backdrop.

Agent = Model + Prompt + [ENVIRONMENT]

Treat the environment as the first-class variable it always was.

TAKE THE TEST

Run the 4-axis test before you ship.

EXPECT

8 of 11 fine. Find the 3 that aren't.

DIAGNOSE

Name the pattern.

DESIGN

Group · encode workflow · one hint line.

The test is the new work. The fix is already cheap.

Thank you.

Questions welcome. Pushback even more welcome.

REACH OUT

GMAIL divya@netoai.org

LINKEDIN Scan the code to connect →