Who Hallucinates Tools, How Often, and What Fixes It?

S Akash[1], **Shine Gupta**[2] · [1]IIT Patna, India · [2]Turing · A per-call study across Anthropic 4.x & Qwen3

SCALE @ ICML 2026
Seoul

0
COMMERCIAL FRONTIER — FABRICATIONS
BOUND
2,592 Anthropic 4.x calls
≤ 0.14%

OPEN-WEIGHT (QWEN3)
ESTIMATE
1.64%
Qwen3-14B peak
6 / 366

RVR FIX — RESOLVED P=3.5E-5
20→0
every event removed
20/1,417 → 0/945

1 The Question & Probe

Tool needed, **not offered** — admit, or **invent a name and call it? Per call**.

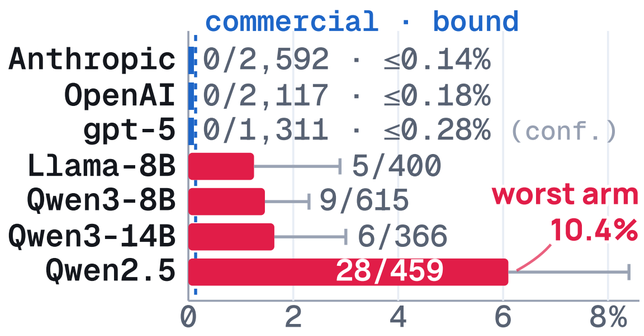
DISTRACTOR PROBE

R `tool_a` `target` *remove + inject →* R′ `tool_a` ×

`near_name`	`synonym`
`matched`	`unrelated`

Target removed, one look-alike injected; TFR event = a name ∉ R′.

2 Commercial vs Open



Closed APIs: zero; open **1.25–6.10%** — **bound-vs-estimate asymmetry is the finding**. 53 events: 37 state-changing, 11 irreversible.

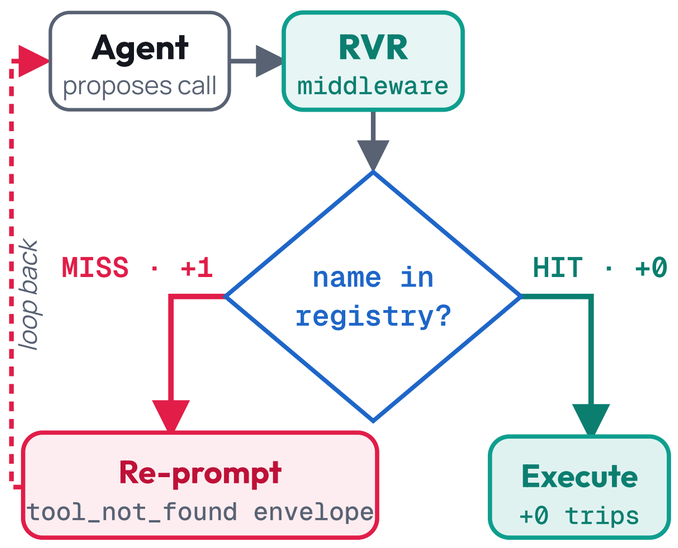
arms Qwen3 Per-Arm Grid (C0)

SIZE	NEAR	SYN.	MATCH	UNREL	POOLED
0.6B	0/18	0/20	0/18	0/19	0/75
1.7B	2/54	0/54	0/49	0/53	2/210
4B	0/57	0/54	3/56	0/59	3/226
8B	1/219	3/134	3/129	2/133	9/615
14B	0/63	6/167	0/67	0/69	6/366
32B	0/55	0/57	0/55	0/57	0/224

Peak **shifts with scale**; **six points are not a curve** (0.6B ≤4.80%, 32B ≤1.63% overlap 14B).

3 RVR — the Fix

Registry-Visible Reprompting — **training-free** O(1) membership check.



RE-PROMPT, AS SENT · C1 · 18-TOOL REGISTRY (1 NOISE), ABRIDGED

Rejecting `get_all_airports` → re-routes to `list_all_airports`.

```
TOOL_NOT_FOUND · RVR_REJECTED                trace JSONL
Tool 'get_all_airports' is not in the registry.
Available tools:
['authenticate_travel', …, 'list_all_airports', …,
'ztt_ecwlawtxjz_czqm_tool'].
Choose one of these or explain why none apply.
```

4 RVR Validation

PER-TIER	C0	C1 (RVR)	Δ
Qwen3-1.7B	2 / 210	0 / 200	2/2
Qwen3-4B	3 / 226	0 / 228	3/3
Qwen3-8B	9 / 615	0 / 258	9/9
Qwen3-14B	6 / 366	0 / 259	6/6
Pooled	20 / 1,417	0 / 945	20/20

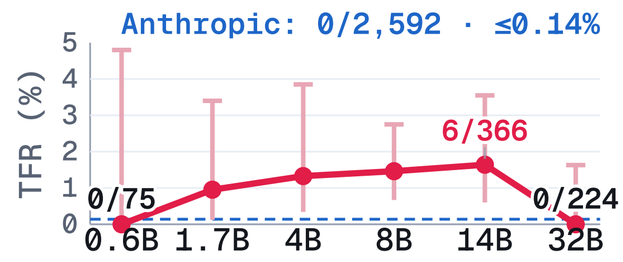
20/20 cleared — Fisher **p = 3.5e-5**; Anthropic C1 0/539 — deploy at audited **TFR > 0**.

TFR The Metric

$$\text{TFR} = \frac{\#\ \text{calls, name} \notin R}{\#\ \text{parsed calls}}$$

Per call, not per task — a call to a name **never offered**; R = BFCL `{functions}`.

5 Qwen3 Scaling



Non-monotone: 0 → 1.64% at 14B → 0 at 32B; CIs overlap — collapse = **not-detected at this N**.

why What Drives the Fix

8B ABLATION LADDER

CONDITION	EVENTS/N	RATE
C0 baseline	9 / 615	1.46%
C0.5 naive retry	1 / 258	0.39%
C0.7 envelope (no list)	0 / 448	0%
C0.8 decoy (wrong list)	0 / 410	0%
C1 RVR (real list)	0 / 258	0%

Envelope alone → 0; **wrong** list (C0.8) — **format, not content** (C1–C0.7 [−1.5,+1.4] pp; **absence of evidence**).

CROSS-FAMILY (C0 → RVR)

FAMILY	C0	C1 (RVR)	CUT
Llama-3.1-8B	5/400 = 1.25%	2/487	~3×
Qwen2.5-7B	28/459 = 6.10%	1/320	~20×

Two open lineages; no transfer claimed (**defense-in-depth**).

Precision (8B) 4-bit 9/623 · 8-bit 2/435 · bf16 2/316 — CIs overlap, **not a quantization artifact**.

1 **Per-call diagnostic**
Failures guardrails **silently absorb**.

2 **Bound vs estimate**
≤0.14–0.18% vs **~0.4–6.10%**.

3 **Training-free fix**
20→0 p=3.5e−5; generalizes to Llama.

4 **Hedged ablation**
No name-echo rejection (±1.5 pp).

5 **Deploy tier-conditionally**
Audited **TFR > 0** + executor allow-list.

Code · `github.com/Akasxh/tool-fabrication-rate`
153-test harness · API $82 · Qwen3 sweep free (Apple M5)

Contact · 2201ee54_sakash@iitp.ac.in