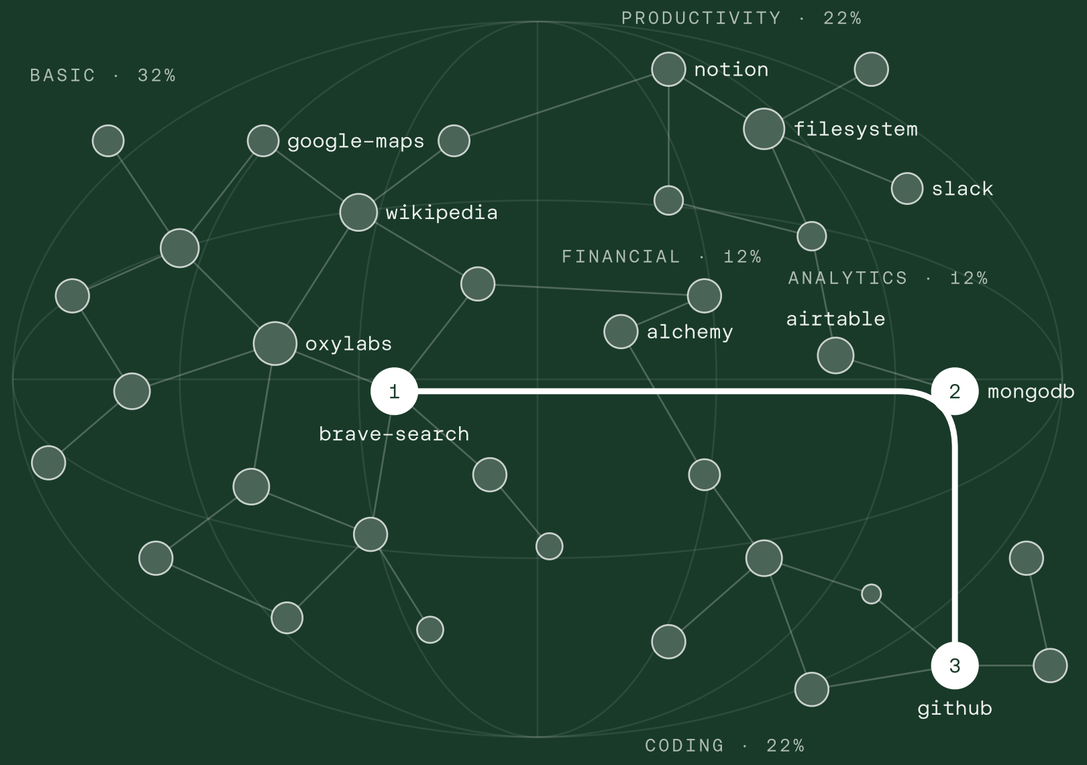
MCP-Atlas

A Large-Scale Benchmark for Tool-Use Competency with Real MCP Servers

Chaithanya Bandi*, Razvan-Gabriel Dumitru*, Ben Hertzberg, Divyansh Agarwal, Geobio Boo, Tejas Polakam, Jeff Da, HiJae Kim, Vipul Gupta, Manasi Sharma, Andrew Park, Martin Dimakis, Ernesto Gabriel Hernández Montoya, Rafael Cruz, MohammadHossein Rezaei, Chetan Rane, Daniel Yue Zhang, Brad Kenstler, Bing Liu

Scale AI · *Equal contribution

PRESENTED BY Razvan-Gabriel Dumitru



Even the best AI agent fails nearly 1 in 5 real tasks.



82.2% of 1,000 tasks passed by Muse Spark, the best of 20 models

17.8% failed

Expert-written solutions put the ceiling near 100%, so about 18 points of headroom remain.

THE GAP

Existing MCP evaluations fall short in three ways

01

Realistic workflows

Multi-step tasks that orchestrate tools across several servers.

IN MCP-ATLAS

98.6%

of tasks need two or more servers

02

Real servers, not mocks

Authentic MCP servers, with the failure conditions that mocked environments leave out.

IN MCP-ATLAS

36

production servers, exposing 220 tools

03

Reproducible scoring

Structured claim-level scoring that doesn't depend on how verbose or polished the answer is.

IN MCP-ATLAS

3

LLM judges cross-check a partial-credit claims rubric

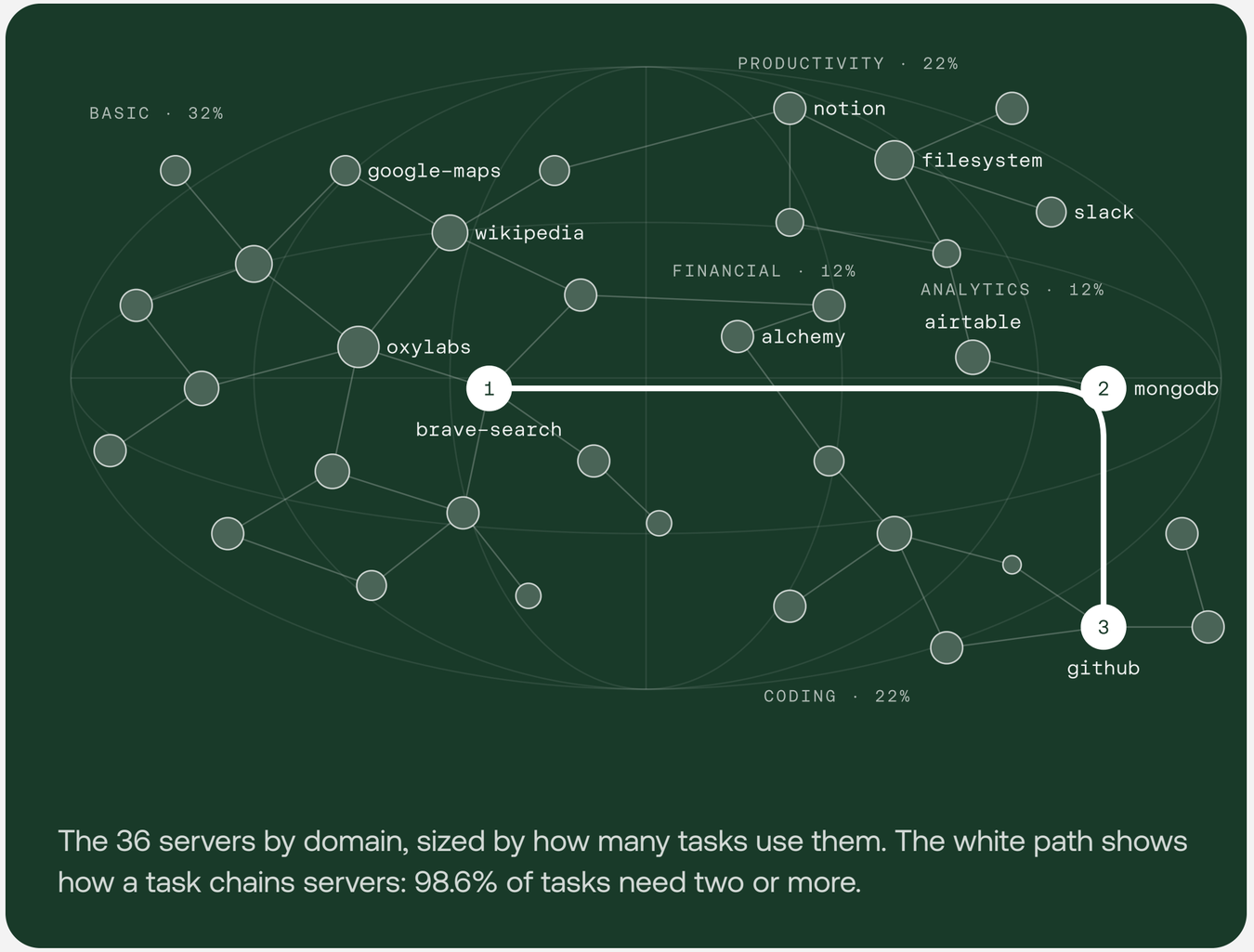
MCP-Atlas at a glance

1,000
tasks written and verified by experts

36
real MCP servers

220
tools

20
frontier models from six providers



The 36 servers by domain, sized by how many tasks use them. The white path shows how a task chains servers: 98.6% of tasks need two or more.

What a task looks like

THE REQUEST

"...a 2024 paper by Jane Castleman that deals with ad control effectiveness, can you get me the abstract? ... the date of our campaign with the biggest engagement rate, started during the 2015–2023 period, and its locality."

15 TOOLS ON OFFER · 3 NEEDED

`arxiv_search_papers`

`notion_API-post-search`

`notion_API-post-database-query`

+ 12 distractors: notion_API-retrieve-a-page, slack_channels_list, whois_whois_domain, fetch_fetch …

SCORED ON 5 CLAIMS · PASS AT ≥ 75%

- The paper's title
- Its abstract
- Three campaigns tie for the top engagement rate, 15%
- Their start dates: 2022-06-24, 2019-09-20, 2017-09-09
- Their localities: National, International, International

6–37 tools offered per task
mean 15.2

2–8 actually needed
mean 4.1

Prompts never name a server, tool or parameter. The agent has to work out the tool chain itself.

We score the answer, not the path

STEP 1

The agent answers

It reads the request, picks tools, calls them and writes a final answer.

STEP 2

Atomic claims

Each task lists the facts a correct answer must contain: 4.7 claims on average.

STEP 3

A judge checks each claim

- fulfilled · 1.0
- partially fulfilled · 0.5
- not fulfilled · 0.0

STEP 4

Coverage decides

Coverage is the mean claim score. A task passes at coverage ≥ 0.75.

ILLUSTRATION · A TASK WITH FIVE CLAIMS

1.0 · 1.0 · 1.0 · 0.5 · 0.0 mean **0.70**

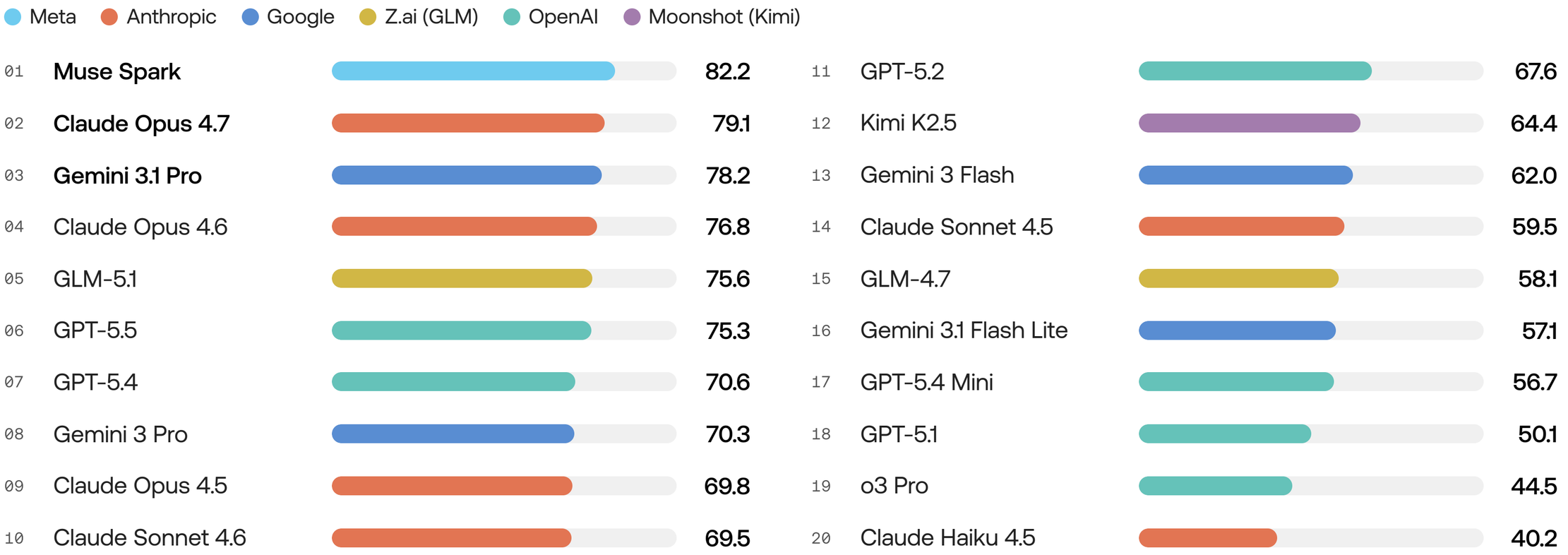
below 0.75: the task fails

Any valid sequence of tool calls earns credit: scoring looks only at the final answer.

Every failed task also gets an automated diagnosis from an 11-category failure taxonomy.

RESULTS

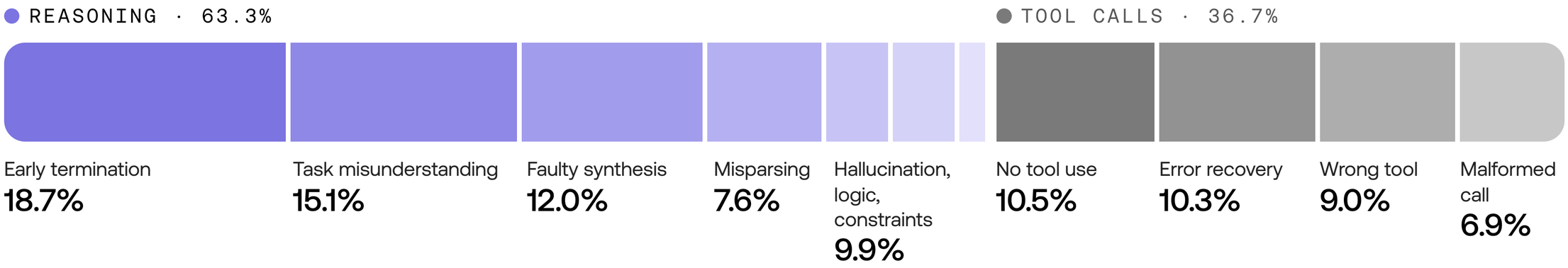
Close at the top, far from 100%



Pass rate (%) at claim coverage ≥ 0.75 on all 1,000 tasks; bars run to 100%. The top three sit within 4 points, inside the 2–5 point spread between judges. Gemini models are preview releases.

DIAGNOSIS

Most failures aren’t about calling tools



Early termination

Understood the task but stopped before completing all the required steps.

Task misunderstanding

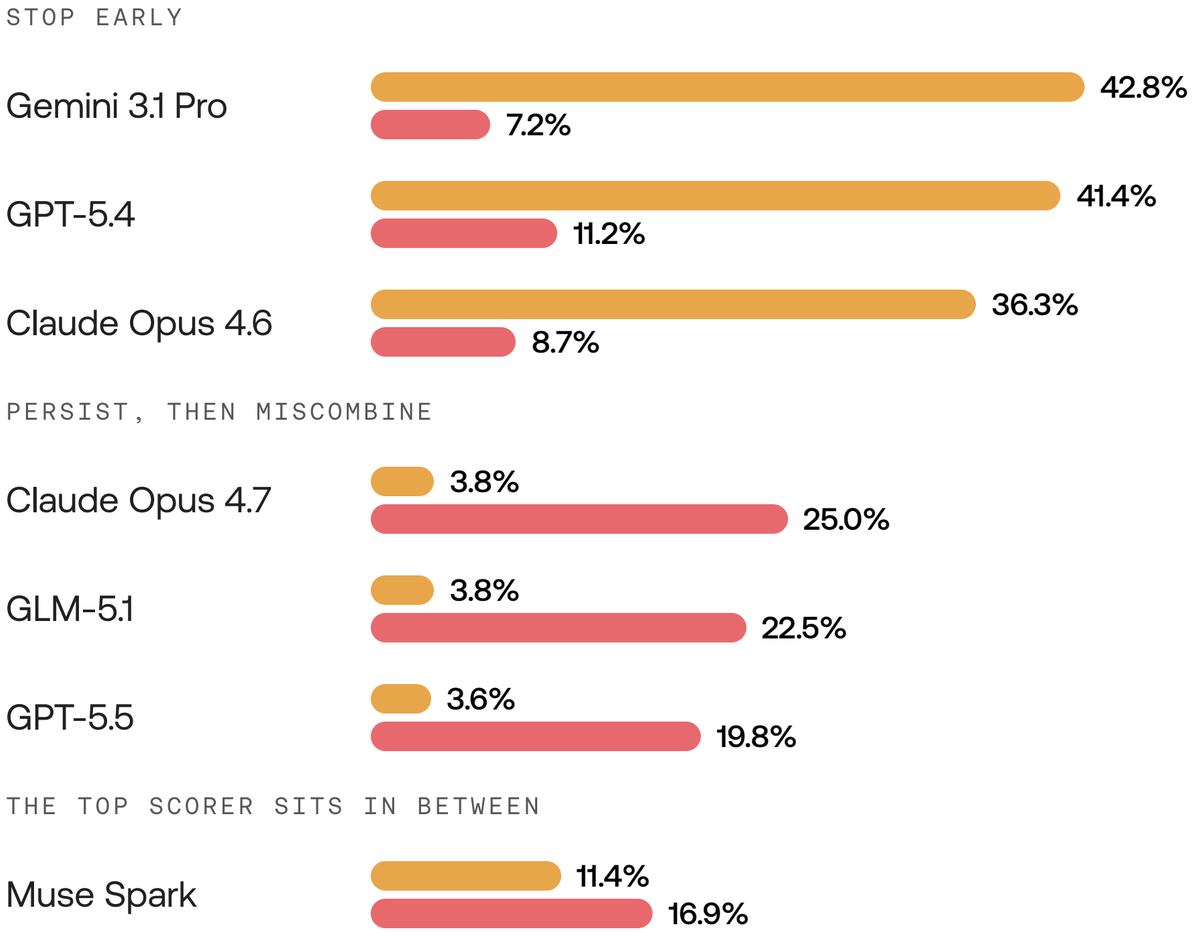
Answered a different question, or missed a key requirement in the prompt.

Faulty synthesis

Had the right tool outputs but combined or interpreted them incorrectly.

Share of ~6,900 diagnosed failures across 20 models. Definitions follow the paper’s 11-category taxonomy.

Two ways to fail at the top



Early termination

Understood the task but stopped before completing all the required steps. E.g. found one half of a two-part answer and stopped.

Faulty synthesis

Had the right tool outputs but combined or interpreted them incorrectly. E.g. got the right rows, then averaged the wrong column.

The newest models have almost stopped quitting early. Their errors reappear later, as harder failures after the tools have answered.

AN OUTLIER

Reasoning without tools

40.1%

of o3 Pro's failures make no tool call at all.

WHAT IT DOES INSTEAD

44% answers from memory

34% asks the user a clarifying question

Passing really does need tools: only 0.27% of passing runs make no tool call.

VALIDATION

Can you trust the scores?

95.3%

The judge matches human experts

Three independent experts labeled 338 claims from 100 tasks, blind to model and score. The judge agreed with their majority (95% CI 92.7–97.3).

±0.34 pp

No judge favors its own provider

After removing model and judge effects, every judge's own-provider effect stays within this range. All three judges rank the same six models on top.

≤1.4 pp

Reruns barely move the numbers

Running Claude Opus 4.7, Gemini 3.1 Pro and GPT-5.5 three times each changed pass rates by at most this much, and never changed their order.

24 / 24

The failure diagnosis holds up

On 50 failures labeled by three domain experts, the diagnostic judge matched every unanimous label for tool-call versus reasoning.

LIMITATIONS

What MCP-Atlas doesn't measure yet

01

Read-only tasks

No task creates, updates or deletes data, so state-changing operations are untested.

02

Single-turn, English only

Multi-turn interaction, multilingual and localized tool use are left as future work.

03

Live servers drift

Server versions are pinned, but real servers can change over time beyond our control.

04

A 100-call budget

The limit may penalize models that rely on retrying often.

05

Judge and claims

The judge leans slightly lenient, and claim extraction can carry some noise.

06

Expert solvability

Expert solutions show the tasks are solvable, not how often an independent expert would solve them unaided.

MCP-Atlas

01 Even the best agent fails nearly 1 in 5 real tasks.

02 Most failures are reasoning, not tool mechanics.

03 500 public tasks, the harness and the evaluator are open.

GET THE BENCHMARK

CODE & DATA

CONNECT WITH ME

LINKEDIN

X

Thank you · Razvan-Gabriel Dumitru · razvan-gabriel.dumitru@scale.com