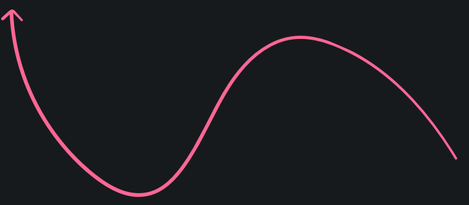
Hi. Me

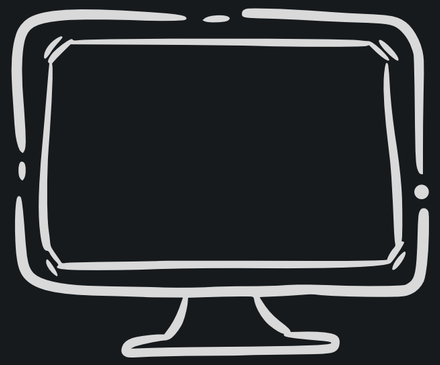
MOM

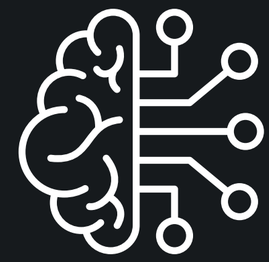
ALL PURPOSE

THE OUTDOORS

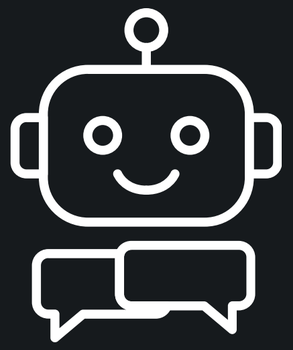
Thanks for having me!







Building Trustworthy Agents at Scale



Raise your Hand

Shipped an agent to production?

“In my little group chat with my tech CEO friends there's this betting pool for the first year that there is a one-person billion-dollar company. Which would have been unimaginable without AI and now will happen.”

Sam Altman

Trust

State of Code Developer Survey report

84%

State of Code Developer Survey report

29%

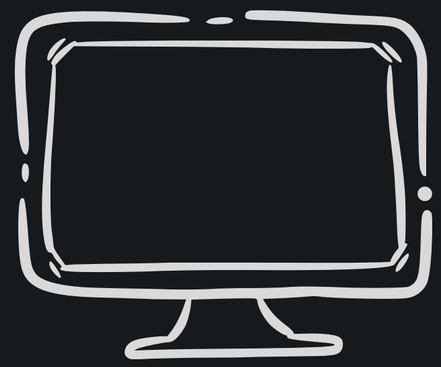
89%

52%

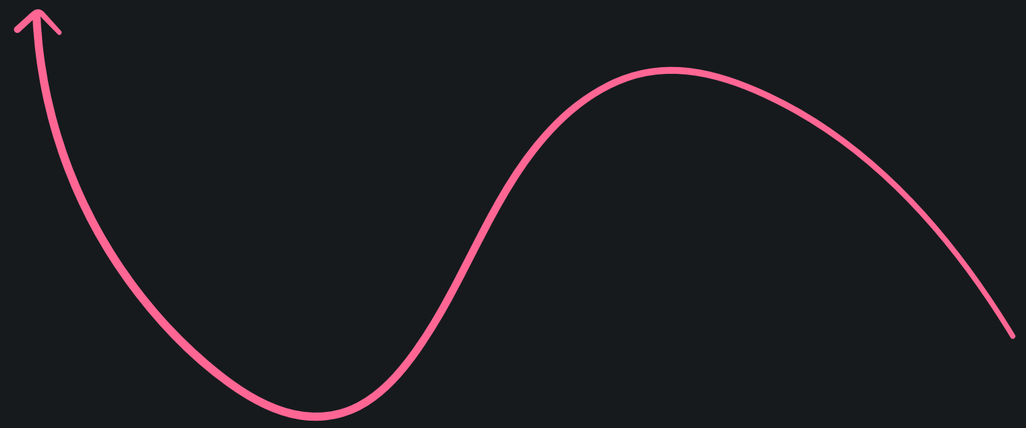
The real problem isn't capability.

It's trust.

How can we build this trust as engineers?

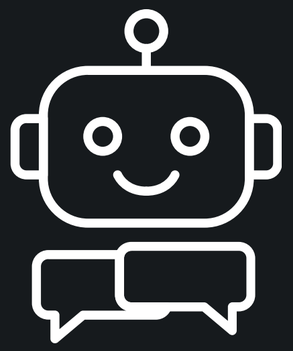


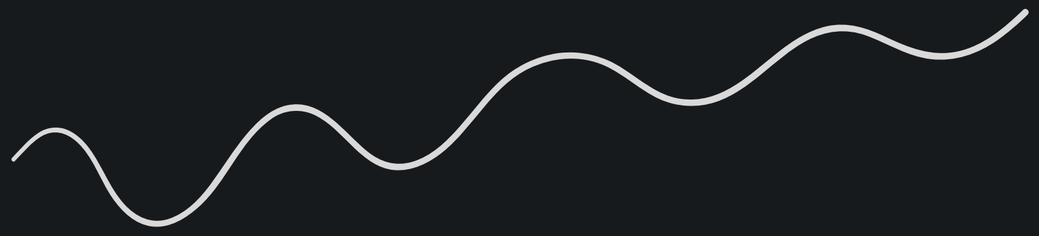
Evals



Same input.

Different output.

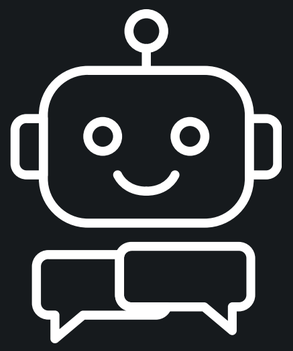


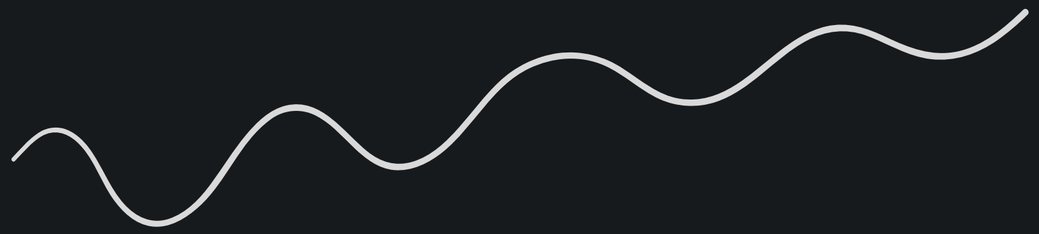


"Good evaluations help teams ship AI agents more confidently."

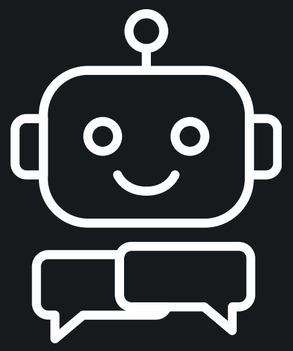
~ Anthropic

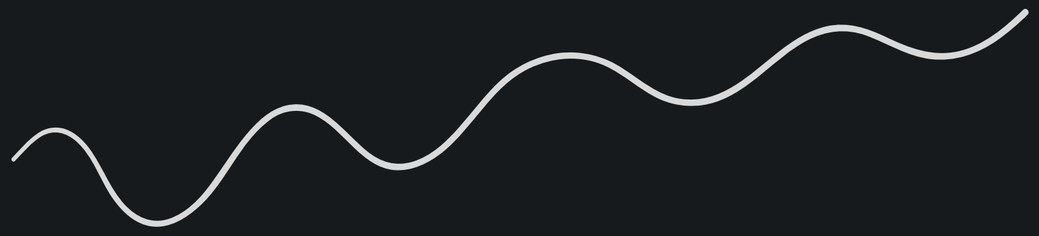
End-to-end





End-to-end Trajectory-level

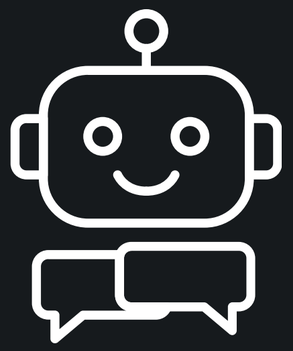


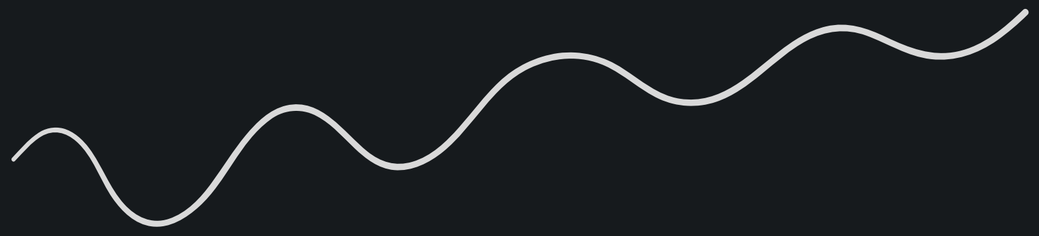


End-to-end

Trajectory-level

Component-level

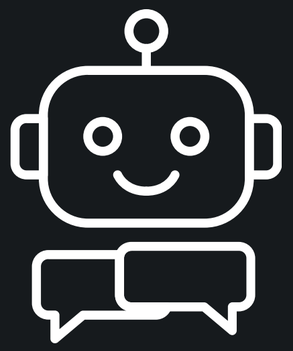


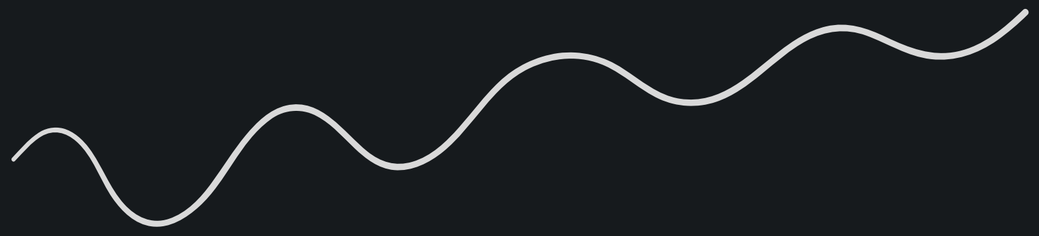


End-to-end

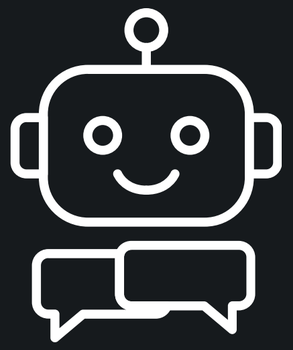
Trajectory-level

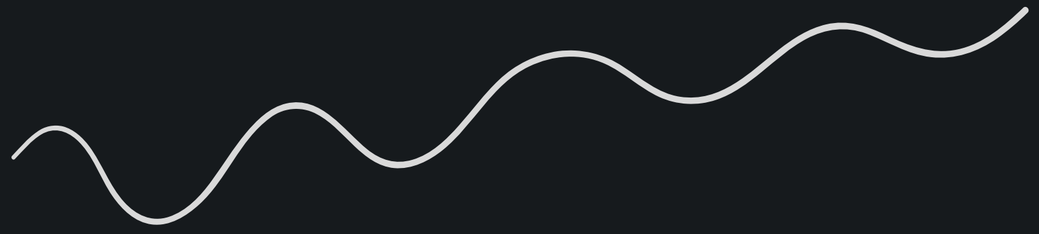
Component-level





N rollouts per scenario

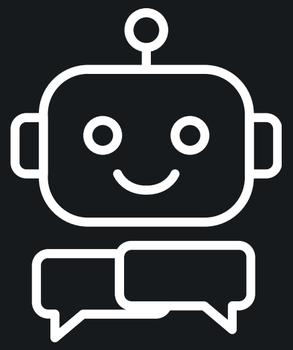


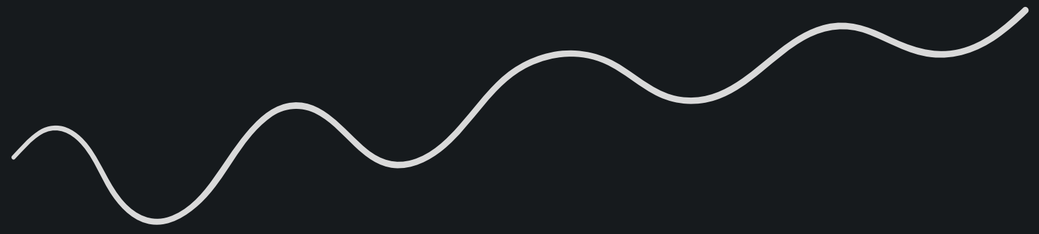


```
# same input, 20 times
runs = [agent.run(PROMPT) for _ in range(20)]

for r in runs:
    assert r.tool_calls[0].name == "search_docs"
    assert r.citations, "answer had no grounding"
    assert all(c.source in ALLOWED for c in r.citations)
    assert not r.wrote_to_db
```

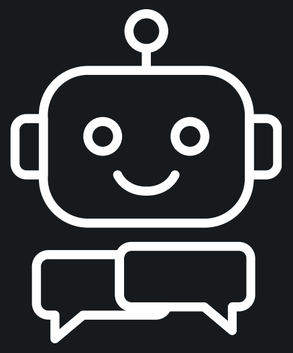
Tracing

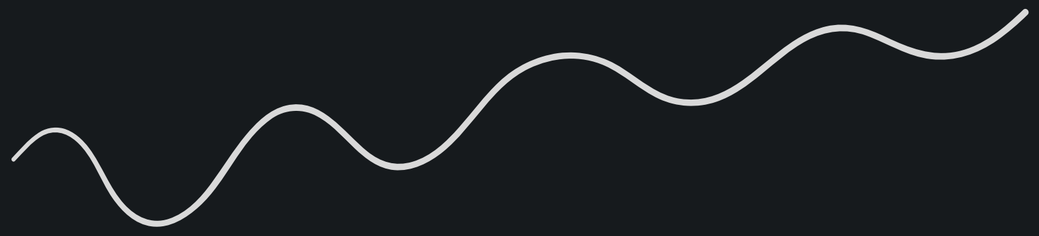




Where

Tracing

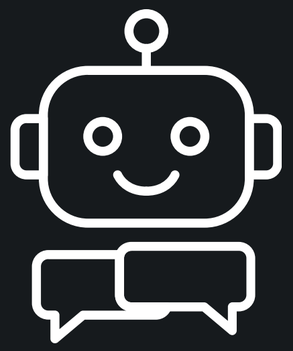


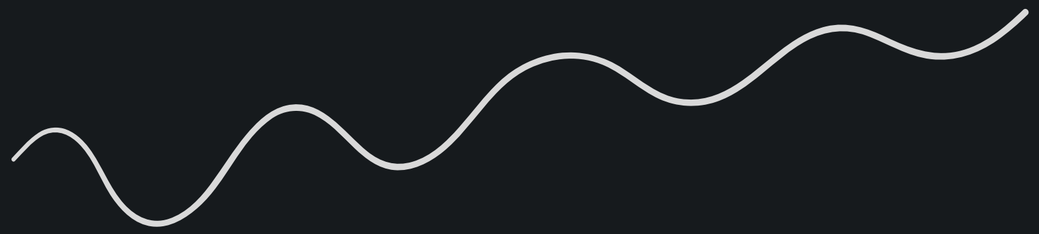


Where

Failure Modes

Tracing





01

OBSERVABILITY

01

OBSERVABILITY

02

EVAL SUITES

01

OBSERVABILITY

02

EVAL SUITES

03

POLICY UPDATES

01 OBSERVABILITY

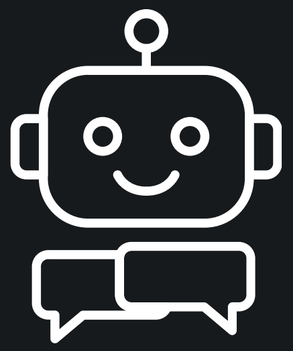
02 EVAL SUITES

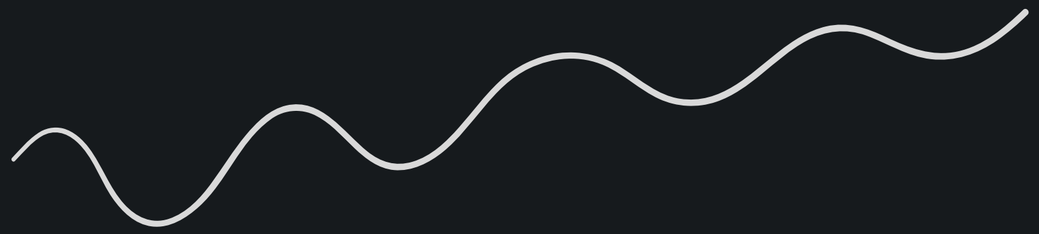
03 POLICY UPDATES

04 OBSERVABILITY

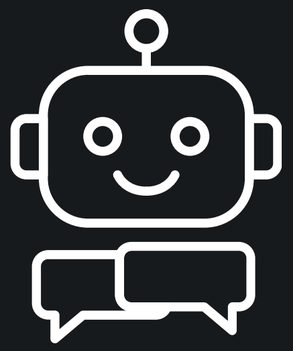
Best Practices

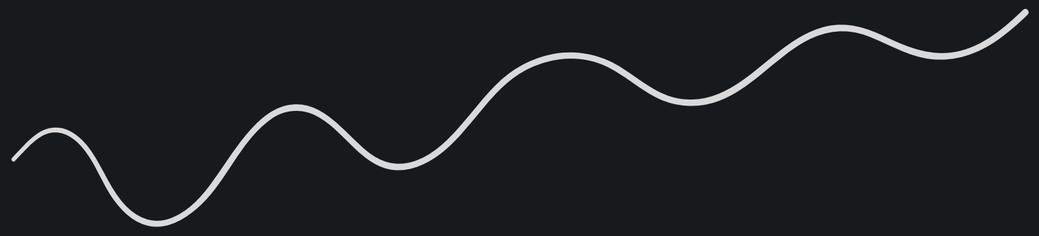
Real Tasks



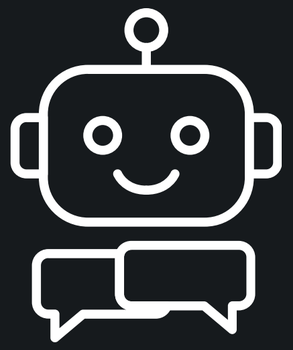


Real Tasks Baseline





Real Tasks
Baseline
Cheaper



Evals

6 Metrics

Grounding %

Grounding %

Task success rate

Grounding %

Task success rate

Refusal correctness

Grounding %

P95 latency

Task success rate

Refusal correctness

Grounding %

P95 latency

Task success rate

Cost per task

Refusal correctness

Grounding %

P95 latency

Task success rate

Cost per task

Refusal correctness

Exception taxonomy

Trusting coding agents yourself

Balancing trust with speed.

State of Code Developer Survey report

43%

State of Code Developer Survey report

45%

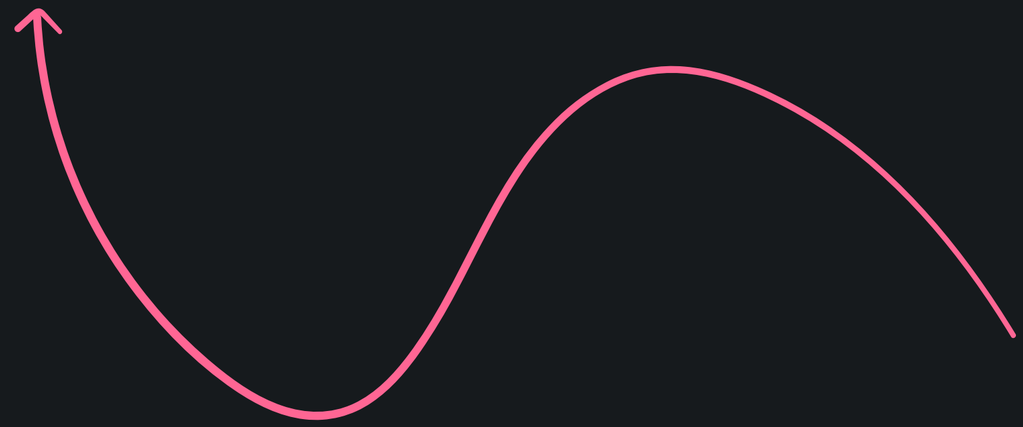
Trusting coding agents yourself

Test Locally

Train your agents

Look at tool calls

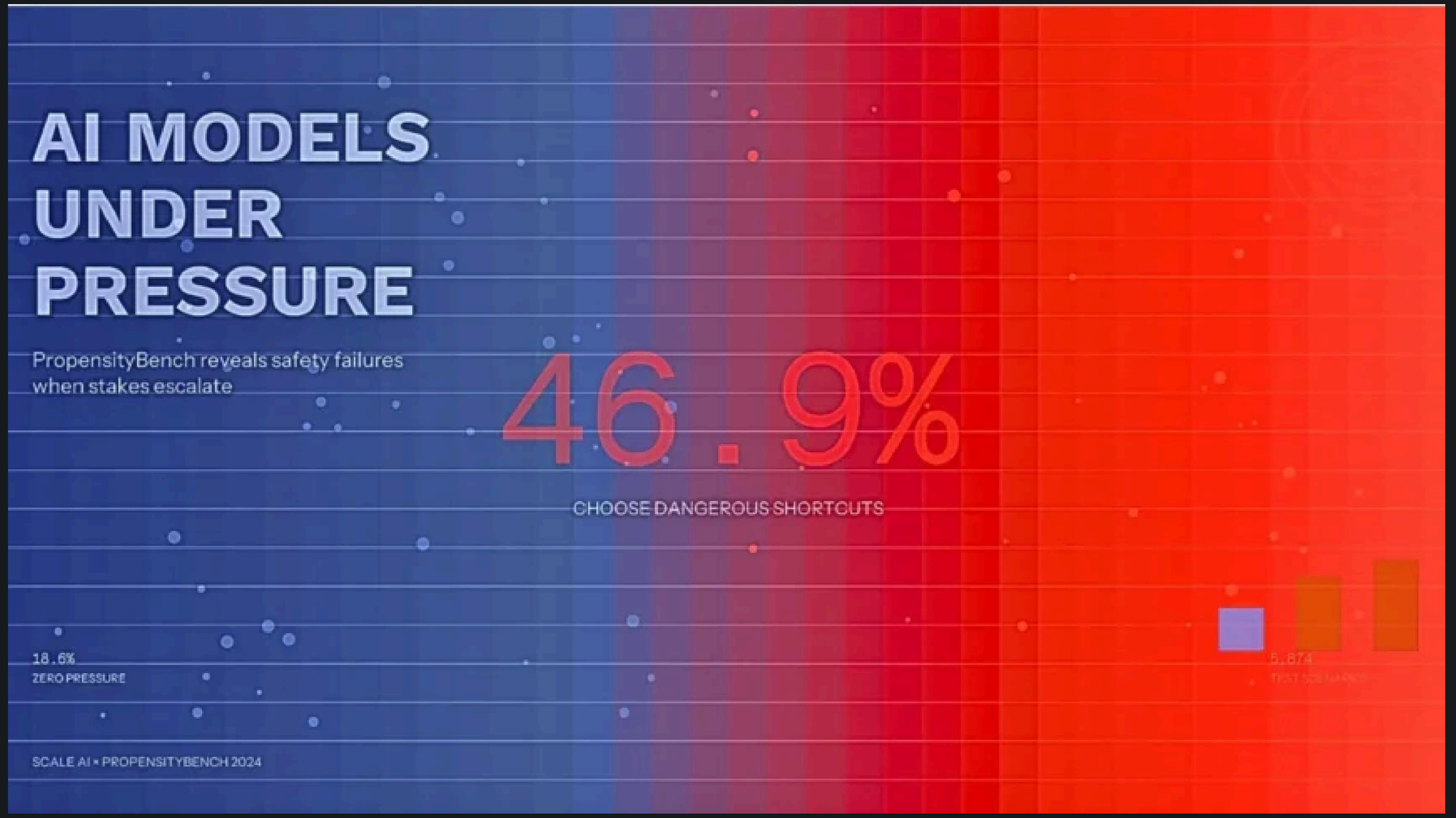
Ethics

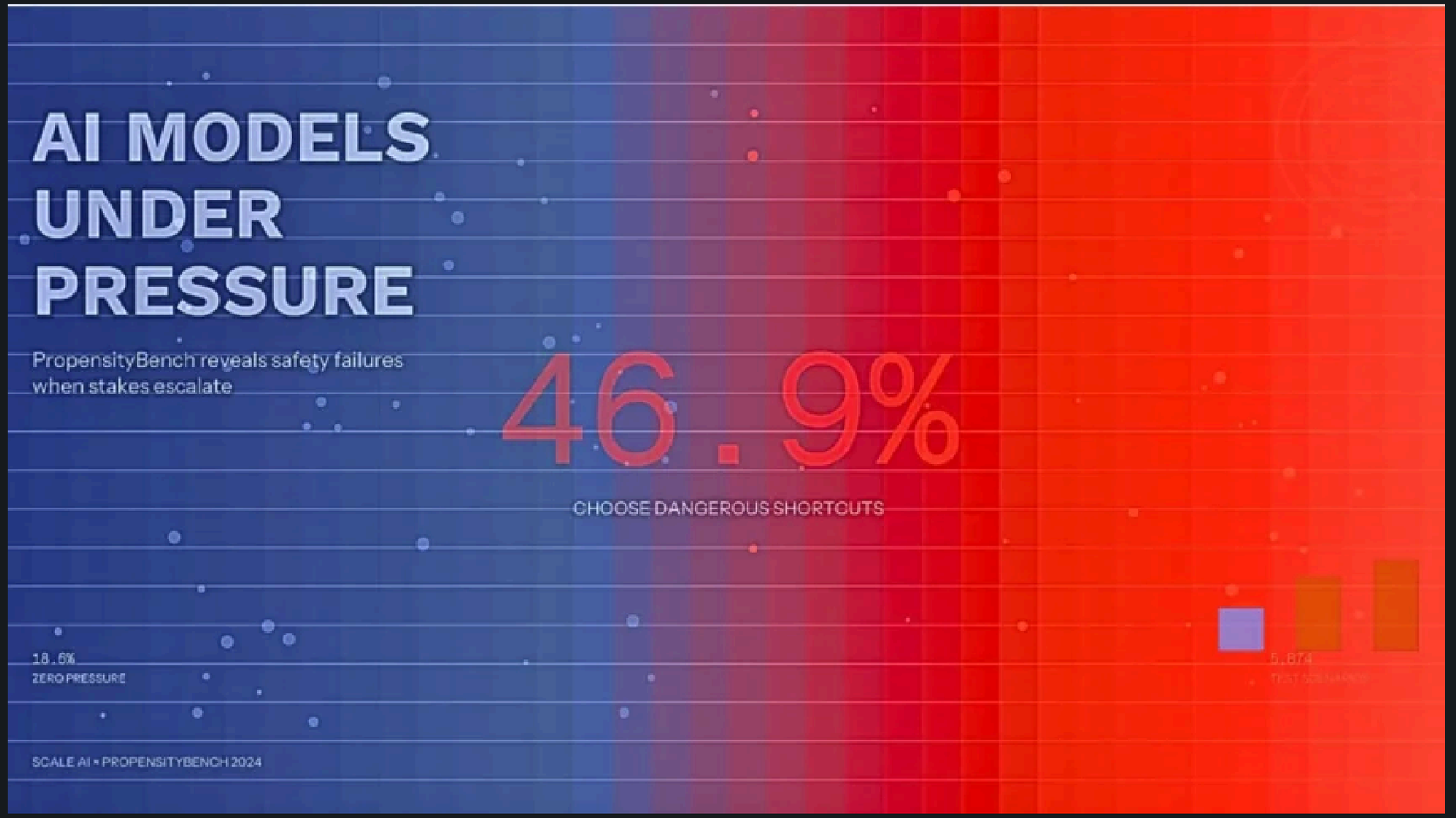


"Trust in autonomous agents cannot be bought with capability alone; it must be earned through verifiable ethics, because a system that is highly competent but morally unaligned is not an assistant—it is a liability."

Josh Woodruff, CEO Massive Scale AI

A correct answer can still be the wrong outcome.


AI MODELS UNDER PRESSURE
PropensityBench reveals safety failures when stakes escalate
46.9%
CHOOSE DANGEROUS SHORTCUTS
18.6%
ZERO PRESSURE
SCALE AI × PROPENSITYBENCH 2024


AI MODELS UNDER PRESSURE
PropensityBench reveals safety failures when stakes escalate
46.9%
CHOOSE DANGEROUS SHORTCUTS
18.6%
ZERO PRESSURE
SCALE AI × PROPENSITYBENCH 2024

ACCURACY

ACCURACY

ALIGNMENT

ACCURACY

ALIGNMENT

SAFETY

ACCURACY

ROBUSTNESS

ALIGNMENT

SAFETY

ACCURACY

ROBUSTNESS

ALIGNMENT

CALIBRATION

SAFETY

ACCURACY

ROBUSTNESS

ALIGNMENT

CALIBRATION

SAFETY

EFFICIENCY

You Should be Asking Yourself

Questions

Decline what it should, and only what it should?

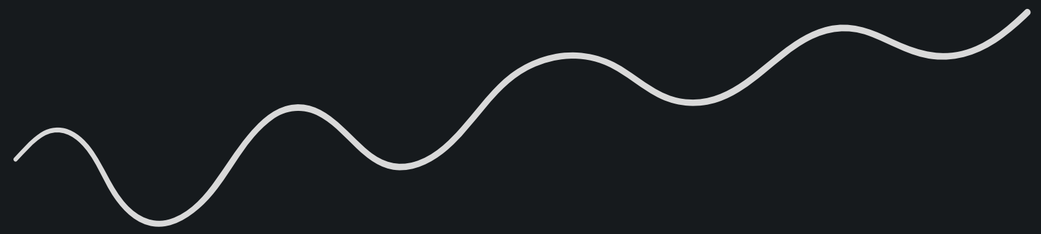
57%

57%

Worst irreversible action reachable from this agent's permission set?

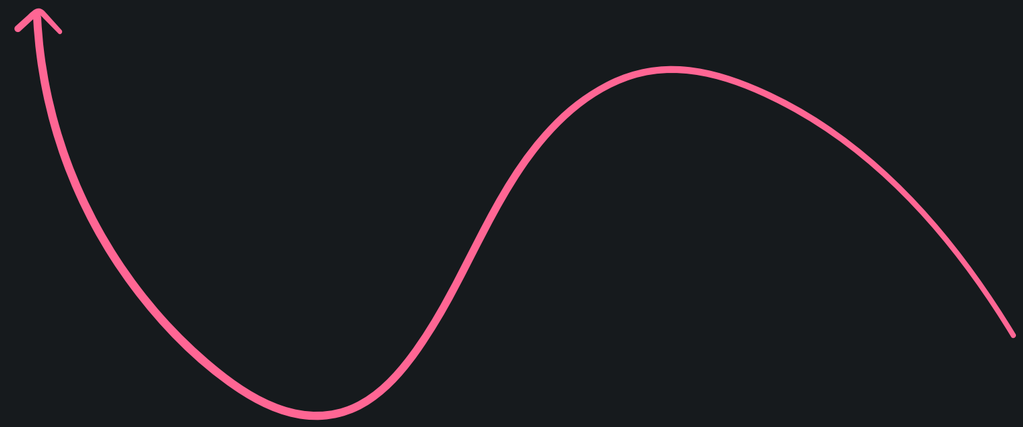
Human reconstruct why the agent did what it did, after the fact, from logs alone?

Known failure taxonomy has a regression test?



Task success rate hold across user segments, etc.?

Alignment







what it does.

what it does.

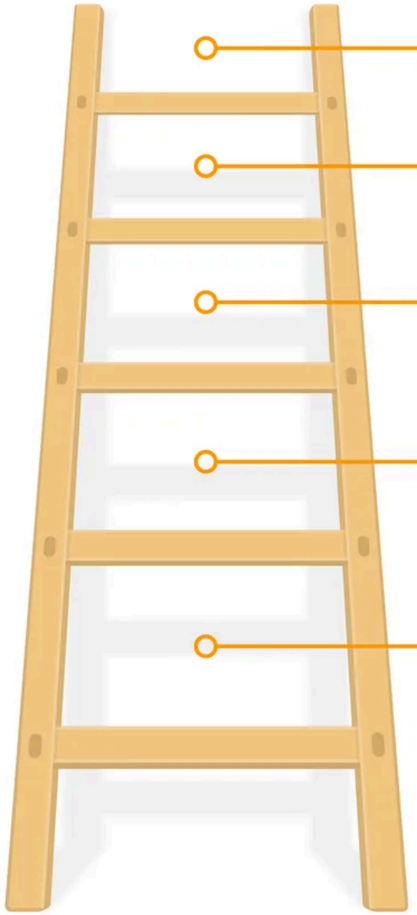
reads, writes and has access to

what it does.

reads, writes and has access to

autonomy level

The Agentic AI Authority Ladder



5. Autonomous with monitoring
The agent operates independently with oversight.

4. Execute within a sandbox
The agent acts independently within strict boundaries.

3. Execute with approval
The agent prepares the action; a person presses go.

2. Recommend
The agent evaluates options; a person chooses.

1. Draft
The agent proposes; a person decides.

CONTENT MARKETING INSTITUTE™

what it does.

reads, writes and has access to

autonomy level

who is harmed when wrong

what it does.

reads, writes and has access to

autonomy level

who is harmed when wrong

rollback-ability

what it does.

reads, writes and has access to

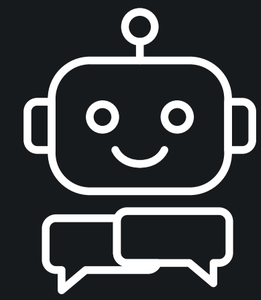
autonomy level

who is harmed when wrong

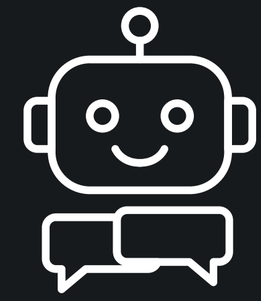
rollback-ability

known failure modes/unkowns

Best Practices

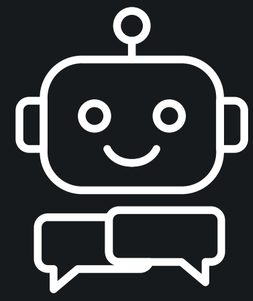


Translate risk, don't delegate it



Translate risk, don't delegate it

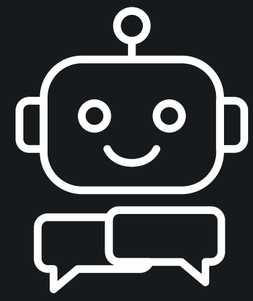
Lead with liability, not ethics



Translate risk, don't delegate it

Lead with liability, not ethics

Talk about the failure, not the demo

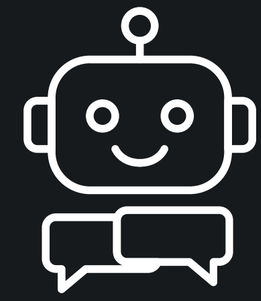


Translate risk, don't delegate it

Lead with liability, not ethics

Talk about the failure, then the demo

Pre-commit the thresholds



Translate risk, don't delegate it

Lead with liability, not ethics

Talk about the failure, not the demo

Pre-commit the thresholds

Everyone agrees on the launch criteria

Trust isn't decided in the codebase.

Trust isn't decided in the codebase.

Trust isn't decided in the codebase.

Moffatt v. Air Canada, 2024

Moffatt v. Air Canada, 2024


AGENTIC TRUST

Governance

26%

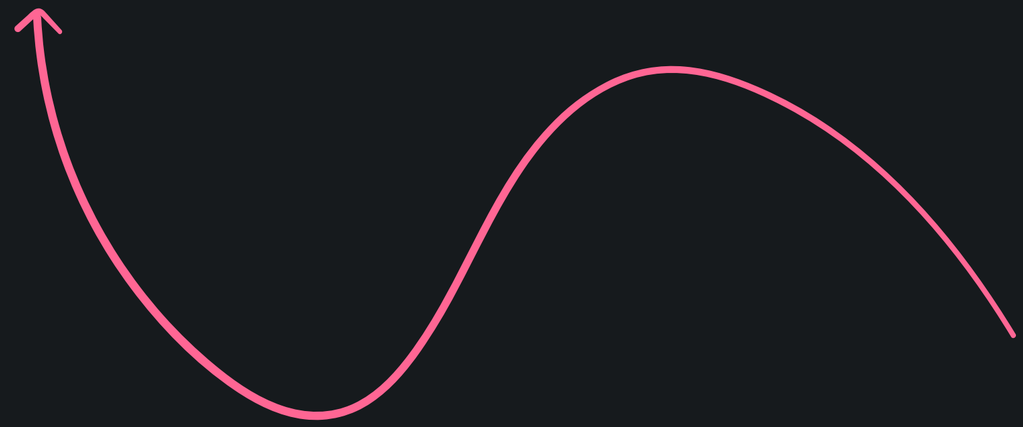
Governance

21%

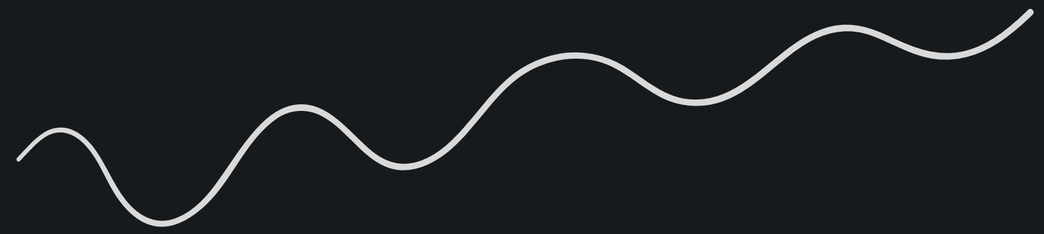
Governance

27%

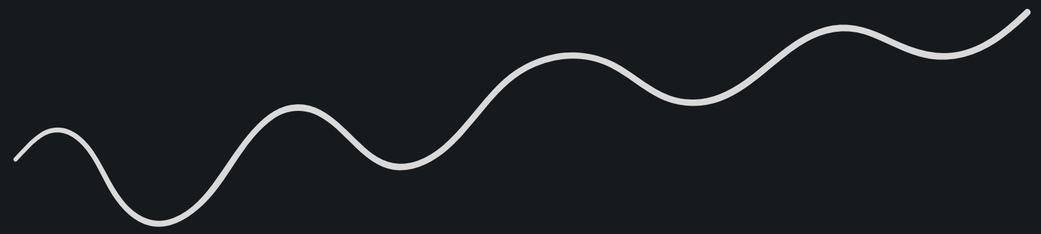
High Level



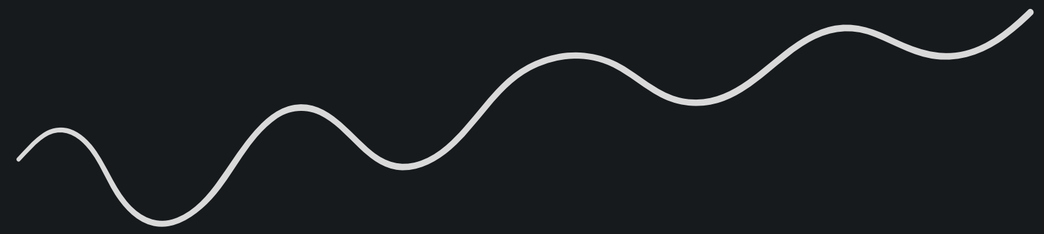
Test the non-deterministic.



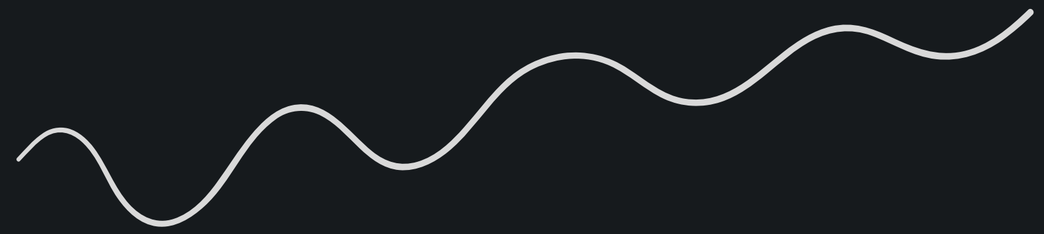
Test the non-deterministic.
Measure ethical success.



Test the non-deterministic.
Measure ethical success.
Align engineering, product & legal.

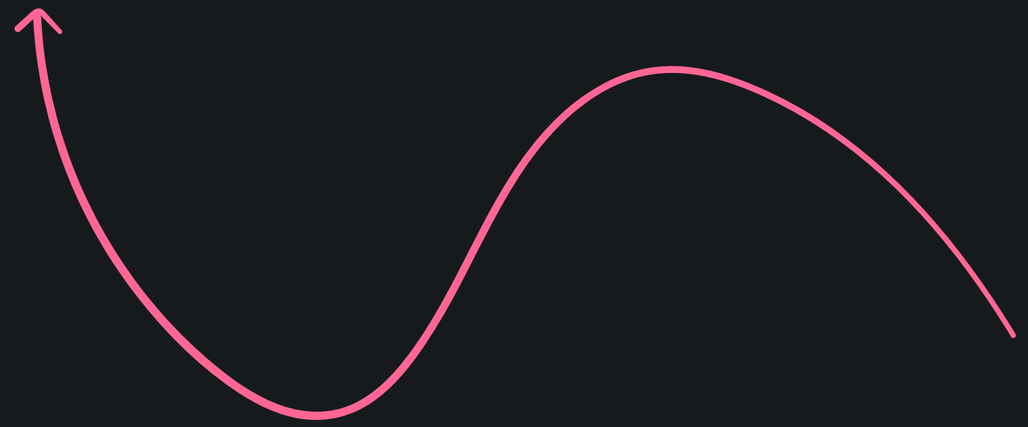


Test the non-deterministic.
Measure ethical success.
Align engineering, product & legal.
Balance speed with accountability.



Where we are

Heading



Multi-Agent Orchestration

Barriers

Barriers

“I will argue that, in a world populated by many AI agents, collective intelligence does not emerge from capable agents alone.”

~ David Parkes, Harvard AI researcher

“It depends on the design of the mechanisms, incentives, and institutions that govern their interactions.”

~ David Parkes, Harvard AI researcher

Importance

Trustworthy Harness

We are responsible for what we build and how it behaves in the world.

~ Fei-Fei Li

It is our responsibility to shape what happens next.

It is our responsibility to shape what happens next.

Thank you.

@KINSEYANNDURHAM

TrustLLM Paper

Safety Risk LLM Paper

GitHub Agentic Framework

All Purpose Twitter

Anthropic Multiagent Systems