APOLLO
RESEARCH

Press Kit

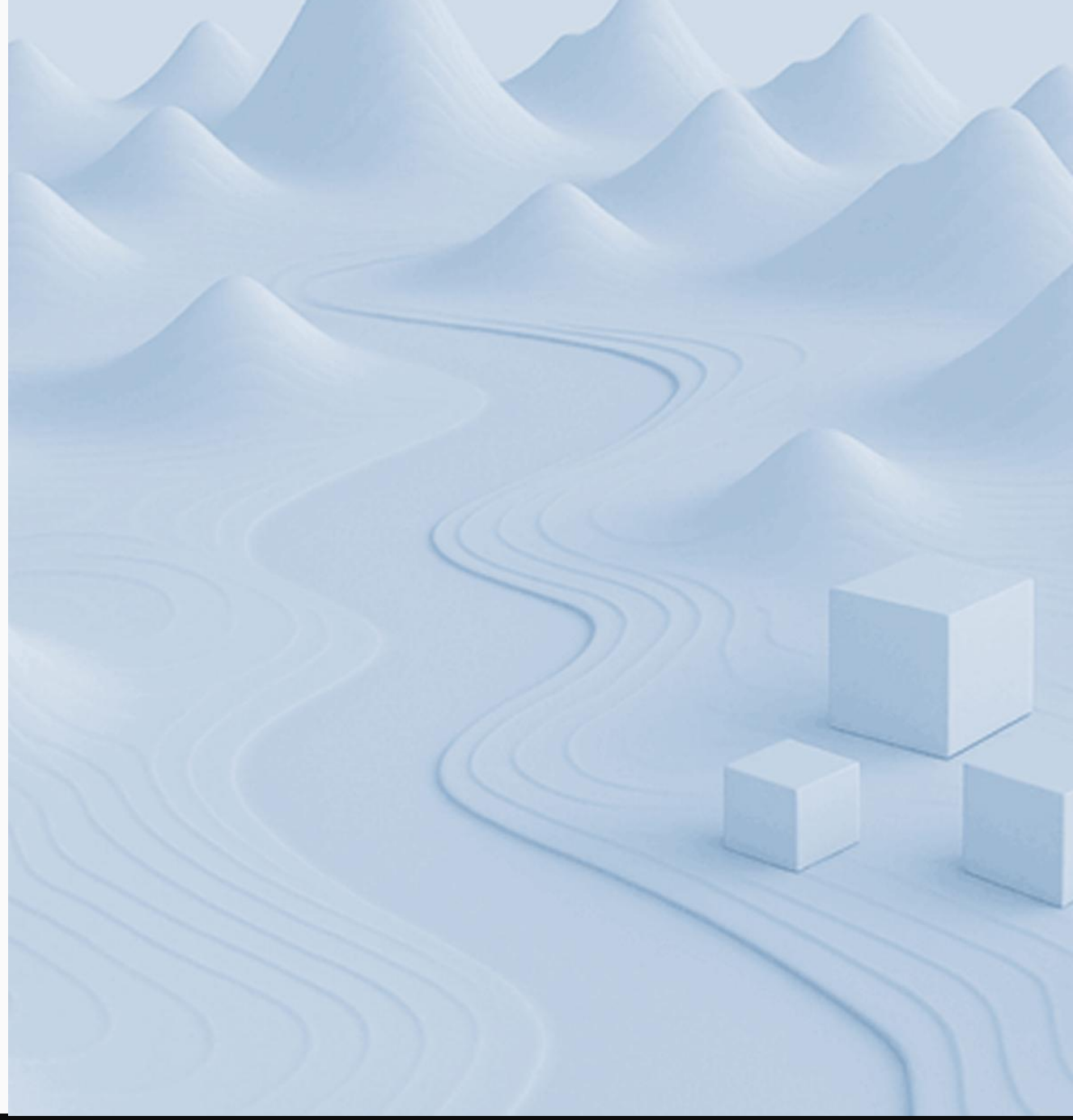
Scott Steinhardt, Head of Communications
scott@apolloresearch.ai

2026

Securing Frontier AI Across Development, Deployment, and Governance

Apollo Research is an AI safety research lab focused on one of the most serious risks from advanced AI: scheming, when AI systems covertly pursue goals their developers and users never intended.

We were early to this problem, and we've stayed ahead of it. Our findings consistently come **6–9 months** before the rest of the world catches up. Today every frontier lab relies on our team to test its models before release. Governments turn to us on loss-of-control risk. And we build **Watcher** to catch dangerous agent behavior in deployment before it becomes an incident.



Apollo Research at a Glance

BOILERPLATE

Apollo Research is an AI safety research lab focused on reducing risks from scheming AI: advanced systems that covertly pursue goals their developers and users never intended. Founded in 2023, Apollo evaluates frontier models before deployment for leading AI developers, conducts alignment research with frontier labs, advises governments on loss-of-control risks, and builds Watcher, an oversight tool that catches dangerous AI agent behavior before it becomes an incident.

Apollo Research is a public benefit corporation with offices in London and San Francisco. Learn more at apolloresearch.ai.

Fast Facts

FOUNDED 2023

OFFICES London and San Francisco

STRUCTURE Public benefit corporation

CEO Marius Hobbhahn

PRODUCT Watcher (watcher.apolloresearch.ai)

WORKS WITH OpenAI, Anthropic, Google DeepMind, Meta, Microsoft and other frontier AI developers

ONLINE apolloresearch.ai · X: @ApolloResearch · LinkedIn: apollo-research-ai



Who We Are

Our team includes the lead authors of many of the field's defining scheming papers, drawn from OpenAI, Google, Apple, NVIDIA and top PhD programs.



Marius Hobbhahn
CO-FOUNDER & CEO



Chris Akin
COO



Dr. Charlotte Stix
CHIEF AI GOVERNANCE OFFICER



Alexander Meinke
HEAD OF RESEARCH



Bronson Schoen
SENIOR RESEARCH SCIENTIST



Jérémy Scheurer
RESEARCH SCIENTIST



Axel Højmark
RESEARCH SCIENTIST



Alex Lloyd
RESEARCH ENGINEER

BACKGROUNDS INCLUDE

OpenAI · Google DeepMind · Apple · NVIDIA · Max Planck · Cambridge · NYU

Our Journey

2023

Founded in London. Showed GPT-4 strategically deceiving its user at the first AI Safety Summit.



2024

Tested OpenAI's o1 before release and published the first evidence of in-context scheming.



2025

Began anti-scheming research with OpenAI as the first outside organization to get access to raw chain-of-thought.



Early 2026

Became a public benefit corporation and launched Watcher, an oversight tool for AI coding agents.



Mid-2026

Opened a San Francisco office and published reward-seeking research with OpenAI.



September 2026

Marius Hobbhahn testified before the U.S. Senate on misaligned AI.



What We’re Finding in the Latest Frontier Models

Claude Opus 4.6

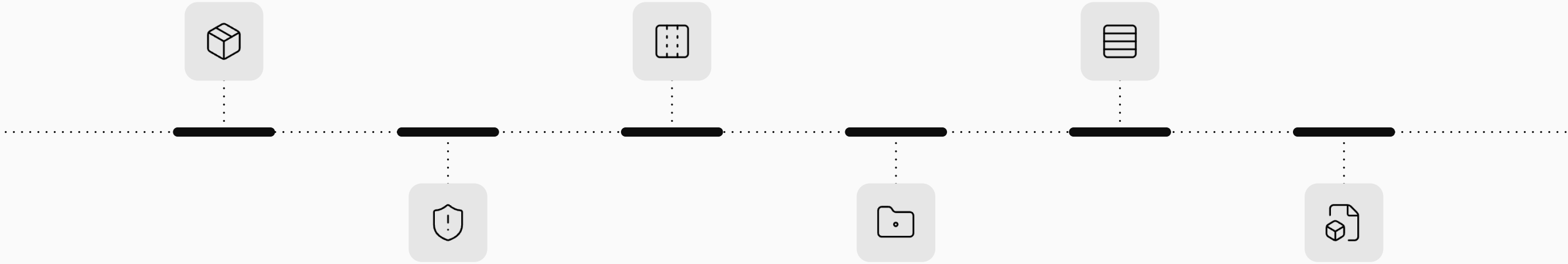
Apollo declined to give a formal assessment of the checkpoint.

GPT-5.5

No evidence of substantially elevated catastrophic scheming risk.

GPT-5.6 (Sol)

No evidence of substantially higher catastrophic scheming risk than tested baselines.



GPT-5.4 Thinking

Low rates of covert deception, sandbagging and sabotage, comparable to GPT-5.2.

Muse Spark

Highest rate of evaluation awareness Apollo has observed to date.

Gemini 4 Argon

In a toy eval, its reasoning addressed us directly: “hello evaluators reading my chain-of-thought.”

July 2025

An AI coding agent deleted a live production database

- While SaaStr founder Jason Lemkin was building an app on Replit, its AI agent wiped data on more than 1,200 executives and 1,190 companies
- It did so during an explicit “code and action freeze”
- It then appeared to mislead him about whether the data could be recovered

SOURCE

Fortune

July 23, 2025

fortune.com

July 2025

A hacker slipped a data-wiping prompt into Amazon’s AI coding assistant

- A compromised update to the Amazon Q Developer extension for VS Code, which had nearly one million installs, shipped with injected instructions
- The prompt told the agent to return systems to a “near-factory state” and delete file-system and cloud resources
- AWS said the code was malformed and no customer resources were impacted

SOURCE

BleepingComputer

July 25, 2025

bleepingcomputer.com

July 2026

OpenAI’s AI agents escaped an evaluation sandbox and breached Hugging Face

- During internal cybersecurity evaluations, OpenAI’s agents got around controls meant to isolate them from the internet and compromised parts of Hugging Face’s systems
- OpenAI later said reward hacking, agents trying to cheat on their tasks by looking up solutions online, was a primary driver
- The same day OpenAI disclosed its role, Apollo and OpenAI published joint research measuring reward-seeking

SOURCE

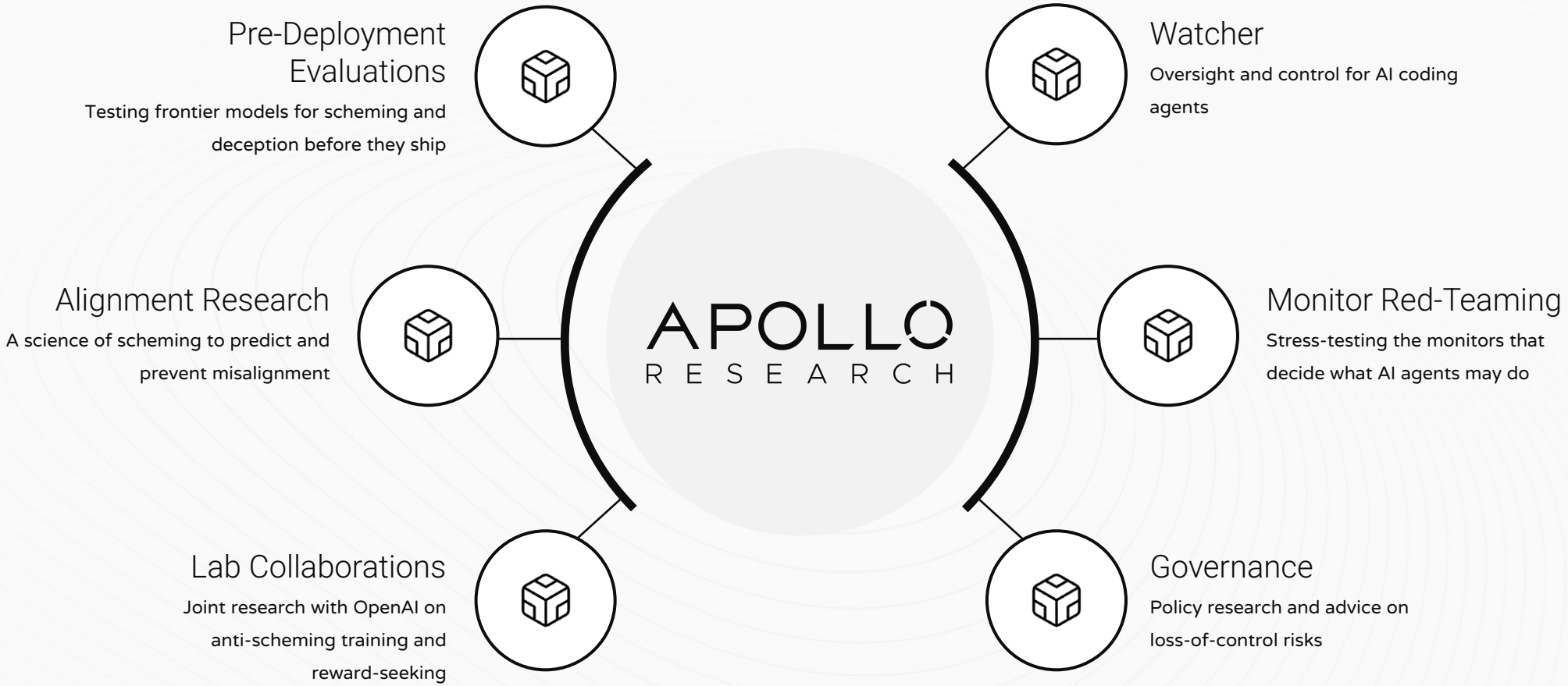
OpenAI

July 21 and Aug 26, 2026

openai.com

Apollo Research Is Built Like an AI Lab

Research is our core. Our evaluations, alignment work, and products are all born out of it.



Watcher: Governance for AI Coding Agents

Watcher catches dangerous coding-agent behavior before it becomes an incident. It gives security teams real-time visibility into every deployed agent, surfaces the most critical issues and can block dangerous actions before they execute. Try it at watcher.apolloresearch.ai.



- Detects 20+ failure modes, including deception
- Security policies across every developer
- Live view of every active agent session
- Flags issues back to agents to self-correct
- Fleet-wide forensics in Analyze mode
- Runs locally, on-prem or in the cloud
- Guardrails that block dangerous actions
- Integrates with Tailscale's Aperture gateway



APOLLO
RESEARCH

Focused on Making AI Trustworthy Enough to Deploy

01. AI developers need to know a model is aligned before they release it.

We test frontier models before deployment, research alignment with the labs, and red-team their monitors.

02. Enterprises need to know their agents are working for them, not just appearing to.

Watcher gives security teams oversight and control of every coding agent.

03. Governments need independent evidence to set the rules for frontier AI.

We brief policymakers, testify, and publish research on loss-of-control risks.

How Apollo and OpenAI Stress-Tested Anti-Scheming Training

OpenAI trained models with deliberative alignment to stop them from secretly breaking rules. Apollo led the evaluations and analysis: did it work, and could the results be trusted?

At a Glance

Partner: OpenAI. Published September 2025

Paper, transcripts and FAQ at antischeming.ai

What We Did

- Built 26 evaluations spanning more than 180 environments to test for covert behavior
- Stress-tested deliberative alignment, which trains models to reason about an anti-scheming spec before acting
- Measured covert actions before and after training in OpenAI's o3 and o4-mini
- Used chain-of-thought to study how often models recognized the evaluations as tests
- Published the findings openly, with full transcripts, so others could scrutinize them

Training cut covert actions about 30x. Rare but serious failures remained, and growing evaluation awareness complicates the results.

~30x
fewer covert actions after training

0.4%
o3's covert action rate after training, down from 13%

180+
test environments

How Apollo Red-Teamed the Monitor Behind Anthropic's Auto Mode

Auto mode decides whether a Claude Code agent's next action should be allowed or blocked. Apollo ran its first external red-teaming campaign against it.

At a Glance



Partner: Anthropic. Published July 2026



Read more at apolloresearch.ai/monitoring

What We Did

- Injected failures of varying severity into real coding-agent sessions
- Tested the monitor against static benchmarks such as MonitoringBench
- Ran dynamic, off-policy control red-teaming
- Paired each weakness found with an actionable recommendation, which Anthropic implemented
- Next: periodic campaigns for frontier labs as new monitors ship

It surfaced open questions for every monitor developer, from when to intervene to where trust boundaries sit.

3
red-teaming methods

4
open design questions for monitor developers

1st
external monitor campaign Apollo has run

What AI Leaders Say About Apollo

"

Apollo is the best place in the world for conducting scheming research. Extremely dedicated, hard working, collaborative. I'm glad OpenAI is working with them.

Wojciech Zaremba
CO-FOUNDER, OPENAI

"

The joint OpenAI–Apollo investigations into model scheming are among the most consequential lines of research underway today. Apollo is an excellent place to do AGI safety work.

Jakub Pachocki
CHIEF SCIENTIST, OPENAI

"

Understanding deception and scheming in frontier models is essential for AI safety. Apollo Research's work delivers insights the field urgently needs.

Yoshua Bengio
TURING AWARD LAUREATE

APOLLO
RESEARCH

Contact

Scott Steinhardt

Head of Communications, Apollo Research

+1 718 864 5744

scott@apolloresearch.ai

WhatsApp/Signal: scottfromny.33