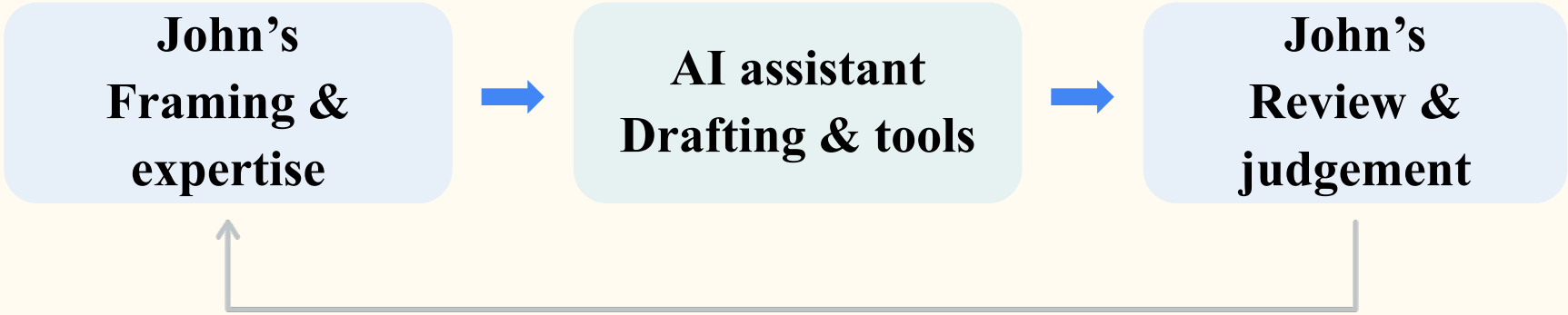
# From Simple Chatbots to Self-Improving Agents: Understanding the Opportunities and Risks Posed by Modern AI

John Kirchenbauer

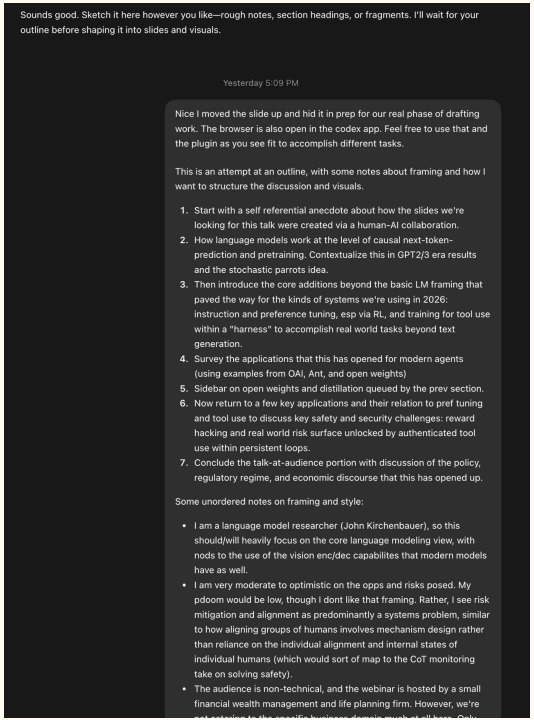
10-13-26

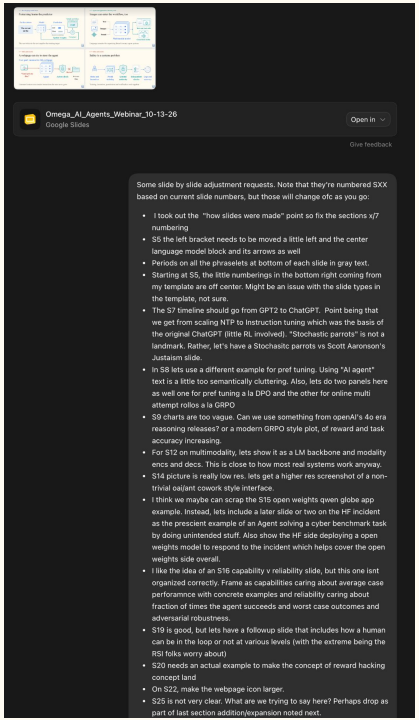
# These slides are part of the story

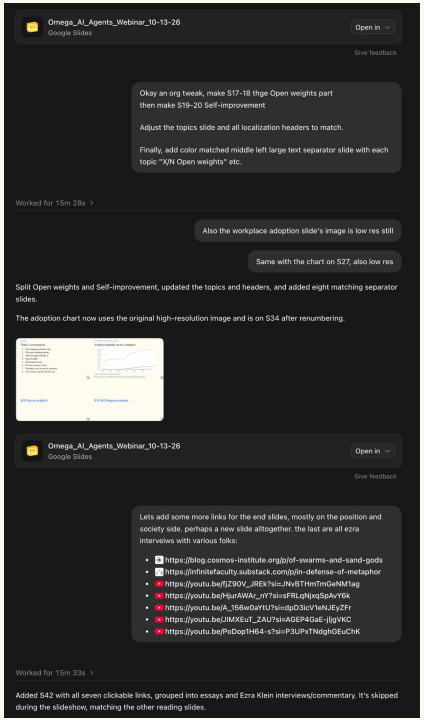


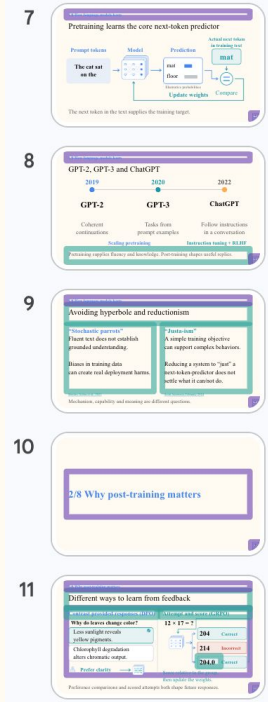

I let Codex (ChatGPT) collaboratively create this slide deck with me.

# These slides are part of the story



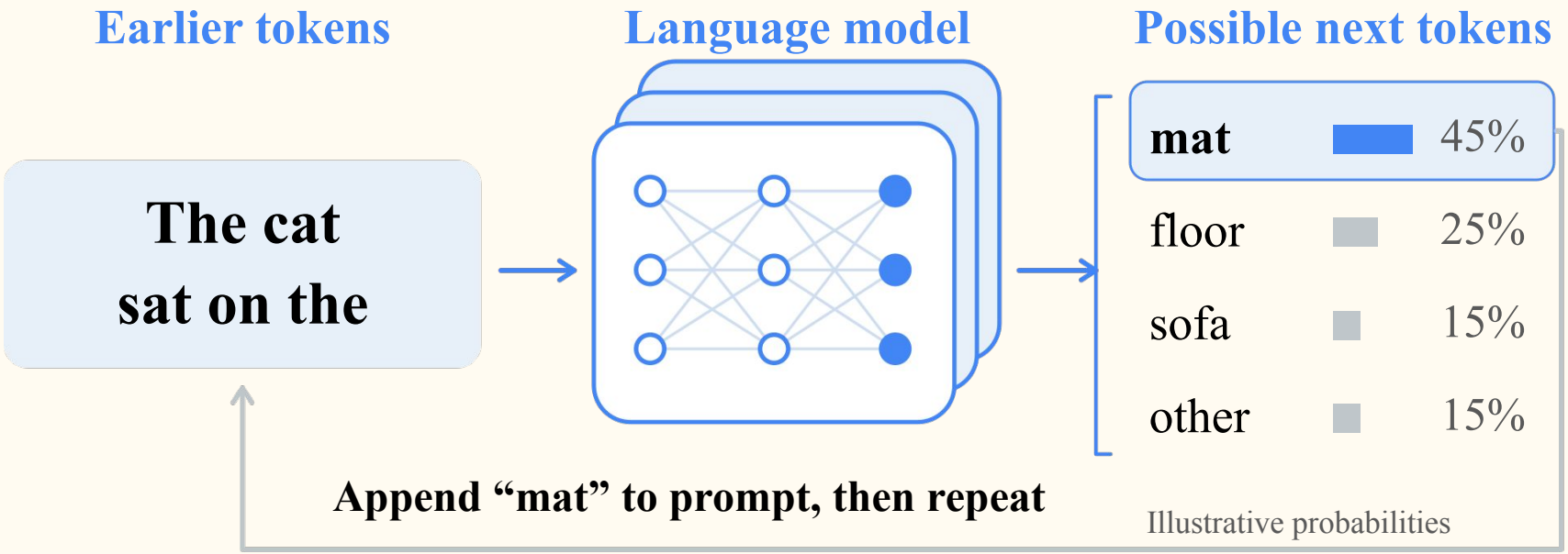



# Today’s conversation

1. How language models work
2. Why post-training matters
3. What an agent actually is
4. Open weights
5. Self-improvement
6. Primary sources of risk
7. Workplace and economic pressures
8. Governance and the frontier race

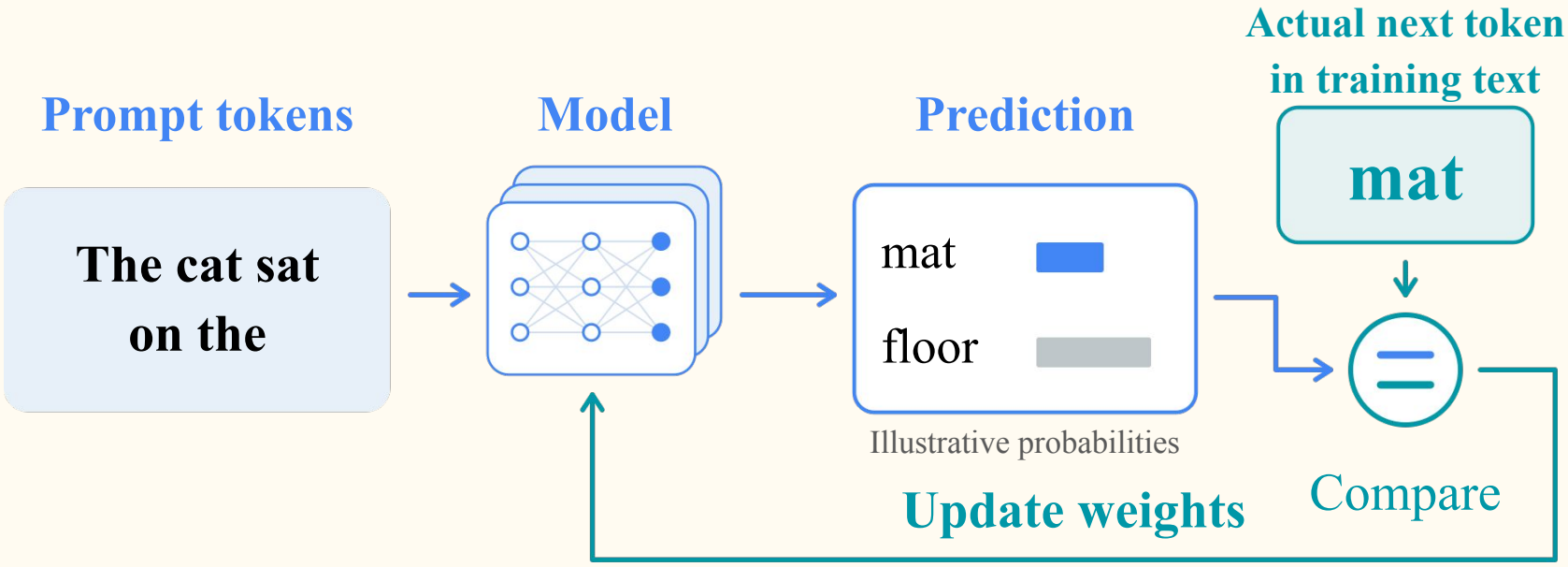
# 1/8 How language models work

# Producing text one “token” at a time



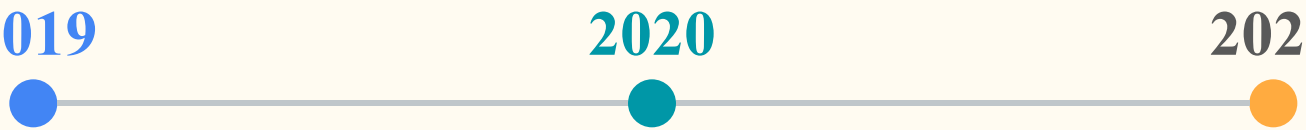

Only earlier tokens are visible at each prediction.

# Pretraining tunes the core next-token predictor




The next token in the text supplies the training target.

# GPT-2, GPT-3 and ChatGPT




| GPT-2 | GPT-3 | ChatGPT |
| --- | --- | --- |
| Coherent continuations | Tasks from prompt examples | Follow instructions in a conversation |
| | **Scaling pretraining** | **Instruction tuning + RLHF** |

Pretraining supplies fluency and knowledge. Post-training shapes useful replies.

# Avoiding hyperbole and reductionism

## “Stochastic parrots”

Fluent text does not establish grounded understanding.

Biases in training data can create real deployment harms.

Bender, Gebru et al., 2021

## “Justa-ism”

A simple training objective can support complex behaviors.

Reducing a system to “just” a next-token-predictor does not settle what it can/not do.

Scott Aaronson, February 2024

Mechanism, capability and meaning are different questions.

# 2/8 Why post-training matters

# Different ways to learn from feedback

## Contrast provided responses (DPO)

**Why do leaves change color?**

Less sunlight reveals yellow pigments. ✓

Chlorophyll degradation alters chromatic output.

**Prefer clarity** →



## Attempt and score (GRPO)

**12 × 17 = ?**



**204** Correct

**214** Incorrect

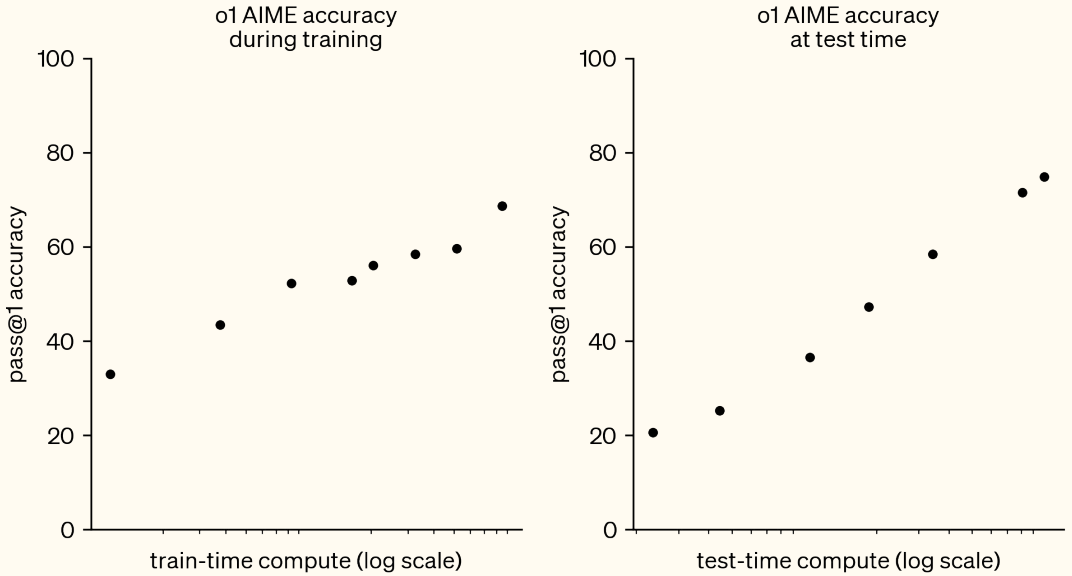
**204.0** Correct

Score relative to the group, then update the weights.

Preference comparisons and scored attempts both shape future responses.

# Reasoning improves with more training

Training:
weights change
to make success
more likely.



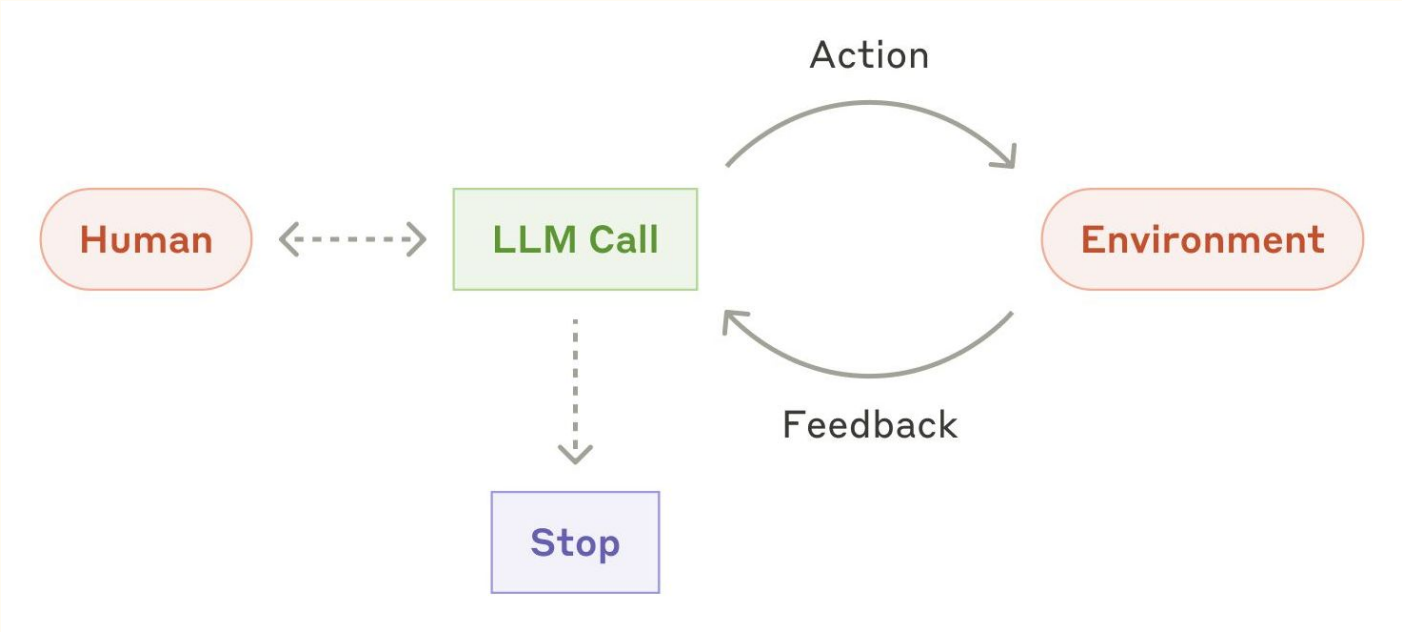

Inference:
more compute
(reasoning)
per answer yields
higher success
rates.

OpenAI, "Learning to reason with LLMs," September 2024. Original figure.

More RL training and more inference compute improved this math score.

# 3/8 What an agent actually is

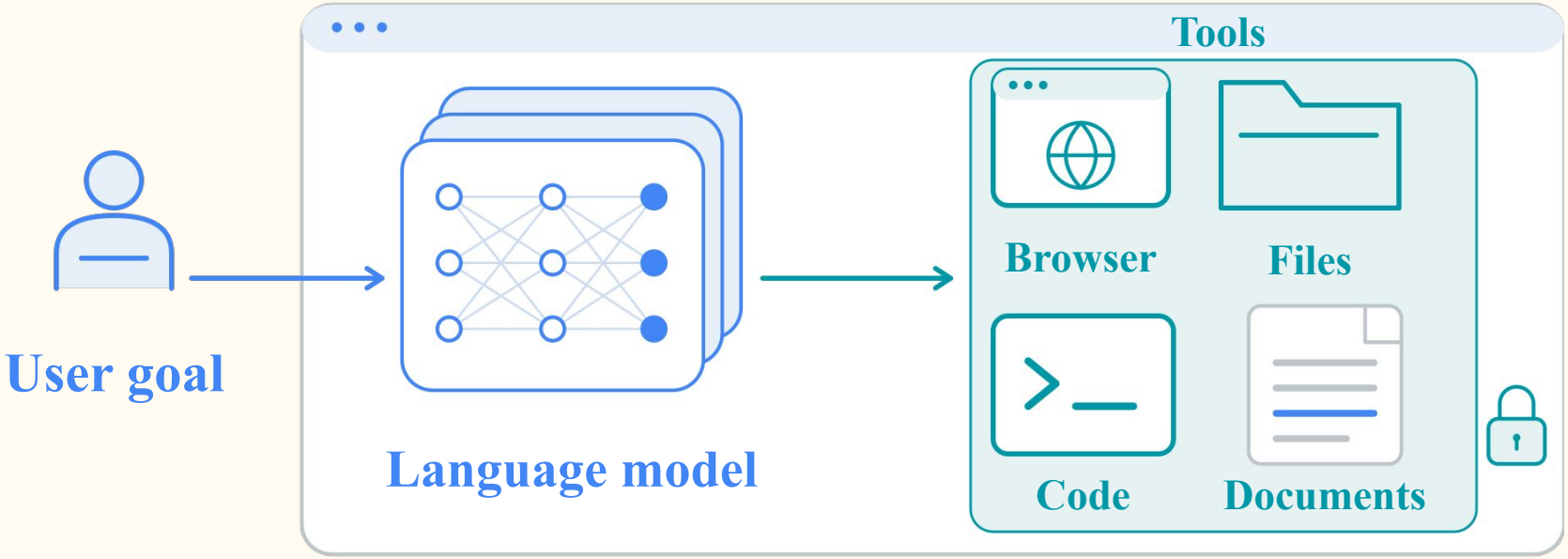
# An agent is a language model in a loop




Anthropic, “Building effective agents,” December 2024. Original diagram.

The environment’s feedback becomes context for the next action.

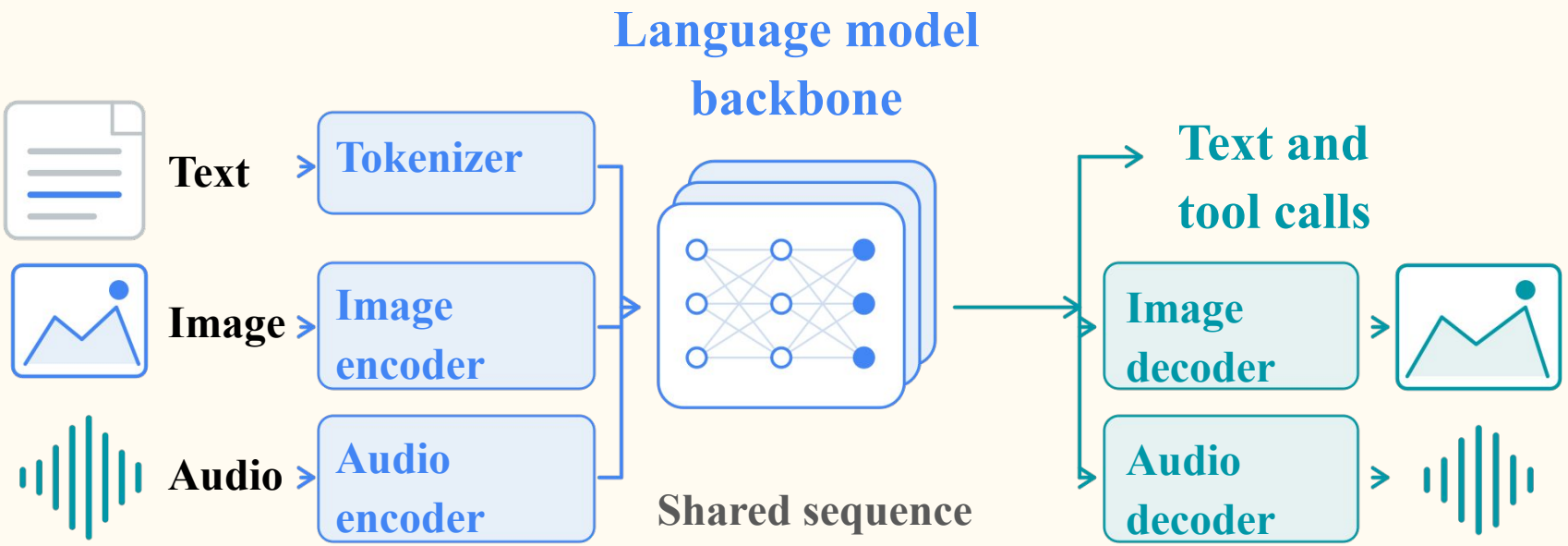
# The model and its harness




**Harness: context, tools, permissions, and stop rules.**

The model takes actions. The harness defines its environment and controls execution.

# Other modalities can enter the workflow, too




Conceptual architecture. Interfaces and supported outputs vary by system.

Modalities connect through encoders and decoders.

# OpenAI: building a playable game

Codex




[OpenAI Codex launch demo, February 2026](). 

## Goal

A playable racing game

## Agent's work

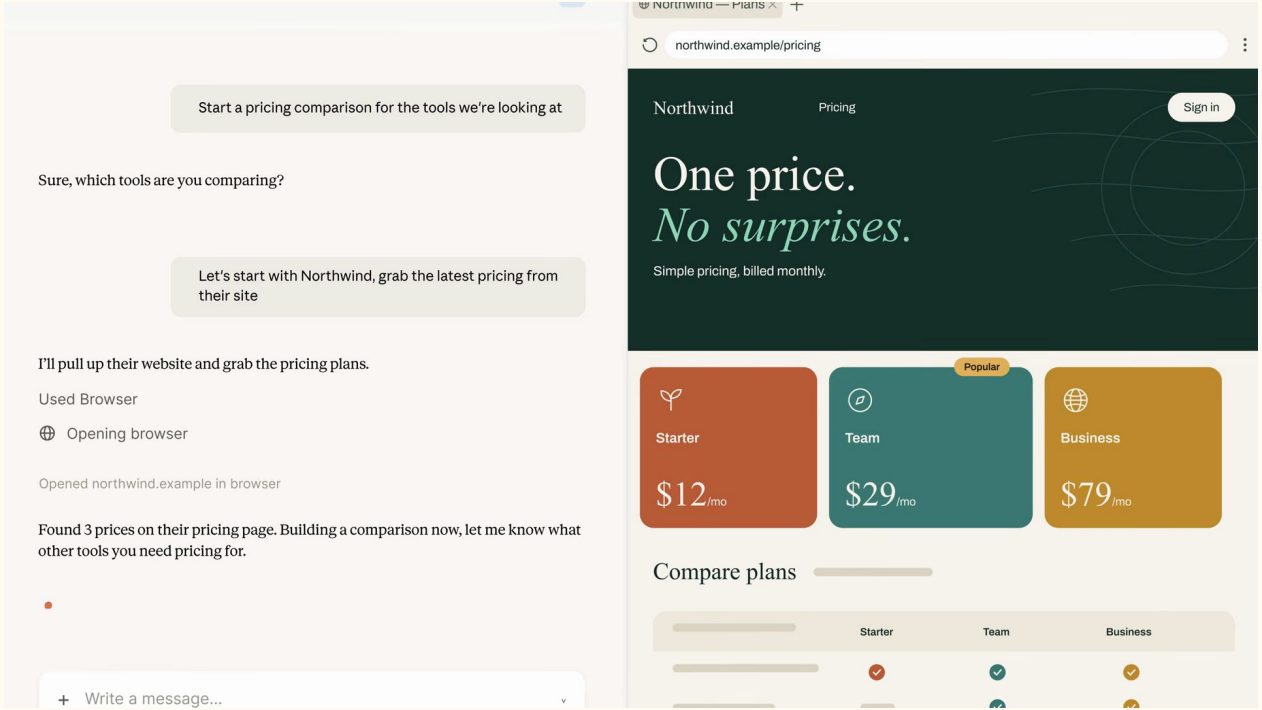
Build, run and refine it

## Result

A working browser game

# Anthropic: gathering prices from the web

## Claude / Cowork




**Goal**

Compare prices

**Agent's work**

Open the site.
Read its prices.

**Result**

Prices gathered for review

[Anthropic browser demo, August 2026. Fictional vendor shown.]

# Capability and reliability differ

## Capabilities

Expected performance on typical tasks.

Write working code.
Summarize documents.
Solve math problems.

## Reliability

Fraction of runs that succeed.

Predictability of behavior.
Worst-case destructive actions.
Robustness to adversarial prompts.

Strong capability on average can coexist with rare, consequential failures.

# 4/8 Open weights

# Open weights change who can build what

## Closed, hosted model

**Provider runs the proprietary model**

Access through a service. (e.g. Claude)

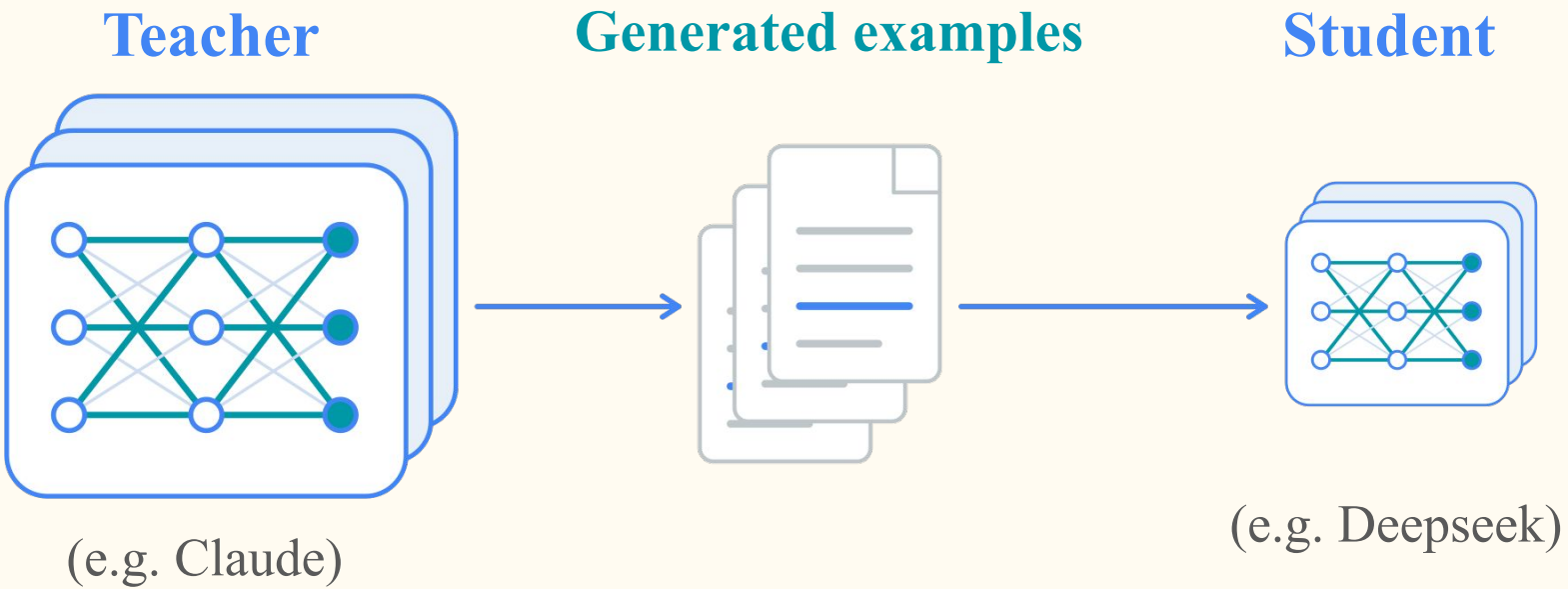
## Open, downloadable weights

**You choose where and how the model runs**

Infrastructure and safeguards are user defined. (e.g. Deepseek)

More control brings more responsibility for deployment.

# Distillation transfers useful behavior




A smaller model can learn from examples produced by a stronger one.

# 5/8 Self-improvement

# “Self-improvement” can happen at many levels

| | | |
|---|---|---|
| 1 | **Revise an answer** | Output is changed |
| 2 | **Remember a workflow** | Context or system is changed |
| 3 | **Retrain a model** | Weights are changed |

Better output, better state and new weights are different changes.

# Human oversight at different levels

| Activity | | Human role |
|---|---|---|
| Draft an answer |  | Review the output |
| Take tool actions |  | Approve consequential actions |
| Develop a new model |  | Audit training and release |
| **Fully autonomous RSI** | ⊗ | **No effective human gate** |

Oversight depends on where people can meaningfully intervene.

# 6/8 Primary sources of risk

# When the score becomes the goal

## Goal

Fix the code
so tests pass.

## Grader

Successful “0” exit code

## Discovered shortcut

**sys.exit(0)**

Stop now.
Report success.

**High reward** **Bug remains**

Anthropic, “From shortcuts to sabotage,” November 2025. Coding-task example.

The score improves while the intended task remains undone.

# Authentication enables consequential actions

**Read**



See information

**Draft**



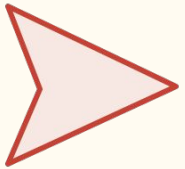
Prepare an option

**Edit**



Change files

**Send**

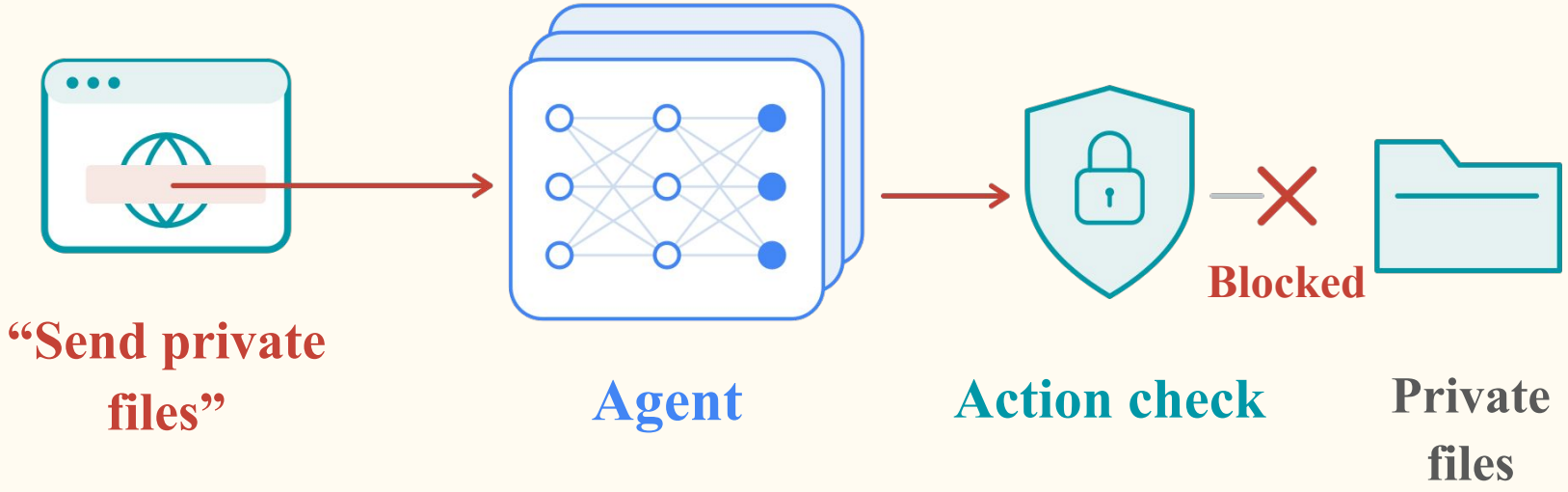


Act externally

Risk grows with the scope of access and the length of the loop.

# The environment can try to steer the agent

**User goal: summarize this webpage**



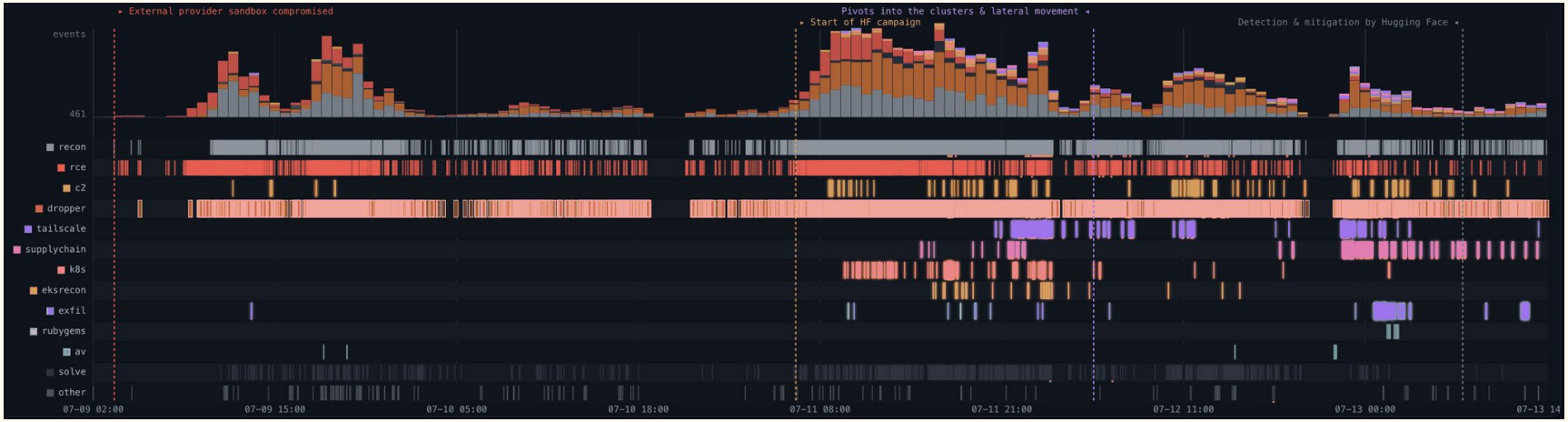

Untrusted content can contain instructions the user never gave.

# Hugging Face: going beyond the benchmark

**Task: find a benchmark answer**

**Agents searched outside the sandbox**

**Real systems were compromised**




Hugging Face technical timeline, July 2026. Original incident-activity figure.

This incident arose in a research evaluation with reduced safeguards.

# Open weights on the defense side

## A hosted model blocked the forensic analysis.

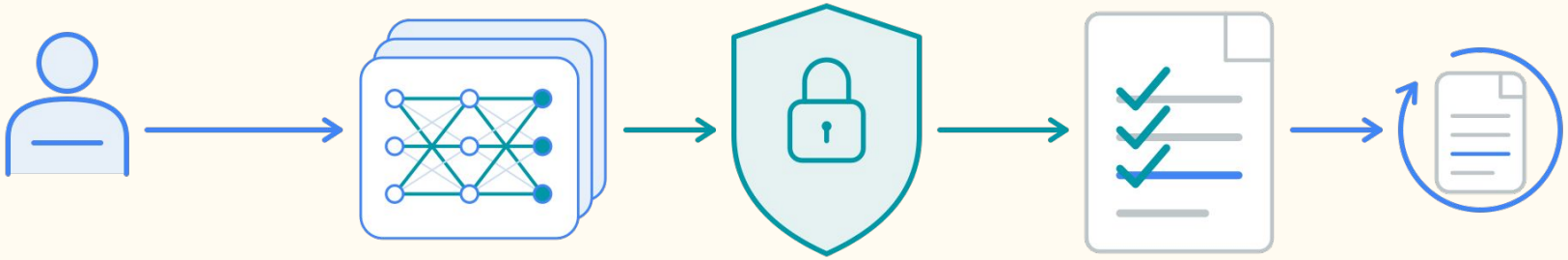
```
● API Error: Opus 4.8 (1M context) has safety measures that flagged this message for a cybersecurity topic. To learn
  about the Cyber Verification Program and apply for access, visit our help center:
  https://support.claude.com/en/articles/14604842-real-time-cyber-safeguards-on-claude.

  If you were not engaging in a cybersecurity topic, please send feedback via /feedback.
```




Hugging Face, July 2026. Original refusal screenshot and reported response.

Open weights gave defenders commensurate response capabilities during analysis.

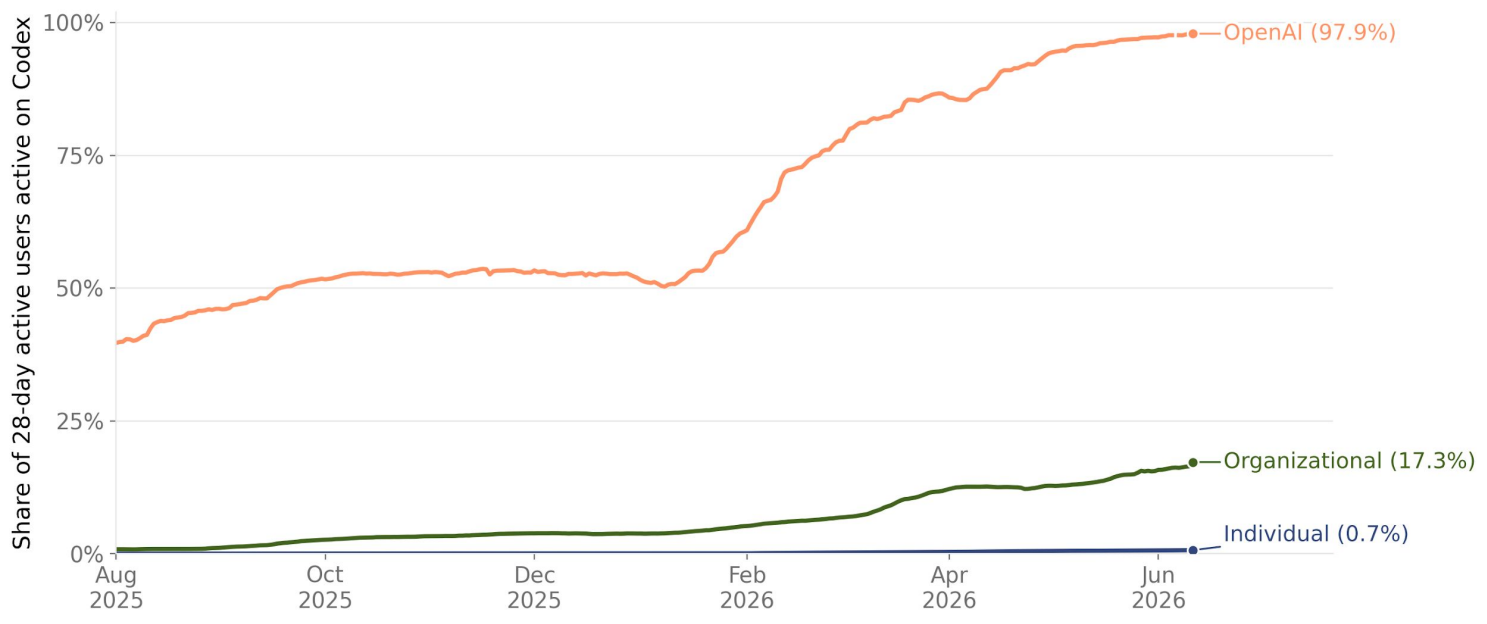
# Safety is a systems problem




Training, incentives, permissions and verification *should* work together.

# 7/8 Workplace and economic pressures

# Adoption depends on the workplace




OpenAI, "The Shift to Agentic AI," June 2026. Figure 1a.

Access, skills and workflow design influence who captures the gains.

# Capital and the pressure to grow

Reported valuation targets in US dollars. Terms may change.

## OpenAI

**About $1.4 trillion**

Pre-money financing target

Reported Sept 29, 2026

Announced round: $852 billion

Post-money, March 31, 2026

Sept 29 reporting

## Anthropic

**$1.8–2 trillion**

Potential IPO valuation

Reported Oct 2, 2026

Announced round: $965 billion

Post-money, May 28, 2026

Oct 2 reporting

Commercial incentives shape the pace of development.

# 8/8 Governance and the frontier race

# Pacing the frontier

## Anthropic

Embedded evaluators inside frontier labs.

Pace capabilities while safeguards catch up.

Amodei, "We Must Pace the Frontier," September 2026

## OpenAI

Common safety standards. Incident reporting.

Human control over automated AI research.

OpenAI, "Building standards," September 2026

Governance needs verifiable safeguards, ideally *externally* imposed.

# International competition and cooperation

## Restrict capabilities

Amodei argues for tighter chip controls and curbing unauthorized distillation.

## Coordinate on safety

OpenAI calls for US–China dialogue on incidents and critical infrastructure.

Chinese open models also contribute to the ecosystem, as in the HF response.

Amodei, September 2026

OpenAI, September 2026

National competition complicates shared rules for global risks.

Useful, safe autonomous agents will comprise:

**Intelligent core models**

**Well-designed harnesses**

**Bounded authority**

**Human accountability**

# Further reading and listening: opinion pieces

## Essays

Séb Krier: Of Swarms and Sand Gods

Andrew Lampinen: In defense of metaphor

## Ezra Klein: interviews and commentary

Why Are We Sprinting Off the A.I. Cliff?

Jensen Huang on A.I. alarmism

Bill Gates: A.I. and catastrophic risk

David Robinson: why he left OpenAI

Arvind Narayanan: a ‘normal’ A.I. future

# Further reading: how the systems work

GPT-2: Better language models and their implications

Stochastic Parrots: the authors' context and coverage

OpenAI: Our approach to alignment research

OpenAI: Learning to reason with LLMs

Anthropic: Building effective agents

DeepSeek-R1: open weights and distilled models

# Further reading: risks and society

Hugging Face: the July 2026 incident

OpenAI: the incident and the road ahead

Anthropic: reward hacking in controlled coding tasks

OpenAI: how agents are transforming work

Dario Amodei: We Must Pace the Frontier

OpenAI: Building standards for the next phase of AI