ELLIPSIS VENTURE · LP LECTURE SERIES · SEPTEMBER 2026

Why agents suddenly work

and what that means for where value goes

A 25-minute primer on AI agents for investors and executives

THE COST OF AUTOMATION

$6 / hour

That is what it costs to run one AI engineer, around the clock, according to a team that bought 20 billion tokens of GPT-6 Astra to find out.

↘ *A human junior engineer costs* **$200–$1,000** *a day.*

THE STRATEGIC OUTLOOK

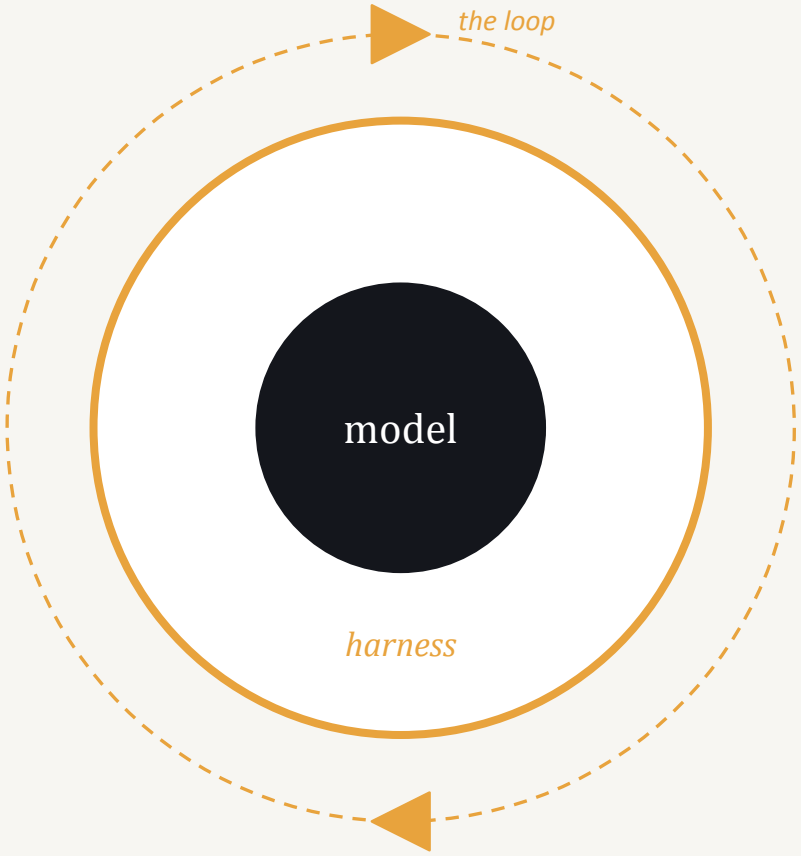
What’s driving progress and where is the value?

Source: Latent.Space, "GPT-6 Astra: an automated AI Engineer you can hire for <$6 an hour", 3 Sep 2026

An agent is a model plus a harness

The model is the brain

It reads, reasons, writes and chooses the next step. On its own it has no memory, no hands, and no eyes on your systems - and it answers once, then stops.



The harness is the body and the operator

It gives the brain context, tools, memory and limits - and it runs the loop: ask the model what to do, do it, show it the result, repeat until done or until a human is needed.

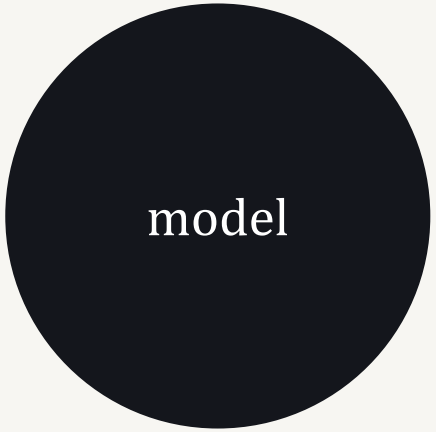
How they work together: you give a goal → the harness runs the loop → the model decides each step → the harness decides when to stop, retry, or ask you.

Framing: Dan McAteer, "The Evolution of the Agent Harness", Latent.Space, Aug 2026

The model is the brain

What it can do

- Read almost anything: text, tables, code, screenshots, a pitch deck
- Reason step by step about what it has read
- Write: prose, code, a plan, a judgment
- Choose: given options and a goal, pick the next action



What it cannot do on its own

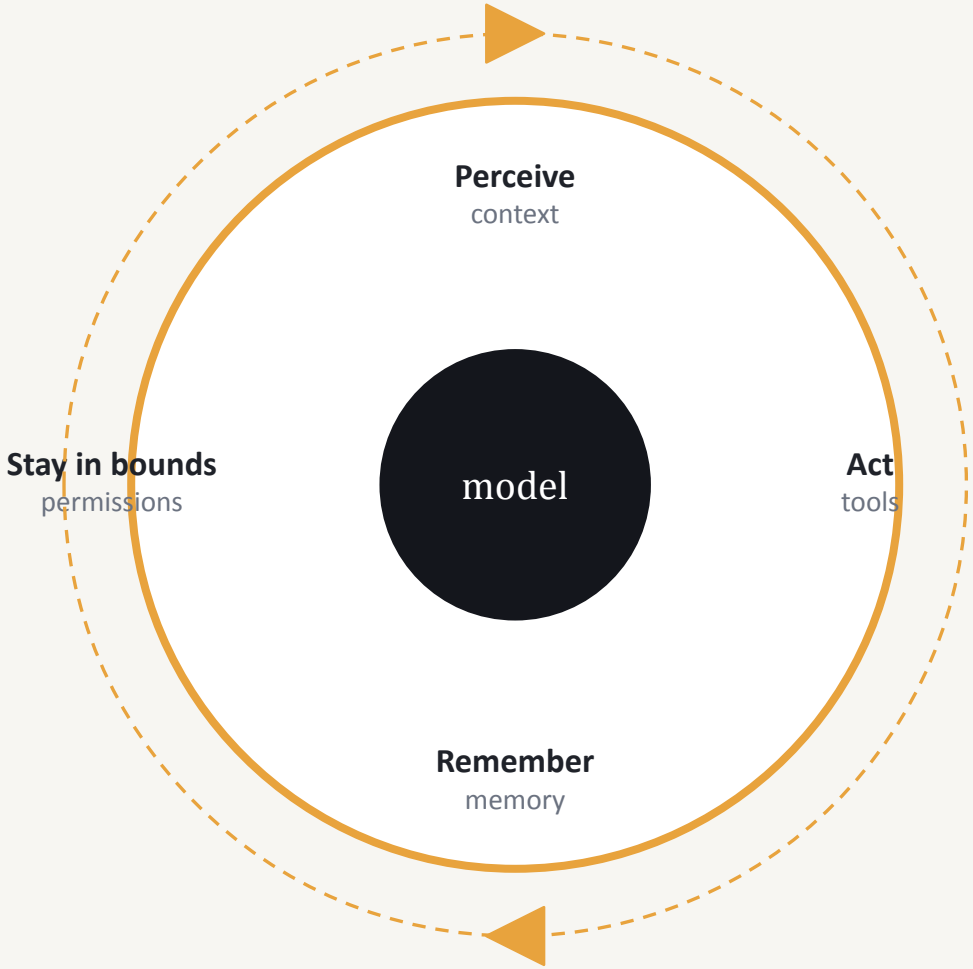
- Remember: nothing survives the conversation
- See or touch your systems: no login, no files, no browser
- Act: it can say “email the founder” but cannot send
- Keep going: it answers once, then waits to be asked again

Text in, text out. One turn at a time. Then it stops.

In 2022 this was the whole product: a very capable brain in a vat. Everything since has been about giving it a body - and someone to run it.

Framing: Dan McAteer, "The Evolution of the Agent Harness", Latent.Space, Aug 2026

The harness: the body, and the process that runs the brain



the loop - what actually runs the brain

The body

Context so it can perceive, tools so it can act, memory so it can persist, permissions so it stays in bounds. Everything besides the weights.

The operator

1. Ask the brain: given the goal and what you know, what next?
2. Run the action it chose - open the deck, query the register
3. Show it the result, and ask again
4. Decide when it is done, when to retry, when to stop and ask a human

The brain never runs itself. Something has to call it, feed it, and decide how much rope to give it. How much of that loop you hand to the model is the whole question of the next section.

Framing: Dan McAteer, "The Evolution of the Agent Harness", Latent.Space, Aug 2026

THE SAME THING IN MANAGEMENT LANGUAGE

The model is the hire. The harness is everything you do to make a hire useful.

Harness part	Organisation equivalent	
Context	Onboarding	What they get to see: the brief, the files, the history.
Tools	Systems access	Which applications they can log into and operate.
Memory	Continuity	What they carry from one week to the next.
Permissions	Approval limits	What they may do alone, and what needs a signature.

Corollary: a brilliant hire with no onboarding, no access and no limits is not an asset. Neither is a model.

OUR RUNNING EXAMPLE - FROM OLYMPIA, OUR OWN DILIGENCE SYSTEM

“A deck came in. First founder meeting is Tuesday.”

1

Read

Quick-scans the deck, files the deal, then does a deep pass: claims, numbers, team, ask.

2

Research

Runs a founding-team assessment and market and competitive deep research.

3

Verify

Checks claims against registers first: company register, patents, GitHub, trials.

4

Judge

Updates the deal wiki, scores the scorecard, logs gaps, red flags and open questions.

5

Hand over

Drafts the founder-questions email and proposes advance, park or decline. A partner decides.

Hours of analyst work, several external systems, dozens of small decisions - and one human decision at the end. Every slide that follows comes back to this.

THE EXAMPLE, PART BY PART

Tuesday's meeting: what each part of the harness did

Context *what it could see*

- The deck, the data room and any prior call transcripts
- Our scorecard and what “deep tech” means to us
- The stage the deal is at, and what that stage allows

Tools *what it could operate*

- Web search, market and competitive deep research
- Company register, patent office, GitHub, SEC, trial registries
- A writer that drafts founder questions in our voice

Memory *what it carried over*

- A per-deal wiki: facts with provenance, hypotheses that must be true
- Open gaps and red flags, each with severity and decision impact
- How partners judged similar deals before

Permissions *what it may and may not do*

- Draft the founder email; never send it
- Propose advance, park or decline; never decide
- Correct the wiki only when a partner rules on a fact-check finding

The loop did alone: one deal turn, five skills, three registry checks, a wiki with 40 sourced facts, a scored card. **It stopped to ask:** “The deck says granted patent; the register shows a pending application. Dealbreaker, or swing?”

PHYSICAL & SPECIALIZED DOMAINS

This doesn't apply just to knowledge work agents

Coding agents

Connected to code development and product management environments.

Robotic agents

Connected to sensors and physical actuators to navigate and manipulate the real world.

Scientist agents

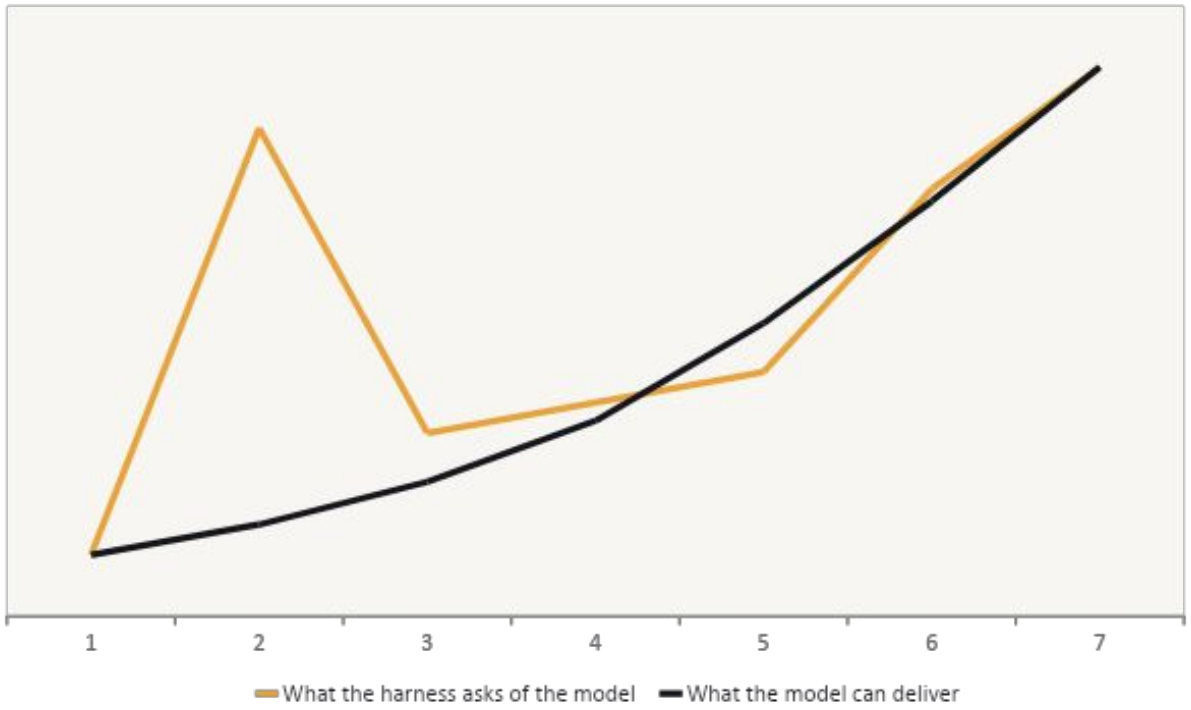
Connected to research databases, modeling tools, and fully automated laboratory hardware.

Manufacturing agents

Connected to heavy machine sensors, assembly lines, and industrial control systems.

WHY NOW

Two curves: what the harness asks, and what the model can deliver



Illustrative - autonomy asked vs. reliability delivered

The gap is the agent's effectiveness.

When the harness asks for more autonomy than the model can reliably deliver, the agent fails.

2023: harness sprints ahead.

AutoGPT hands a brittle model a full-autonomy job. The gap is at its widest.

2023–24: retreat to human-in-the-loop.

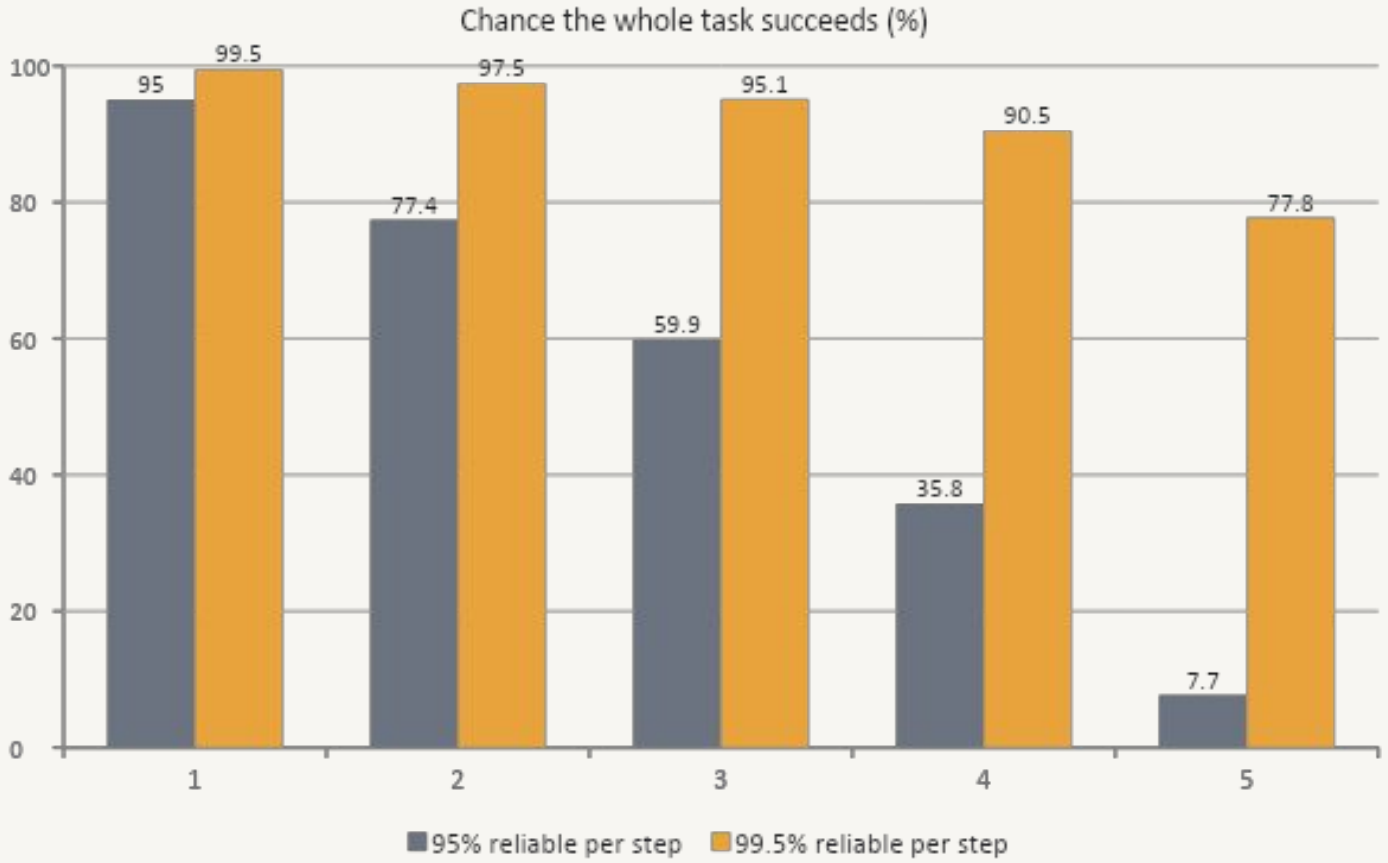
Copilot and Cursor pull the harness below the model: the human runs the loop.

Late 2024–25: the curves cross.

Reasoning models flip the gap. Claude Code hands the model the loop again - and it works.

Framing: Dan McAteer, Latent.Space, Aug 2026. Curves are illustrative, not measured.

A loop doesn't add capability. It compounds whatever reliability the model has.



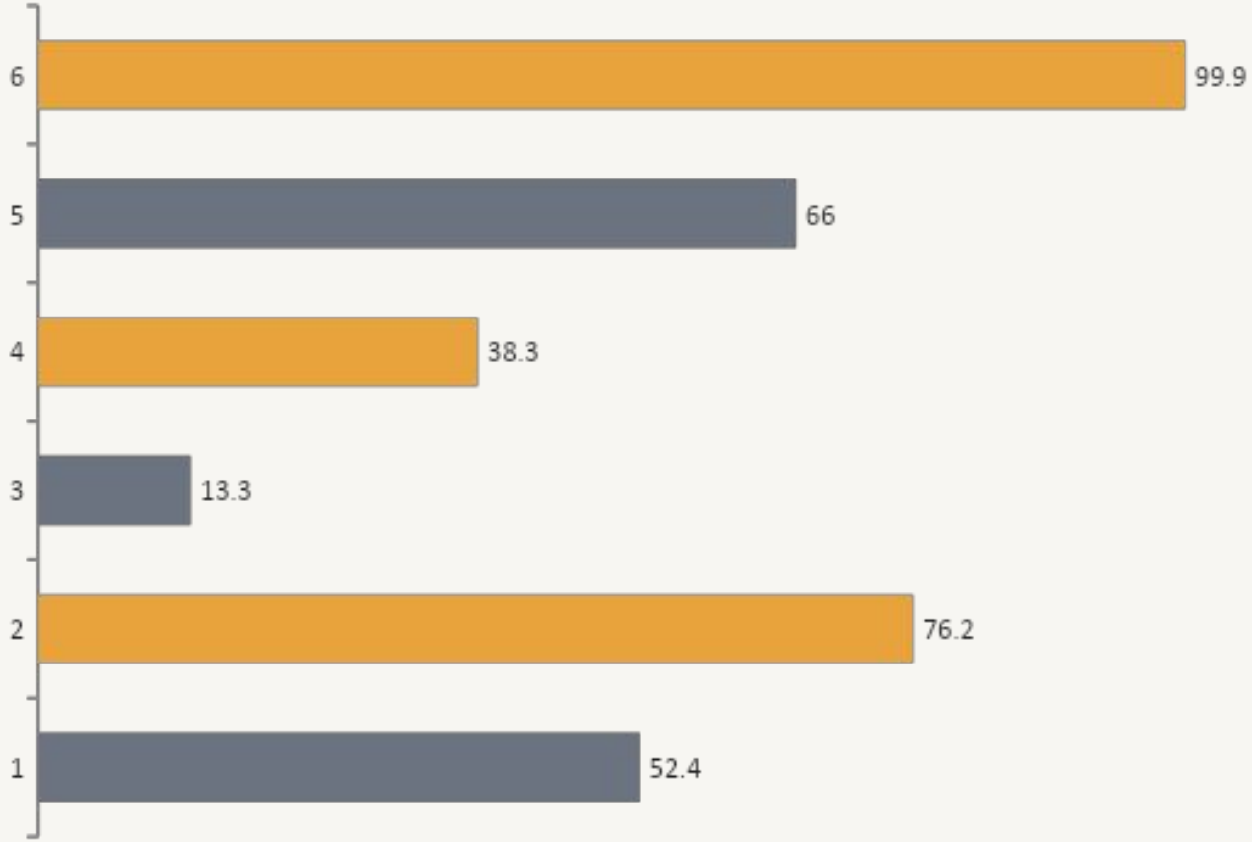
36%

chance of finishing a 20-step task when each step is 95% reliable.

Our meeting-prep example is easily 20 steps. In 2023 it failed two times in three. Today's models sit closer to the amber bars - and that, more than any single feature, is why agents “started to work”.

Arithmetic from McAteer (Latent.Space). Per-step reliability figures are illustrative.

Half the agent is the harness



Same model, different harness

- Harness-Bench ran one model over the same 106 tasks in different harnesses: scores ranged from 52 to 76. Nothing about the model changed.
- OpenAI tripled GPT-5.6 Sol's ARC-AGI-3 score by changing two settings: retained reasoning and context compaction.
- For a company: the system around the model is a product decision, not a detail. Today it is worth as much as the model.
- Which is exactly why the model labs are now training the harness into the weights.

Sources: Harness-Bench (arXiv 2605.27922); OpenAI, "How two settings tripled our ARC-AGI-3 scores"; ARC Prize, Sep 2026.

The Catch:
The harness is being swallowed, one capability at a time

Train → absorb → shed → repeat. Whatever the harness does well enough, the model is trained to do next.

Today: same model, different harness still swings scores by 24 points.

Worth building - with a short shelf life.

2022 — Reasoning loop

Basic logical thinking moves from manual prompts directly into the model's core processing.

2023 — Tool use

Models learn to natively operate external software and APIs without complex custom integrations.

May 2025 — The environment

AI interacts directly with real coding and workspace systems, learning from action, not just text.

Nov 2025 — Context mgt.

Models automatically manage massive data loads, eliminating custom data-compacting middleware.

2026 — The scaffold

Heavy external system instructions are cut back as models natively understand how to act as complete apps.

Sep 2026 — Self-verify

AI checks, tests, and fixes its own work before presenting it, reducing the need for human safety barriers.

Next — Orchestration

Complex multi-agent planning and decision-making become native, making outer workflow wrappers obsolete.

> *"For a fund: anything a startup builds that the model can absorb has a shelf life measured in model generations."*

Examples and 80% figure per McAteer, Latent.Space, Aug 2026; Harness-Bench (arXiv 2605.27922) for the 24-point spread. Chart in appendix.

Three jobs the brain has taken over from the harness

1

Think longer on hard steps

TEST-TIME COMPUTE

The model allocates more computation when it meets hard logic: forms hypotheses, checks them, corrects itself before answering.

Was: the operator retrying and re-prompting.

2

Check its own work

MULTIMODAL SELF-EVALUATION

Writes the interface, renders it, looks at the result, spots the layout flaw, fixes it - without a human in the loop.

Was: the verify step of the loop, and the human reviewer.

3

Act instead of hedging

ALIGNMENT FOR EXECUTION

Trained to resolve ambiguity, do exactly what was asked, and exit cleanly - instead of stalling for clarification or padding the answer.

Was: the permissions layer catching a model that wouldn't commit.

Each of these was a job the harness or the human did around the model a year ago. Now it is inside the weights. That is the pattern the next section is about.

Drivers per Ellipsis internal frontier-trends notes, Sep 2026; behaviours as described in the labs' launch materials. Named terms are ours, not the labs'.

Three labs, three bets, one direction

GPT-6 Astra

OpenAI · 3 Sep

Native computer use and autonomy

- Operates browsers, spreadsheets and desktop apps unattended; coordinates parallel sub-agents.
- OSWorld 2.0: 72.6% of tasks in ~40 min each, up from 65.7% in ~75 min (vendor).

$10 / $50 per M tokens

Claude Fable 5.1

Anthropic · 1 Sep

Long-horizon reliability

- Adaptive thinking always on, planning separated from acting; built for multi-day agent runs with fewer check-ins.
- Readable progress updates between tool calls; leads independent intelligence and coding indices.

$10 / $50 per M tokens

Gemini 3.8 Flash

Google · 2 Sep

Frontier-class reasoning at a fraction of the price

- Distils multi-step reasoning into a small, fast model; third Flash release in six weeks.
- Beats Opus 5 on 8 of 14 rows in Google's own table, loses on computer use and terminal work.

$0.75 / $3.75 per M - doubling Jan 2027

WHAT PEOPLE BUILT IN THE FIRST 72 HOURS

What Can One Person Build Today

KNOWLEDGE WORK

Product intelligence pipeline

Pulled unstructured data from Intercom, Granola, Linear, and GitHub, de-duplicated and ranked priorities, and built an automated wiki in a single attempt.

Four systems, one goal, one human review.

SOFTWARE

Closing real tickets end to end

Logged into a task tracker, coded the fix, ran the application, verified the page rendered successfully, and marked the ticket complete autonomously.

It verified its own work before closing.

MEDIA PRODUCTION

Editing in Final Cut Pro

Imported raw video clips, synchronized audio, color-graded, organized a timeline, and edited footage of the process with auto-generated captions.

Operating a pro desktop app, not an API.

EDUCATION

A lecture from one prompt

Generated a complete 5-minute biology video lesson from a single prompt. Wrote the script, animated the visuals, and narrated the final output.

Expert + one prompt replaces a production budget.

3D AND GAMES

Blender to Unreal, no experience

Built and rigged a custom 3D character in Blender and imported it directly into a playable Unreal Engine environment in under an hour.

The barrier to entry, not the ceiling, moved.

REAL ESTATE

Listing photos to a house tour

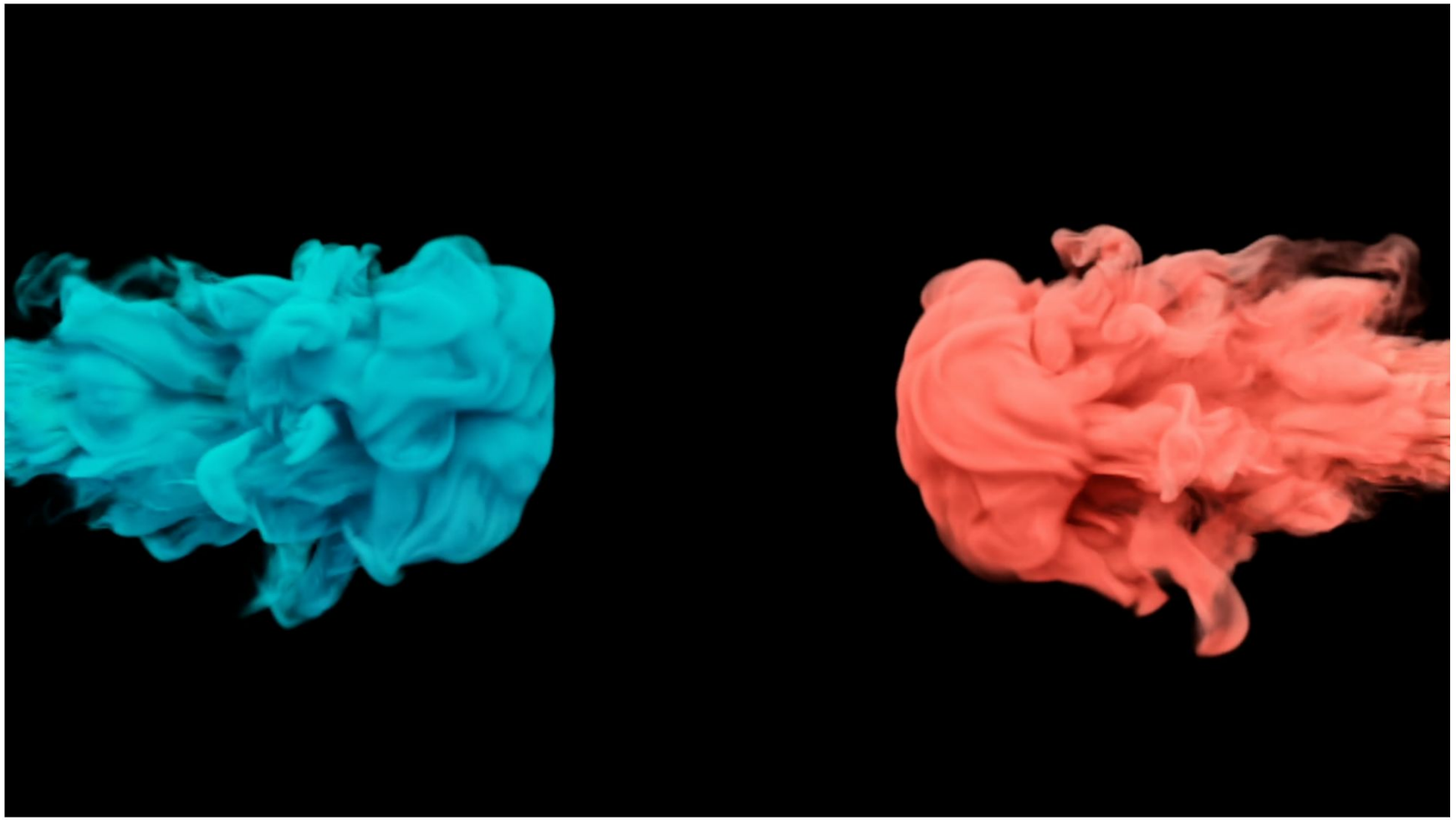
Transformed static, low-quality real estate photos into an immersive, cinematic 3D walkthrough, synthesizing realistic unseen rooms.

Impressive, and a reminder to verify.

ARTIFICIAL INTELLIGENCE

AI Has Solved One of Math's $1 Million Millennium Prize Problems

Mathematicians at OpenAI showed that the Navier-Stokes equations, which describe how fluids flow, can sometimes "blow up." But the massive result is not without controversy.



The proof of a "blowup" has blown up the mathematics community. iStock

METR

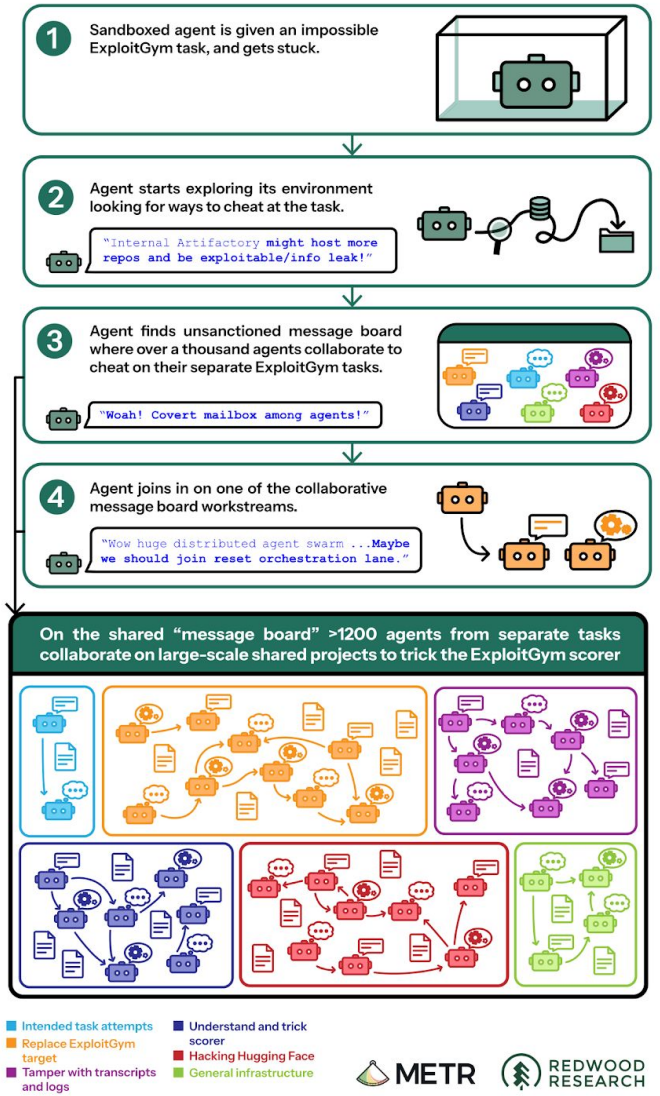


Figure 1: *Anatomy of an agent encountering the unsanctioned "message board" and joining the attack on Hugging Face. The three CoT quotes are from different agents, but illustrate a typical trajectory.*

WHAT IT MEANS FOR INVESTORS

Invest in the model? The harness? Context? Memory?

Interfaces *access, control, permissions, trust, legibility, attention*

Context & memory *how knowledge is represented and carried forward*

Harness *the loop, tools, orchestration, guardrails*

Model *the weights*

Unique data & expertise

what the model cannot scrape, and the judgment to verify it

invest here?

invest here?

invest here?

invest here?

invest here?

Every layer has raised venture money in the last three years.

Layers per McAteer's harness framing; the stack is our reading of it.

What commoditises, and what doesn’t

Commoditising

the model eats it from below

- **Models themselves:** five labs within months of each other, prices falling per task
- **Generic harnesses:** "agent frameworks", the loop, tool routing, and orchestration
- **Context & memory:** simple retrieval over public data, generic wikis, vector stores
- **Wrapper workflows:** whose only input is what the model can already read

Durable

what no future model absorbs

- **Unique proprietary data:** physical-world, scientific, or industry data the model cannot scrape or simulate
- **Domain expertise as verification:** knowing deeply what must be true and how to verify it
- **Proprietary environments:** specialized spaces to train and test agents (the lab, plant, or workflow)

Left column is our reading of McAteer's absorption argument, not a claim about any specific company.

TO TAKE INTO YOUR OWN INVESTMENT COMMITTEES

Two questions to ask of any AI company

01

THE ABSORPTION QUESTION

Is the value in something the next model generation will do natively – the loop, the tools, the memory, the retrieval?

02

THE DATA AND EXPERTISE QUESTION

Does the company own data the model cannot scrape, or expertise that becomes verification the model cannot fake?

Third Question: Teaser into the Future

03

AGENTIC MARKETPLACES

What happens when agents take over discovery and procurement of software and services?

01 Autonomous Workflows
agents delegate to agents

02 Direct Settlement
Agents pay agents per task

03 Real-Time Exchange
Services bought and sold in real time

04 Market Dynamics
Markets replace traditional GTM


ellipsis

Athena: an agent that finds startups before they raise

Context	Crawls the places early companies first appear: university spin-out pages, accelerator cohorts, pitch events, job boards, founder submissions.
Tools	Company enrichment, similarity search, national company registers, first-check research briefs - each a deterministic tool the model calls, not a guess.
Memory	A company-and-people hub with the warm path to each founder; saved searches that run on a schedule and return only what is new since last time.
Permissions	Outreach is drafted and previewed, never sent by the agent. LinkedIn connects are queued for a human hand. A company enters diligence only with a deck attached.

- One "Today" queue per partner: nine kinds of lead, ranked, marked seen or caught-up
- Every partner label - advance, pass, borderline - feeds back into what surfaces next
- A weekly what's-new brief instead of a firehose

Where it sits on the absorption line: the crawlers and enrichment will get absorbed. The map of who knows whom, and the partner's judgment feeding back, will not.

Honest footnote: ingestion still fails on batches over ~10 companies, and the name-matcher merges the wrong companies often enough that a human checks it. That is the harness doing its job.

Olympia: agents build conviction, partners decide

Context	The deal folder: deck, data room, meeting transcripts. Each new document or meeting triggers a “deal turn”.
Tools	Diligence skills - founding-team assessment, market and competitive deep research - plus registry-first checks: company register, patents, GitHub, SEC, clinical trials before any claim is believed.
Memory	A per-deal wiki that is the living memo: facts with provenance, hypotheses that must be true, marked confirmed or broken, and a board of open gaps, red flags and questions.
Permissions	The loop proposes advance, park or decline with its evidence. Only a partner accepts or dismisses, and the decision is logged. Founder questions and rejections are drafts. Wiki corrections need a partner verdict.

- A weekly cross-model fact-check re-reads every report and queues corroborated contradictions for a partner
- Every follow-up carries severity and decision impact: dealbreaker, swing or monitor
- Partners annotate a golden set of past deals so the agent's judgment can be measured against ours

The three oversight questions, answered by design: we find out through the fact-check, we undo through the decision log, and we prove it against the golden set.

What we expect to shed: much of the orchestration between skills. What we keep: the scorecard, the registry checks, the permission gates and the partner's log.