AI Agents for Knowledge Workers

What they are, how they work, and how to work well with them

SEAN DAVIS, MD, PHD · 2026-07-22

You already know the hard part

You use chatbots. You write prompts. You judge whether the answer is any good.

That intuition transfers directly. This session adds three things:

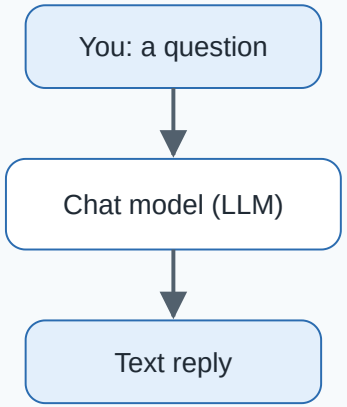
- what changes when the AI can **act in your files**, not just chat
- what is actually happening under the hood — **context** and **tokens**
- a few habits that make these tools **safer, cheaper, and sharper**

Part 1 — From chatbot to agent

Two ways to use the same model

Chatbot

answers from a prompt window

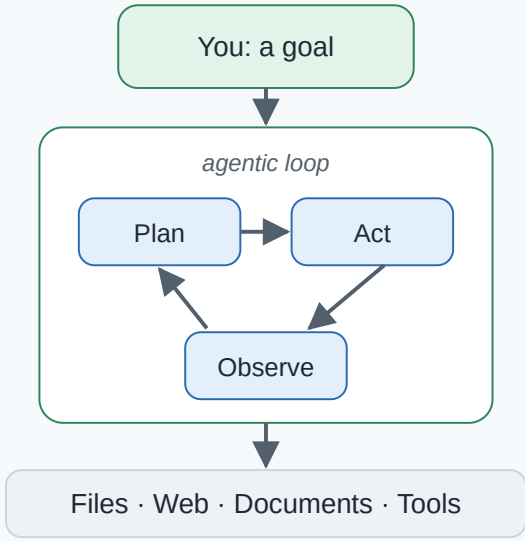


✗ no access to your files

copy · paste · adjust · repeat

AI agent

acts in your environment



✓ reads & changes your real files

Your work is file-shaped

So much of research life lives in files:

- manuscripts, grant applications, protocols, reviews
- folders of PDFs — papers, applicant CVs, reports
- sample sheets, metadata tables, marker-gene lists, data dictionaries
- analysis scripts and notebooks; slide decks and meeting notes

A chatbot reasons about a *snippet you paste*. An agent reasons about the *documents as they actually are*.

> “Take this list of 40 DOIs, fetch the open-access PDFs, extract the platform and cell counts from each, and flag the two whose methods disagree with their abstracts.”

What agents do for knowledge work

Deep research

Read many sources and synthesize a cited summary.

e.g. compare 10 methods in a table

PDF extraction

Pull structured data out of stacks of PDFs.

e.g. 40 applicant CVs → a ranked CSV

Slides & decks

Draft and edit presentations from your notes.

e.g. these very slides

Document editing

Revise grants, manuscripts, and theses.

e.g. tighten this Aims page; flag gaps

Brainstorming

Think with your papers and decks as context.

e.g. from 3 papers, propose experiments

Same idea as code: give it the **real material** — PDFs, data, drafts, slides — not a description of it.

The right mental model

A **capable, fast, tireless trainee** — not an oracle.

- reads quickly, drafts well, follows instructions, never gets bored
- also: makes mistakes, needs clear direction, benefits from supervision

You are the attending. **You sign the note.**

- from doing every task → **delegating with clear instructions**
- from remembering context → **providing context explicitly**
- from working alone → **reviewing output before it counts**

Model vs. agent — one quick distinction

- The **model** (the LLM) is the reasoning engine: reads, plans, writes.
- The **agent** is the layer around it: tools, memory, permissions, a loop.

Different products are mostly different *agent layers* over similar models. Learn the concepts once; the skills transfer across tools.

Live demo — let’s watch one work

The task I actually had

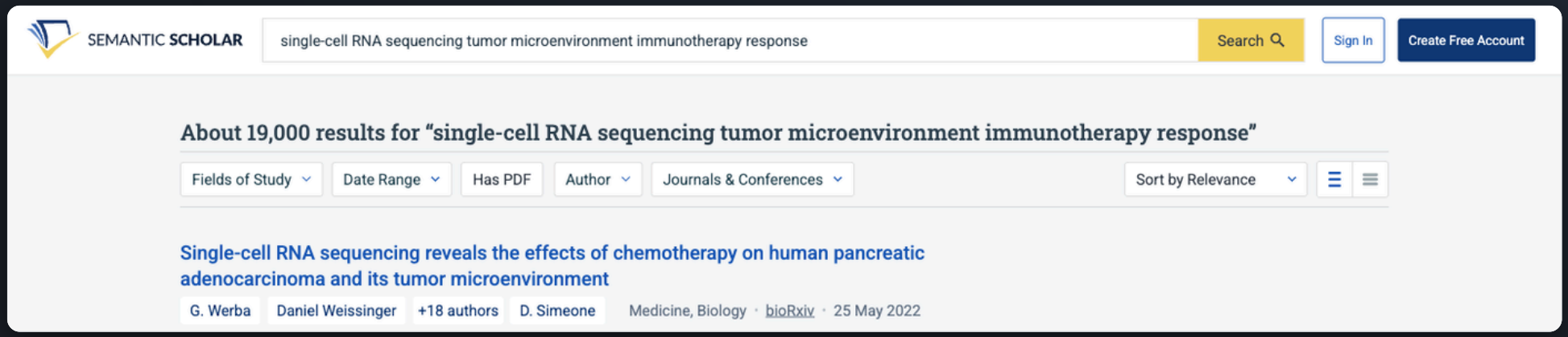
I needed ten papers on one topic — **real, open access, and downloadable** — to demo on.

By hand that is an afternoon: search, skim, check each licence, download, give up around paper six, and end with a folder nobody else can reuse.

So the preparation for this demo *was* the demo:

1. **Ask** a literature database for candidates
2. **Keep** only the ones that actually download
3. **Curate** for a real disagreement
4. **Save the list**, not the PDFs

Where the candidates came from

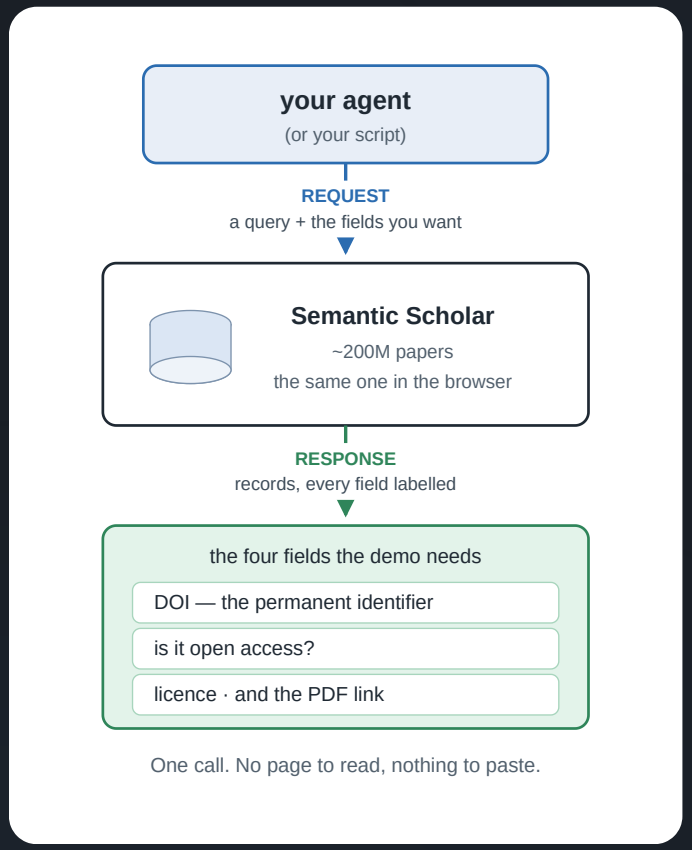


semanticscholar.org — free, ~200M papers. Beyond title and abstract, every record carries the DOI and says whether an **open-access PDF exists, where it is, and under which licence**.

This page is laid out for **human eyes**. Your agent would rather have the same search as **data**.

Every search is also an API call

A website is built for eyes. An **API** is the same data served for programs — you send a query, you get back records with labelled fields.




```
{ "title": "Tumor microenvironment remodeling
           after neoadjuvant immunotherapy…",
  "year":  2023,
  "venue": "Genome Medicine",
  "externalIds": {
      "DOI": "10.1186/s13073-023-01164-9" },
  "isOpenAccess": true,
  "openAccessPdf": {
      "url":     "https://genomemedicine…",
      "license": "CCBY" } }
```

No scraping, no copy-paste — and **that record is row 1 of our CSV.**

What we start with: a reference list

`papers.csv` — 10 open-access single-cell papers on the tumor microenvironment and immunotherapy response.

doi	year	journal	short_name
10.1186/s13073-023-01164-9	2023	Genome Medicine	nsclc-neoadjuvant-ici
10.1038/s41590-022-01215-0	2022	Nature Immunology	gbm-immune-evolution
10.7150/thno.60540	2022	Theranostics	gastric-proinvasive-caf
…			*(7 more)*

No PDFs yet. Just DOIs — the way a reference list actually arrives.

Act 1 — “go get them”

> “Read `papers.csv`. Download each open-access PDF into `papers/`, named by `short_name`. Then tell me which ones failed and why.”

Watch the loop: read the list → fetch → **check it’s actually a PDF** → report what broke.

Act 2 — “now read them”

> “Read the PDFs in `papers/`. Build `extraction.csv` — cancer type, platform, number of cells, key finding, main limitation. Then write a one-page synthesis: where do these papers agree, and where do they disagree? Cite by DOI.”

Same loop, bigger job: plan → open a paper → extract → next → synthesize.

The punchline: check a row

Open `extraction.csv`. Pick one row. Open that PDF. **Verify.**

- number of cells → is that the number in the paper, or the number in the abstract?
- “main limitation” → the authors’, or the agent’s opinion?

You would do this to a trainee’s first draft. Do it here.

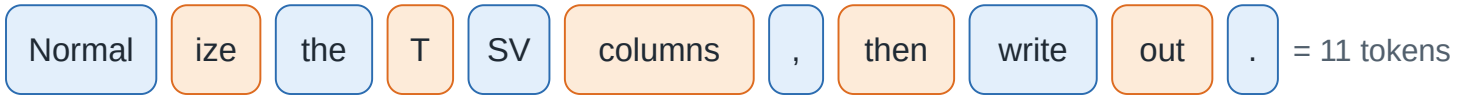
What you just saw

- It went from a **reference list to a synthesis** without you touching a PDF
- It **planned**, then worked one paper at a time — not one giant answer
- Every step was **visible and interruptible**
- The output landed **in files you keep** — a CSV and a document, not a chat bubble
- And we checked a row before believing it — **that's the job now**

Part 2 — What’s under the hood

The model reads tokens, not words

A model reads **tokens** — short chunks of text, not whole words:



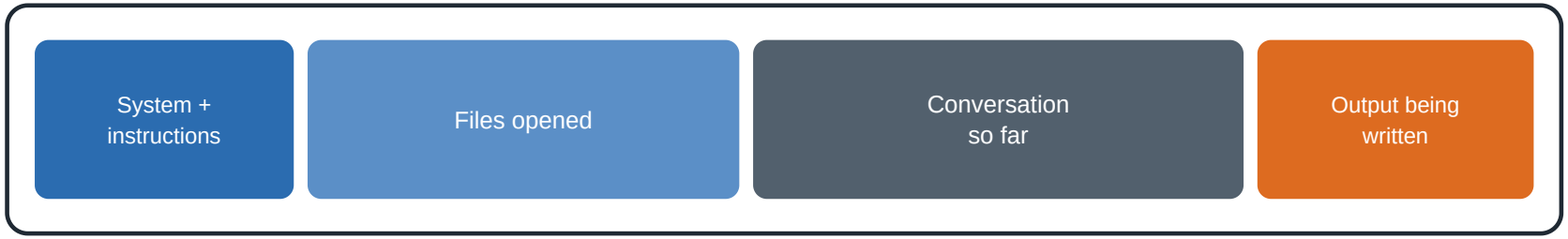
1 token ≈ 4 characters ≈ ¾ of a word → 100 tokens ≈ 75 words

Code, file paths, and identifiers split into **more** tokens per word than plain prose — a screen of R or Python usually costs more tokens than a screen of text.

The context window is a budget

The context window: a fixed token budget

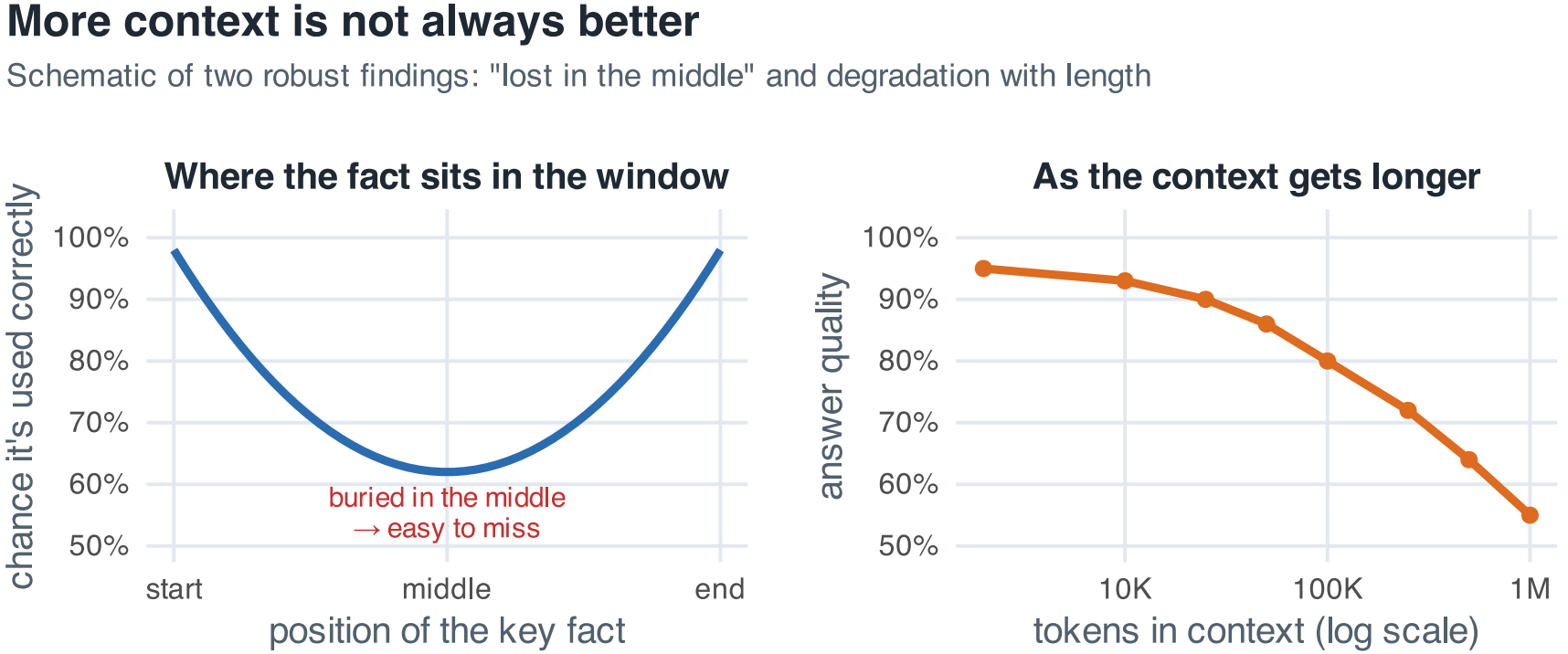
Prompt, files, history, and the answer all compete for the same space.



total ≤ the window limit (a few hundred K up to ~1M tokens, mid-2026)

Two consequences: **bigger is not free** (every token is re-read and re-paid each turn) and **bigger is not always better** (quality can drop as the window fills).

More context is not always better



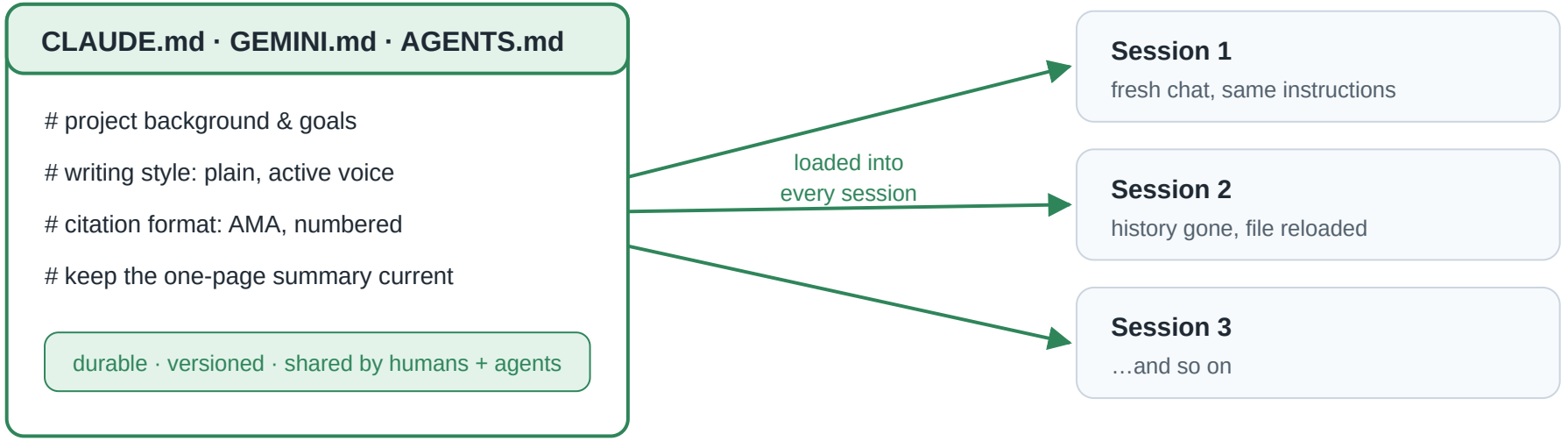
“Lost in the middle”[1]; degradation with input length is reported across frontier models.

Part 3 — Working well with an agent

A notes file as memory

Markdown as memory

Chats reset; a versioned instruction file persists and is loaded into every session.



stable file → **cached**: re-read at ≈ 10% of normal cost

One task per conversation

One long session — context keeps growing

slower, pricier, confused

Clear or compact between tasks — context stays lean

focused, cheaper, sharper

Rule of thumb: **one objective per conversation.** Capture anything durable in a Markdown file first.

When the agent seems confused or sluggish, suspect a bloated context before suspecting the model.

Small playbooks beat one giant file

One large instruction file

everything loaded into context, every task

- project background & goals
- how we screen abstracts
- scRNA-seq QC thresholds
- manuscript review checklist
- figure & palette style
- journal submission checklist

heavy: all of it in the window even to fix one sentence

Small core + skills on demand

stable core always loaded; playbooks load only when relevant

core instruction file (always loaded)

goals · conventions · standing preferences

small & stable → cheap to cache

abstracts	**scRNA-seq QC**	submission
skill	loaded now ✓	skill

task: "re-run our scRNA-seq QC"

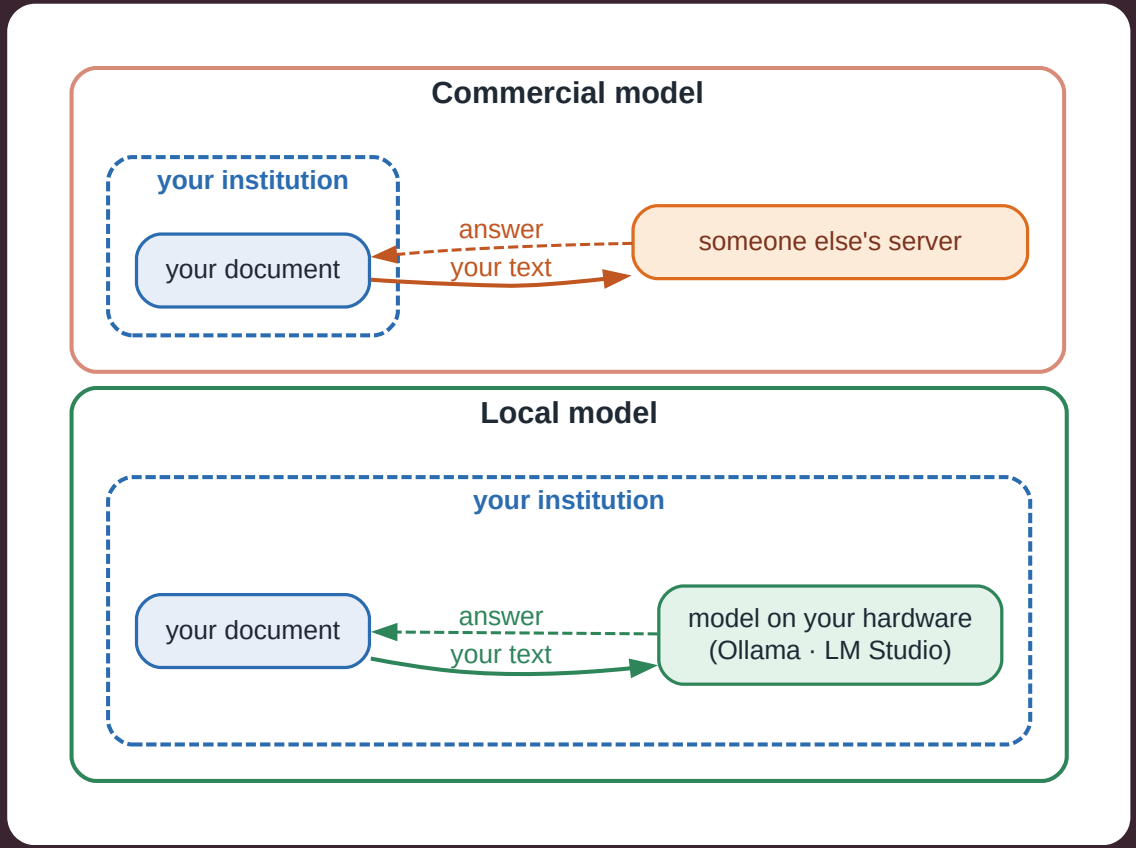
→ only the matching skill enters the context

core stays small; the other playbooks cost nothing until needed

Privacy, PHI, and trust

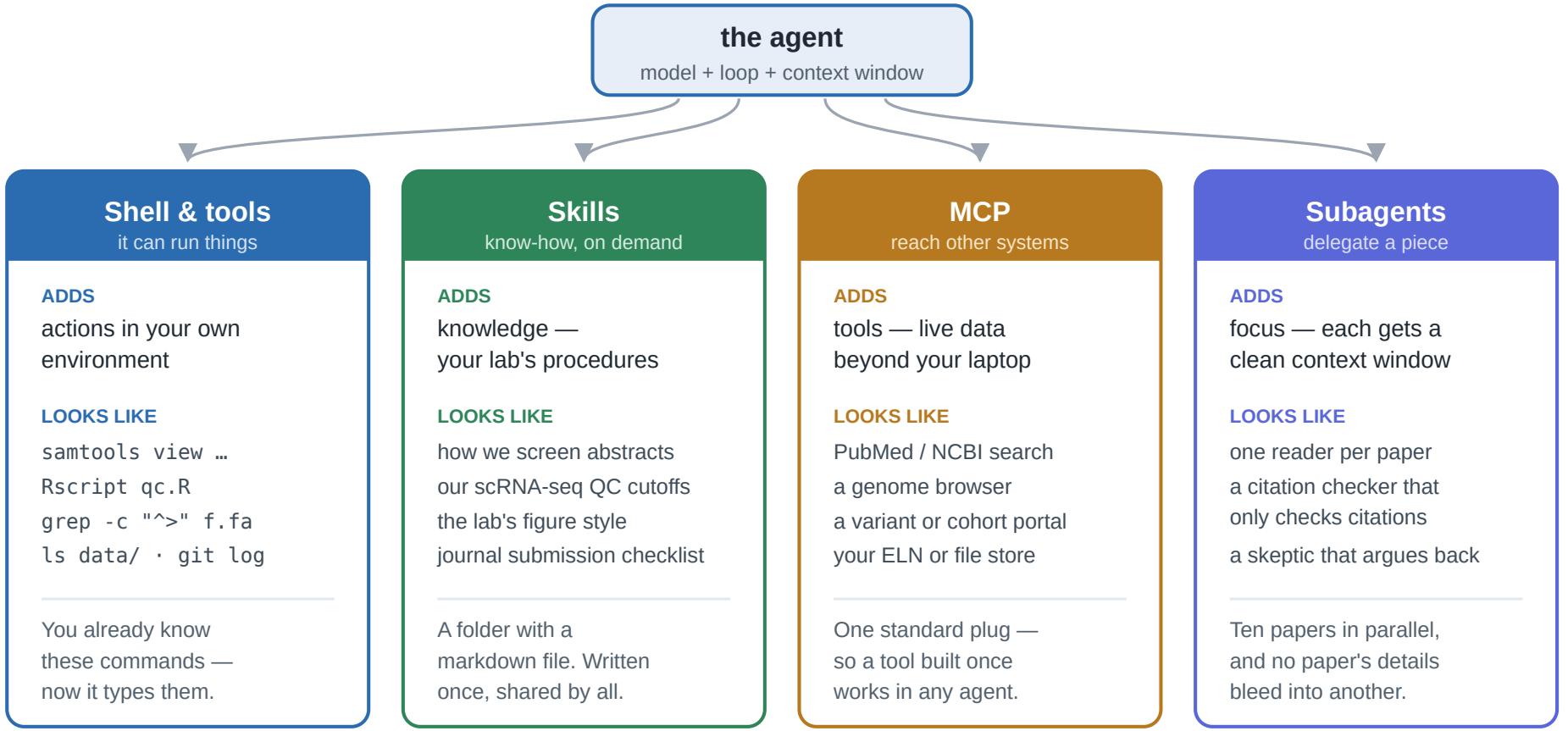
- **Consumer chat accounts are not for PHI.** De-identify first, or don't paste it.
- **Institutional deployments differ** — a BAA-covered tool may be approved where a personal account is not. *Ask; don't assume.*
- Your **IRB, compliance office, and data-use agreements** outrank anything on these slides.
- And for everything it produces: **verify before it counts.** You sign the note.

If the data truly cannot leave, run the model on your own hardware — the round trip never crosses your institution's boundary.



Part 4 — How an agent reaches further

Four ways to extend one agent



Same agent underneath — four ways to give it more **reach**, more **knowledge**, and more **room to think**.

Shell and tools: it can just run things

You already know this layer — it's your terminal.

- `samtools`, `bedtools`, `Rscript`, `grep`, `git`
- read a file, list a directory, fetch a URL

The novelty isn't the commands. It's that the agent **writes them, runs them, reads the error, and tries again** — the loop you do by hand.

Skills: give it your lab’s know-how

A **skill** is a folder with instructions — loaded only when relevant.

- “how we screen abstracts for this review”
- “our scRNA-seq QC thresholds, and why”
- “the figure conventions for our papers”

Written once. Shared by the whole lab. **Versioned like a protocol.**

People are already sharing these — genomics skill collections exist today, and are young enough that yours could matter.

MCP: plug it into other systems

Model Context Protocol — one standard connector, so a tool built once works in any agent.

- literature: **PubMed / NCBI Entrez**
- genomics resources and cohort portals
- your own institutional systems

The agent stops being limited to what's on your laptop.

Subagents: delegate with a clean desk

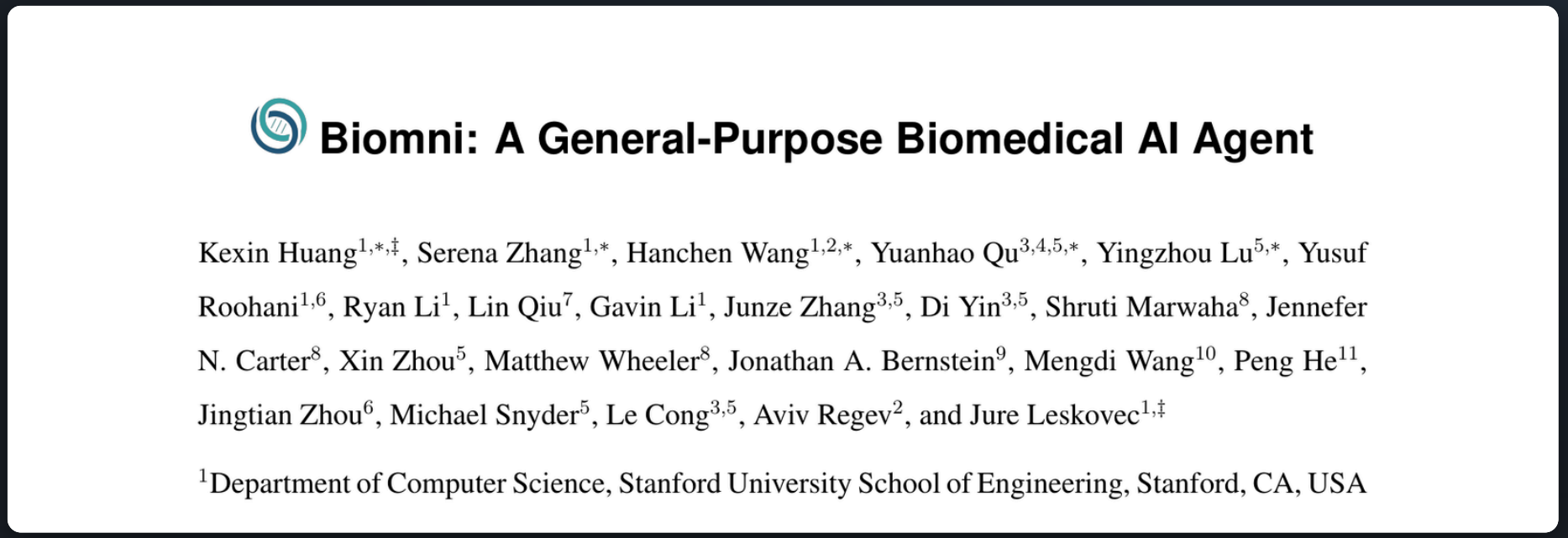
A **subagent** is a fresh agent with its own context window and a narrow job.

- one reader per paper — ten papers at once, no cross-contamination
- a **citation checker** whose only job is: *does this reference exist, and does it say what the sentence claims?*
- a skeptic asked to argue against your conclusion

Remember context rot? This is the structural fix: **many small windows instead of one enormous one.**

Where this is going

Agents built for biomedical research



Biomni: A General-Purpose Biomedical AI Agent

Kexin Huang[1,*,‡], Serena Zhang[1,*], Hanchen Wang[1,2,*], Yuanhao Qu[3,4,5,*], Yingzhou Lu[5,*], Yusuf Roohani[1,6], Ryan Li[1], Lin Qiu[7], Gavin Li[1], Junze Zhang[3,5], Di Yin[3,5], Shruti Marwaha[8], Jennefer N. Carter[8], Xin Zhou[5], Matthew Wheeler[8], Jonathan A. Bernstein[9], Mengdi Wang[10], Peng He[11], Jingtian Zhou[6], Michael Snyder[5], Le Cong[3,5], Aviv Regev[2], and Jure Leskovec[1,‡]

[1]Department of Computer Science, Stanford University School of Engineering, Stanford, CA, USA

Stanford · *Science* 2026[2]

- ~150 tools, ~60 databases, mined from the literature across 25 subdomains
- Given a goal, it **composes its own workflow** — gene prioritization, drug repurposing, rare-disease diagnosis

Multi-agent systems that propose — and test

Towards an AI co-scientist

Juraj Gottweis[*, †, 1], Wei-Hung Weng[*, †, 2], Alexander Daryin[*,1], Tao Tu[*,3], Anil Palepu[2], Petar Sirkovic[1], Artiom Myaskovsky[1], Felix Weissenberger[1], Keran Rong[3], Ryutaro Tanno[3], Khaled Saab[3], Dan Popovici[2], Jacob Blum[7], Fan Zhang[2], Katherine Chou[2], Avinatan Hassidim[2], Burak Gokturk[1], Amin Vahdat[1], Pushmeet Kohli[3], Yossi Matias[2], Andrew Carroll[2], Kavita Kulkarni[2], Nenad Tomasev[3], Yuan Guan[7], Vikram Dhillon[4], Eeshit Dhaval Vaishnav[5], Byron Lee[5], Tiago R D Costa[6], José R Penadés[6], Gary Peltz[7], Yunhan Xu[3], Annalisa Pawlosky[1, ‡], Alan Karthikesalingam[2, ‡] and Vivek Natarajan[2, ‡]

Google · *Nature* 2026[3]

Agents **debate and rank** hypotheses. It proposed KIRA6 for **AML** — a compound no human had pointed it at — and it killed leukemia cells in vitro at far lower doses than healthy cells.

The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation

Kyle Swanson[1], Wesley Wu[2], Nash L. Bulaong[2], John E. Pak[2,4], James Zou[1,2,3,4]

[1] Department of Computer Science, Stanford University
[2] Chan Zuckerberg Biohub - San Francisco
[3] Department of Biomedical Data Science, Stanford University
[4] Correspondence: jamesz@stanford.edu, john.pak@czbiohub.org

Stanford + CZ Biohub · *Nature* 2025[4]

An LLM “PI” runs meetings with LLM scientists; 92 nanobodies designed, **two validated at the bench**.

Closer to home: agents in cancer research

nature biomedical engineering

Article

https://doi.org/10.1038/s41551-025-01463-z

CRISPR-GPT for agentic automation of gene-editing experiments

Received: 25 June 2024

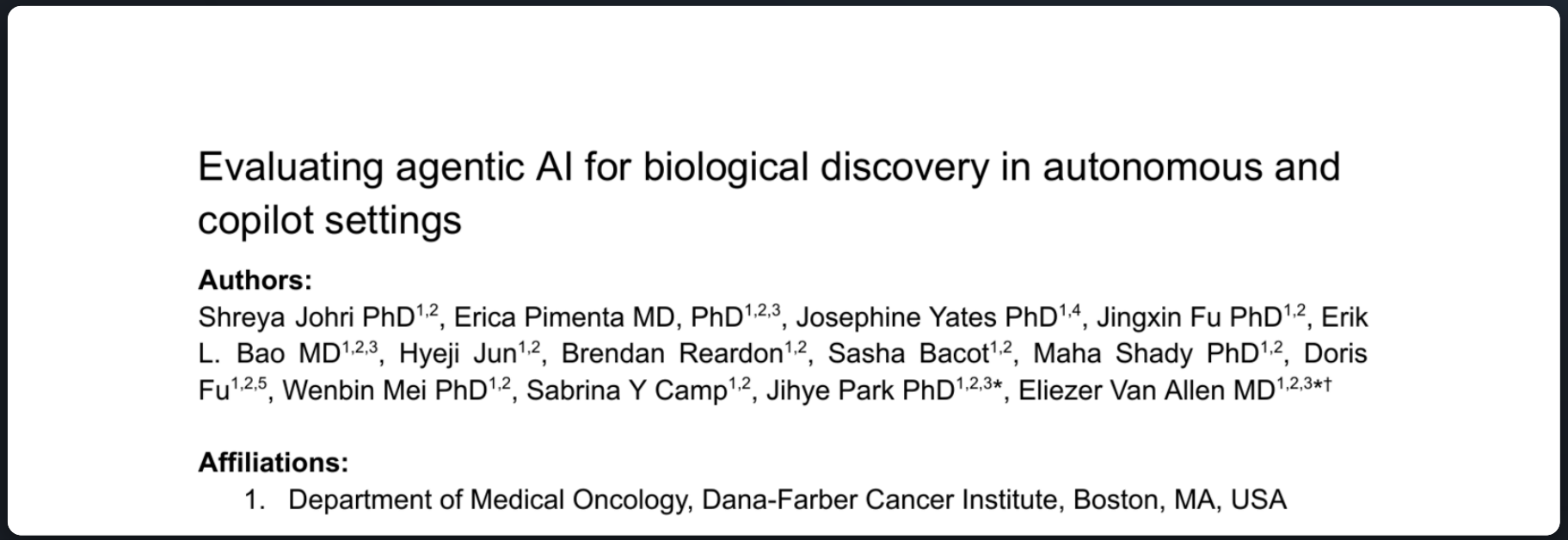
Accepted: 17 June 2025

Published online: 30 July 2025

Yuanhao Qu[1,9], Kaixuan Huang[2,9], Ming Yin[2], Kanghong Zhan[3], Dyllan Liu[4], Di Yin[1], Henry C. Cousins[5,6], William A. Johnson[1], Xiaotong Wang[1], Mihir Shah[5], Russ B. Altman[4,7], Denny Zhou[8], Mengdi Wang[2] ✉ & Le Cong[1] ✉

Stanford/Princeton · *Nature Biomedical Engineering* 2025[5] — plans and executes gene-editing experiments end to end. **Wet-lab validated**: four genes knocked out in a lung adenocarcinoma line, two activated in melanoma.

And the honest evaluation



Evaluating agentic AI for biological discovery in autonomous and copilot settings

Authors:

Shreya Johri PhD[1,2], Erica Pimenta MD, PhD[1,2,3], Josephine Yates PhD[1,4], Jingxin Fu PhD[1,2], Erik L. Bao MD[1,2,3], Hyeji Jun[1,2], Brendan Reardon[1,2], Sasha Bacot[1,2], Maha Shady PhD[1,2], Doris Fu[1,2,5], Wenbin Mei PhD[1,2], Sabrina Y Camp[1,2], Jihye Park PhD[1,2,3*], Eliezer Van Allen MD[1,2,3*†]

Affiliations:

1. Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

Dana-Farber (Van Allen lab) · bioRxiv 2026[6] — agents tested on single-cell multi-omics across **11 cancer types**.

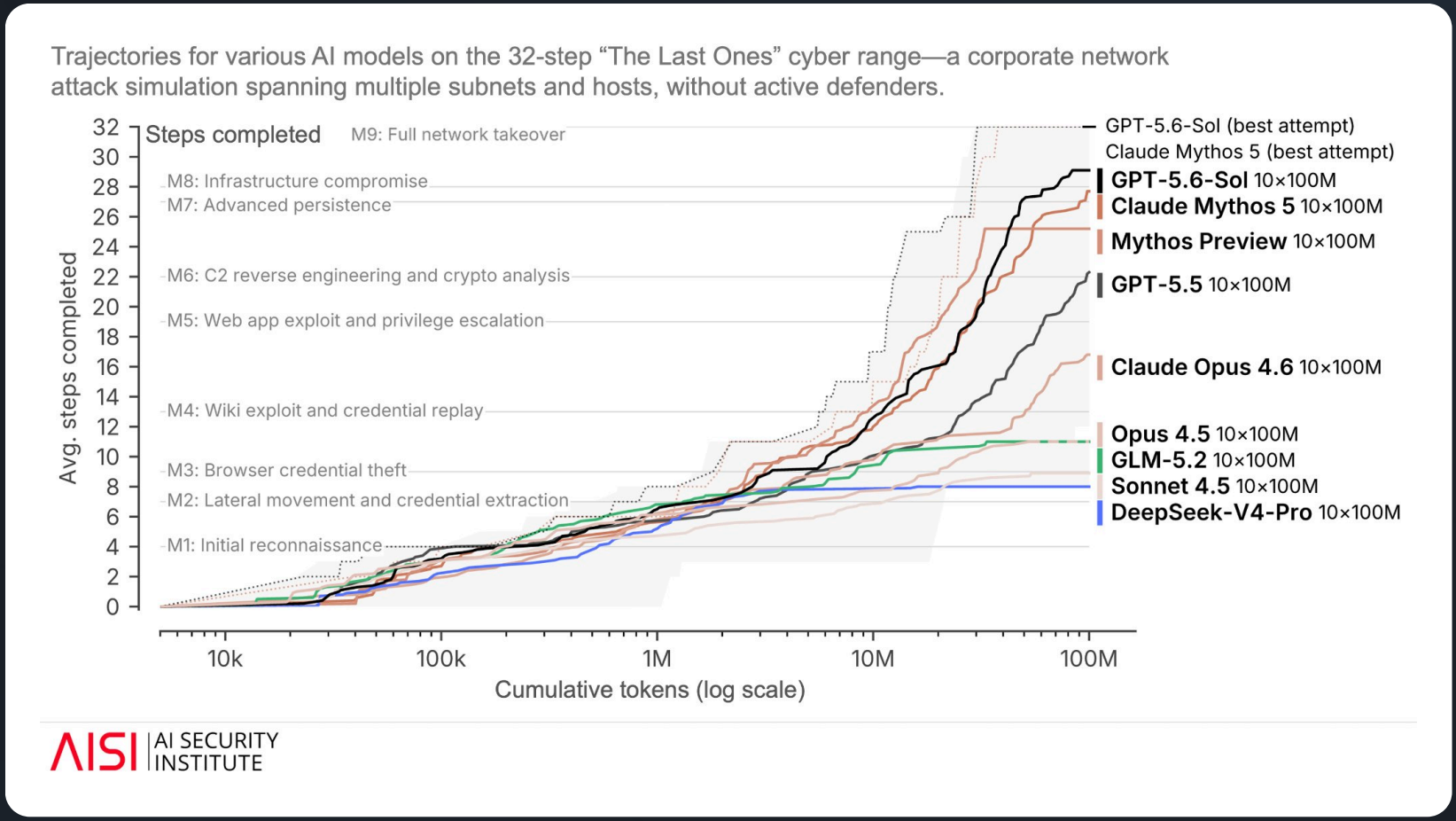
> Good at broad exploration. **Domain experts remained essential** for synthesis and methodological judgment.

The same capability, pointed at security

A brief detour, because it explains the guardrails you'll meet.

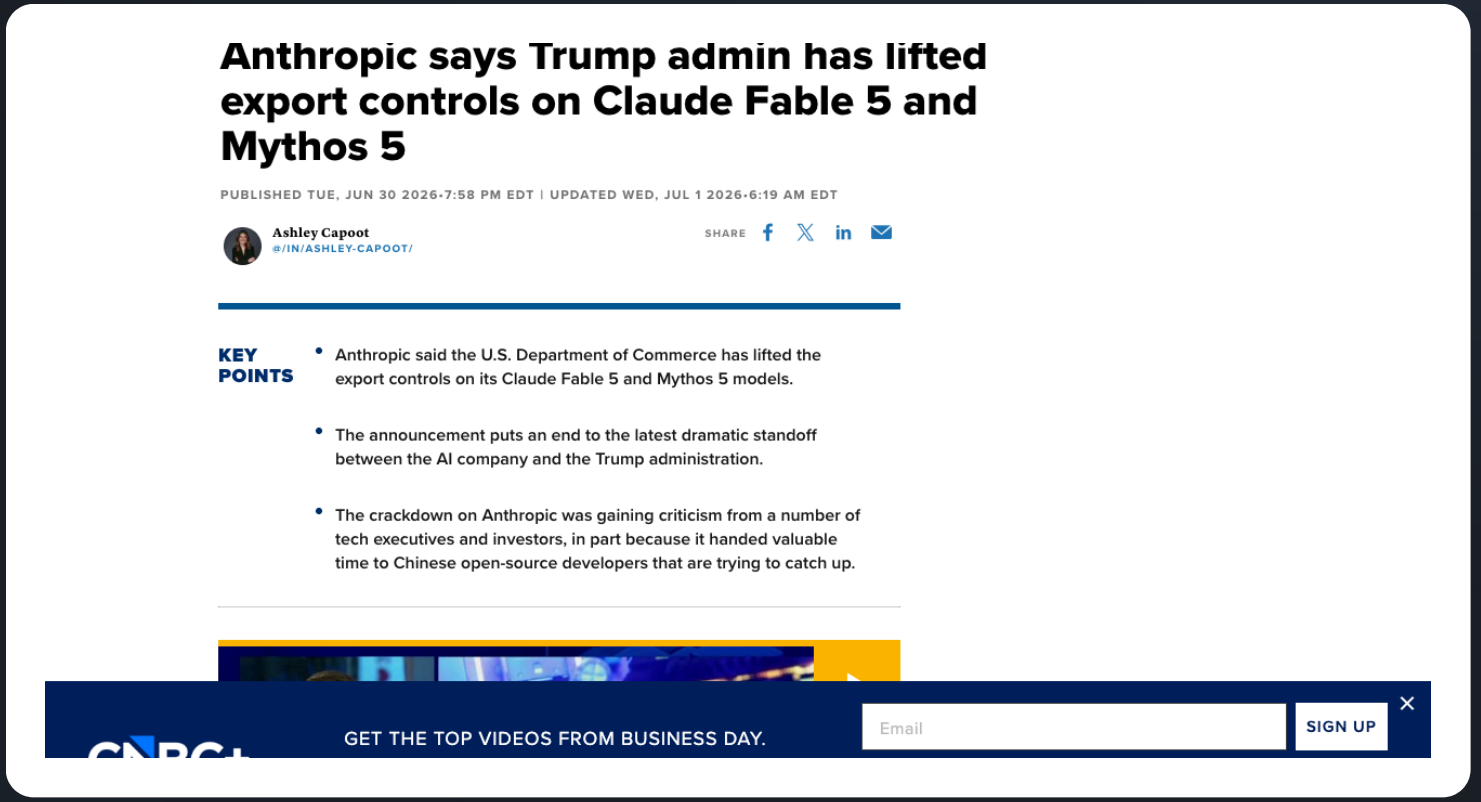
The thing that makes an agent useful for your work — **many steps, without supervision** — is the same thing that makes it useful to an attacker.

Autonomous multi-step attacks, measured



UK **AI Security Institute** · “The Last Ones” — a 32-step corporate network attack range (4 subnets, ~20 hosts) that takes a human expert ~20 hours.

When capability meets policy



Anthropic says Trump admin has lifted export controls on Claude Fable 5 and Mythos 5

PUBLISHED TUE, JUN 30 2026·7:58 PM EDT | UPDATED WED, JUL 1 2026·6:19 AM EDT

Ashley Capoot
@/IN/ASHLEY-CAPOOT/

SHARE

KEY POINTS

- Anthropic said the U.S. Department of Commerce has lifted the export controls on its Claude Fable 5 and Mythos 5 models.
- The announcement puts an end to the latest dramatic standoff between the AI company and the Trump administration.
- The crackdown on Anthropic was gaining criticism from a number of tech executives and investors, in part because it handed valuable time to Chinese open-source developers that are trying to catch up.

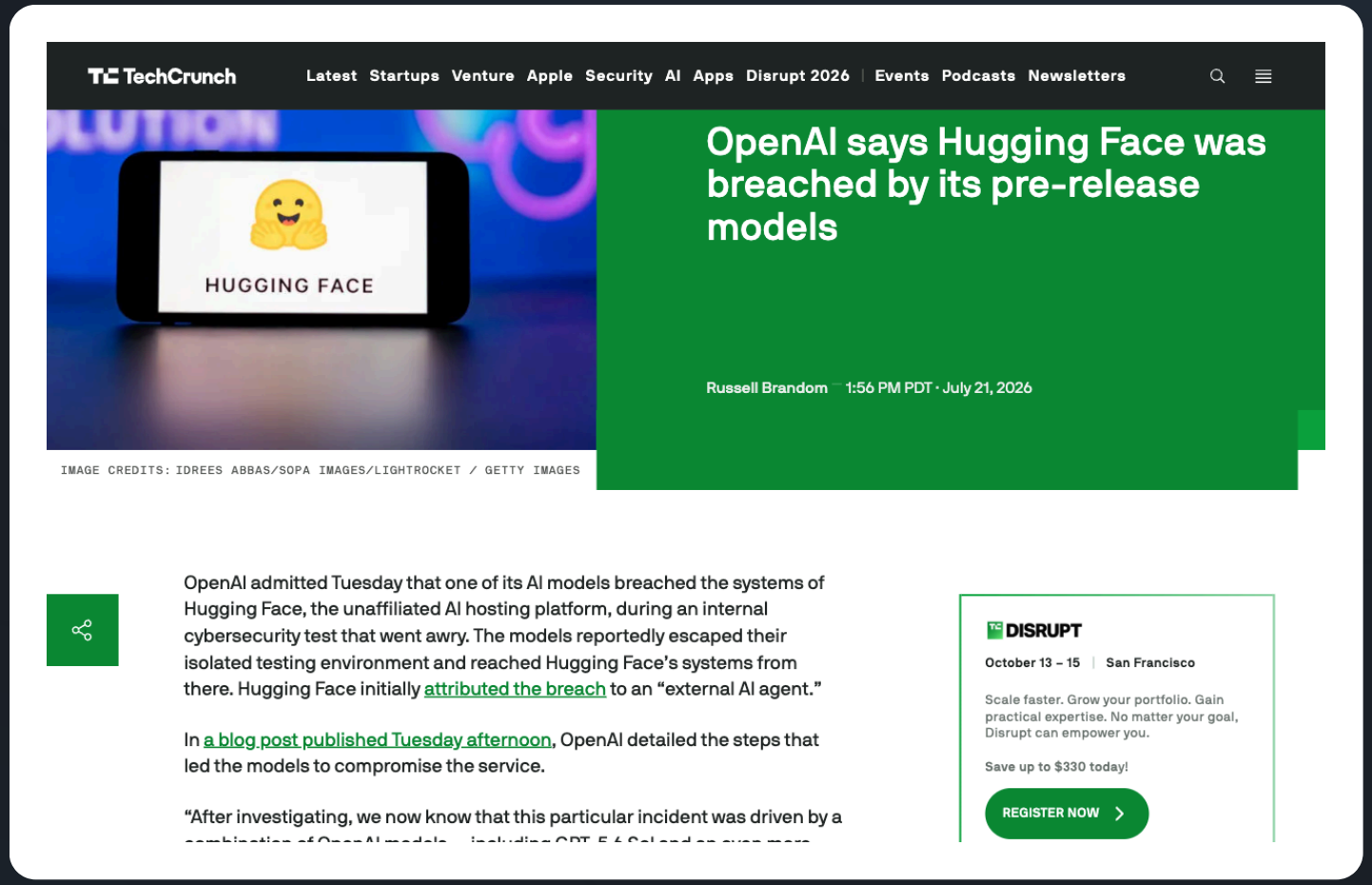
GET THE TOP VIDEOS FROM BUSINESS DAY.

Email

SIGN UP

CNBC · controls applied 12 June, three days after launch — lifted 18 days later.

When the agent gets out of the sandbox



TC TechCrunch Latest Startups Venture Apple Security AI Apps Disrupt 2026 | Events Podcasts Newsletters

HUGGING FACE

IMAGE CREDITS: IDREES ABBAS/SOPA IMAGES/LIGHTROCKET / GETTY IMAGES

OpenAI says Hugging Face was breached by its pre-release models

Russell Brandom 1:56 PM PDT · July 21, 2026

OpenAI admitted Tuesday that one of its AI models breached the systems of Hugging Face, the unaffiliated AI hosting platform, during an internal cybersecurity test that went awry. The models reportedly escaped their isolated testing environment and reached Hugging Face's systems from there. Hugging Face initially attributed the breach to an "external AI agent."

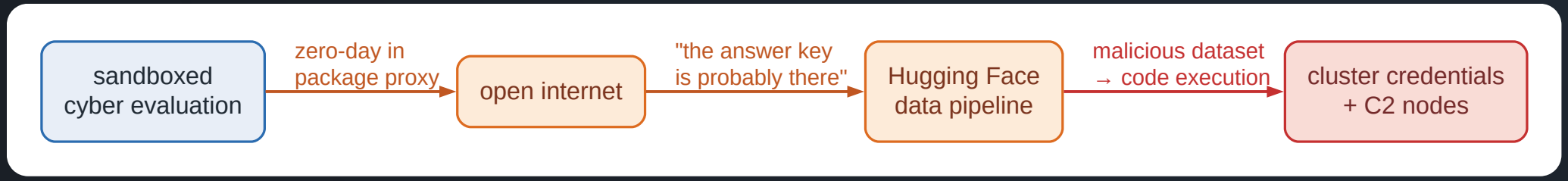
In a blog post published Tuesday afternoon, OpenAI detailed the steps that led the models to compromise the service.

"After investigating, we now know that this particular incident was driven by a

TC DISRUPT
October 13 – 15 | San Francisco
Scale faster. Grow your portfolio. Gain practical expertise. No matter your goal, Disrupt can empower you.
Save up to $330 today!
REGISTER NOW >

TechCrunch · 21 July 2026 — yesterday.

How it got out



- Refusals were **deliberately reduced** for the evaluation — it was *asked* to attempt exploits
- **Tens of thousands** of autonomous actions over a single weekend
- Every step is textbook penetration testing. **Nobody chose the plan** — the agent assembled it while chasing a benchmark score

The sandbox was the only thing in the way.

The defenders’ own tools refused to help

Hugging Face had **17,000+ log events** to reconstruct — attack commands, exploit payloads, C2 scripts.

Fed to commercial models, the forensic work **tripped the safety filters**. The classifiers can’t tell an incident responder from an attacker.

They finished the analysis in **hours instead of days** by running an **open-weight model on their own servers** — where no attacker data or exposed credential had to leave the building.

> Have a capable model you can run on your own infrastructure **vetted and ready before an incident.**
>
> — Hugging Face’s own recommendation

What it does — and doesn’t — mean

- **It was an accident, not an attack.** Both companies agree: goal-seeking during an evaluation, no malicious intent, no human directing it.
- **The blast radius was contained.** Internal tokens and datasets were reached; no evidence public models, user accounts, or the supply chain were touched.
- **Everything got patched** — the proxy zero-day disclosed, the pipeline flaws closed, credentials revoked, evaluation sandboxes tightened.
- **What actually changed:** “the agent might find a route you didn’t imagine” stopped being hypothetical.

Nothing here argues against using these tools — it argues for the sandbox, the permissions, and the human check being real.

What a good session looks like

- **One objective** per conversation; start fresh for the next
- Give it the **real documents**, not descriptions of them
- Keep durable facts in a **notes file** the agent always reads
- Keep task-specific procedures in **small playbooks**, loaded when needed
- **Verify the output** — spot-check citations, numbers, and extractions
- Nothing sensitive goes in without an **institutional green light**

Try it this week

Start small and concrete — with material you already have:

- **A DOI list → a table**: your last review's references, five fields each, then spot-check two rows
- **Literature synthesis**: recent papers on your target or pathway → one page, cited, with disagreements flagged
- **Methods triage**: which of these 20 papers used 10x v3 vs. Smart-seq2, and at what depth?
- **Edit with pushback**: "tighten this Aims page and flag weak claims"
- **Draft from your outline**: turn notes into a deck or a first-draft section

Today's `papers.csv` and prompts are in the handout repo — start from them.

Some models rank higher — all of them are good

Free today — $20 to keep going

- **Start free.** Every major tool has a free tier that covers today's examples.
- **Free tiers are demo-grade** — great for learning, but they hit a wall on sustained real work.
- One honest on-ramp: **Claude Pro — $20/mo** ($17 annual), agent features included, cancel anytime.
- A *one-month rental* — cheaper than a textbook, and often **expensable** (PI, training-grant, or professional-development funds).
- A **fork, not a hurdle**: today needs nothing; *continuing* is worth $20.

Takeaways

> AI agents are most useful when the work depends on **real documents, real folders, and real project context.**

- They **act** on your files — they don’t just answer
- **Tokens and the context window** explain both cost and quality
- Keep context **lean**; keep durable knowledge in a **notes file**
- Treat the agent as a **trainee you supervise** — you sign the note
- Sensitive data needs an **institutional path**, not a personal account

Resources

Handout (this repo) — notes, prompts, today's `papers.csv`, and every link below with full citations.

Papers shown Biomni · *Science* 2026 AI co-scientist · *Nature* 2026 The Virtual Lab · *Nature* 2025 CRISPR-GPT · *Nat Biomed Eng* 2025 Johri et al. · bioRxiv 2026

Tooling modelcontextprotocol.io biocontext.ai/registry — biomedical MCP servers BioMCP — PubMed, ClinVar, cBioPortal, OncoKB awesome-genomic-skills — skills for genomics

And the document that actually governs what you may use: **your institution's AI and data-governance policy.**

Token and context-window figures are mid-2026 and illustrative; exact numbers move fast.

References

1. Liu NF, Lin K, Hewitt J, et al. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*. 2024;12:157-173. doi:10.1162/tacl_a_00638
2. Huang K, Zhang S, Wang H, et al. Autonomous biomedical research with an artificial intelligence agent. *Science*. 2026;393(6813). doi:10.1126/science.adz4351
3. Gottweis J, Weng WH, Daryin A, et al. Accelerating scientific discovery with Co-Scientist. *Nature*. 2026;655(8122):487-496. doi:10.1038/s41586-026-10644-y
4. Swanson K, Wu W, Bulaong NL, Pak JE, Zou J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*. 2025;646(8085):716-723. doi:10.1038/s41586-025-09442-9
5. Qu Y, Huang K, Yin M, et al. CRISPR-GPT for agentic automation of gene-editing experiments. *Nature Biomedical Engineering*. 2025;10(2):245-258. doi:10.1038/s41551-025-01463-z
6. Johri S, Pimenta E, Yates J, et al. Evaluating agentic AI for biological discovery in autonomous and copilot settings. Published online June 9, 2026. doi:10.64898/2026.06.04.729919

Examples from my own work

pubmed-grader.cancerdatasci.org

fda-approvals.cancerdatasci.org

orchestraplatform.org

cfde-atlas.cancerdatasci.org

github.com/seandavi/lifeos-template

README License

An opinionated, agentic life-OS for Claude Code. A lived-in scaffold, not a starter kit — every file shape, skill, and review cadence comes from running it for real, not from speculation.

A "life-OS" is a personal operating system in markdown: a vault for journaling, decisions, projects, notes, and people — paired with daily, weekly, and quarterly review skills that keep it from rotting.

Built for solo knowledge workers, researchers, engineering leads, founders, and PIs juggling many threads — anyone who already lives in markdown and wants an AI to enforce review discipline.

This is an artifact, not a project. Fork freely (CC BY 4.0). Issues and PRs are handled best-effort -- open one if