A 35-Second Off Switch for AI Research, Google Splits Its Chips by Job, and Anthropic Puts a Price on Its IPO

OpenAI's Control Room can halt research model workloads within 35 seconds and fails closed, Google's TPU 8tx and 8i Max target confidential and long-context agent inference, Anthropic reportedly sets a $70-$78 IPO price range, and Vercel's agent-desktop lets coding agents drive desktop apps without screenshots.

Themes: Governance & Safety · Infrastructure & Economics

Note: direct page fetches were blocked today, so this issue relies on search summaries. OpenAI's Control Room figure comes from OpenAI's own post; the IPO terms and lease details are secondary reports and unverified. Check the SEC filing and company pages before relying on them.

1) OpenAI Control Room: A Fail-Closed Off Switch

OpenAI described Control Room, a central system that tracks model workloads across its research infrastructure (reinforcement-learning runs, evaluations, large sampling and data-processing jobs) and stops them when safety or security concerns arise. It works alongside OpenAI's misalignment monitors, and OpenAI says it halts model output within 35 seconds of a decision to stop or pause a job. It is fail-closed: if Control Room is down, disabled or tampered with, every research inference request is denied until it recovers. It covers internal research, not ChatGPT or API customers. Coverage dates the rollout to about November 9; outside observers note OpenAI has not said whether 35 seconds is a maximum, an average or a test result.

The concept, explained simply: Think of the big red stop button on a factory conveyor belt. A fail-open button is one where, if the wire to the button is cut, the belt keeps running. A fail-closed button is wired so that a cut wire stops the belt. Control Room is the second kind: losing contact with the safety system means the machine stops, not that it carries on unsupervised.

Why it matters: Over the past weeks this series covered staged releases, gated cyber models and agent monitoring as a product feature. Control Room moves the same idea inside the lab: being able to stop a misbehaving model quickly, and making 'no supervisor' mean 'no output', is the practical core of AI control. Teams running their own agents can copy the pattern: a central kill switch, a measured time-to-stop, and a default of deny when the guard is missing.

2) Google Splits Its TPU Line by Job

Google added two systems to its eighth-generation TPU family on November 9, joining the TPU 8t (training) and TPU 8i (inference) chips launched in April. TPU 8tx uses the 8t silicon with hardware-level encryption for confidential inference and light training, pairs with AMD EPYC host CPUs, claims up to 2x lower latency than 8t, and can be deployed by select customers in their own data centres. TPU 8i Max packages two 8i chips together, doubling compute, on-chip SRAM and memory bandwidth for long-context agent inference; one report puts its high-bandwidth memory at 384 GB. No independent benchmarks or prices were found.

The concept, explained simply: A general-purpose chip is a family car. Google is now building a removal van (training), a taxi (fast inference), an armoured car (8tx: data stays encrypted even from the operator) and a long-haul truck (8i Max: lots of memory for agents that keep huge documents and long conversations 'in mind'). Confidential computing means data is encrypted not only when stored or sent but also while it is being processed, so even the cloud provider cannot peek.

Why it matters: Agents re-read long histories on every step, so memory capacity and bandwidth, not raw compute, are often the bottleneck. Specialised silicon is how Google competes with NVIDIA on cost per token. And on-premises encrypted TPUs answer the question that blocks many banks, hospitals and governments: 'can we use frontier models without handing our data to someone else's computer?'

3) Anthropic Prices Its IPO; a Compute Deal Wobbles

An amended S-1 filed around November 10 reportedly sets a range of $70-$78 per share for 825 million Class A shares on Nasdaq under the ticker ANT, implying a fully diluted valuation of roughly $1.3-1.4 trillion at the top end; pricing is expected around November 18. The same filing reportedly says a roughly $34.9 billion compute agreement with Lambda became terminable by either side after Lambda missed a deadline to secure a facility lease (DCD earlier reported the Lambda deal as $41.3 billion through 2034); Anthropic reportedly says it has no plans to walk away. Total disclosed cloud and data-centre commitments are reported at about $895 billion. A digest also reported new Claude office tools announced in San Francisco on November 10; this could not be confirmed from a second source.

The concept, explained simply: A price range is the shop putting a sticker on the box before opening day: bankers ask big investors what they would pay, then set the final price inside (or outside) the range. A 'terminable' contract is like a flat lease with a clause saying either side can walk away if the landlord never finishes the building. Nothing has been cancelled, but the commitment is no longer locked in.

Why it matters: The S-1 is the most detailed public picture yet of what a frontier lab costs to run: nearly a trillion dollars of compute commitments set against what public investors will pay. The Lambda clause shows the physical risk under those numbers: AI capacity depends on someone actually getting buildings, power and leases on time. For anyone planning on a provider's capacity, delivery dates matter as much as the headline deal size.

VIRAL / OPEN-SOURCE SPOTLIGHT

Vercel Labs agent-desktop, desktop control for AI agents

agent-desktop is an open-source (Apache-2.0) Rust command-line tool and C library from Vercel Labs that lets AI agents read and operate native desktop apps on macOS, Windows and Linux. Instead of taking screenshots and guessing where to click, it reads each app's accessibility tree (the structured list of buttons, fields and menus that screen readers use), gives every interactive element a stable reference, and lets the agent act on that reference. Every command prints one JSON result and exits, so it works with Claude Code, Codex or any agent that can run shell commands. It contains no model of its own; for web pages Vercel points to its sister tool, agent-browser.

Why it is taking off: version 1.0 shipped on October 28 and reports put it at about 7,200 GitHub stars by November 2, near the top of November's trending lists. Screenshot-based computer use is slow, expensive in tokens and brittle when a window moves; reading the accessibility tree is faster, cheaper and more precise. Caution: a similarly named project (lahfir/agent-desktop) exists, so check you have the right repo, and remember that granting accessibility permissions gives an agent real control over your machine.

MARKET SIGNAL

The frontier-AI trade is turning from private rounds into public prices. Anthropic's reported $70-$78 range, with a valuation of roughly $1.3-1.4 trillion against about $895 billion of compute commitments, will be the first time public markets directly price a frontier lab. In the same week Google widened its custom-chip line to four TPUs, a reminder that the hyperscalers are also building their own chips to lower their cost per token. The signal: the market will reward labs that can show revenue growing faster than compute commitments, and will price delivery risk (leases, power, chips) into every 'trillion-dollar' figure.

PRACTICAL TAKEAWAYS

1) Make your agent guard fail closed.
If your monitoring or policy service is unreachable, deny agent actions by default instead of letting them run unsupervised.

2) Measure your time-to-stop.
Test how many seconds it takes to halt a running agent job in practice, and set a target the way OpenAI did.

3) Match the chip to the job.
Compare training, low-latency, confidential and long-context inference options separately instead of buying one generic tier.

4) Ask about confidential computing.
For regulated data, ask your AI provider whether data stays encrypted while being processed, and whether on-premises options exist.

5) Read contract delivery clauses.
When you depend on a provider's capacity, check which deadlines let either side walk away, not just the headline size.

6) Prefer structure over screenshots.
For desktop or browser automation, try accessibility-tree tools like agent-desktop before screenshot-based computer use.

SELF-TEST

Q1. What does 'fail-closed' mean for OpenAI's Control Room?

A1. If Control Room is down, disabled or tampered with, all research inference requests are denied until it recovers.

Q2. What does OpenAI's 35-second figure measure?

A2. The time from a decision to intervene until model output from that research job is halted.

Q3. How do TPU 8tx and TPU 8i Max differ?

A3. 8tx adds hardware encryption for confidential inference; 8i Max doubles the 8i's compute and memory bandwidth for long-context agent inference.

Q4. What is confidential computing?

A4. Keeping data encrypted even while it is being processed, so the infrastructure operator cannot read it.

Q5. Why does agent-desktop use the accessibility tree instead of screenshots?

A5. It gives agents structured, stable references to buttons and fields, which is faster, cheaper and more reliable than guessing from pixels.

TOMORROW

What We're Watching: final pricing and first-day trading of Anthropic's IPO around November 18-19. What We're Watching: whether OpenAI publishes how the 35-second Control Room figure was measured. What We're Watching: pricing and availability for TPU 8tx and 8i Max. What We're Watching: whether Anthropic and Lambda confirm or renegotiate their compute agreement.

TOPICS COVERED SO FAR (RECAP)

Recent days covered the Fable/Mythos/Haiku 5.x model refresh, gated cyber releases from Google and OpenAI, Anthropic's S-1, attacks on AI developer platforms and secrets hygiene, soft versus hard AI law in

Australia, AI screening in cardiology, AI rebounding in the workforce, agent gateways, GPU-backed debt, the science-usage study, the shared robots.txt opt-out, the Agent Client Protocol and always-on agents. Today added fail-closed kill switches for AI research (OpenAI Control Room), specialised and confidential TPUs, IPO price ranges and terminable compute contracts, and accessibility-tree desktop automation for agents.

Sources: openai.com/index/control-room (Control Room), scalevise.com, runtimewire.com and unite.ai (Control Room coverage), datacenterknowledge.com, constellationr.com, datacenterdynamics.com and aiweekly.co (TPU 8tx / 8i Max), kellyabbott1.substack.com (AI Frontier Daily, Nov 11: IPO range and Lambda clause), datacenterdynamics.com (Anthropic $895.4bn commitments), sec.gov (Anthropic S-1), theneuron.ai (Nov 10-11 digests), marktechpost.com, aicybr.com and aitoolly.com (vercel-labs/agent-desktop). Figures are secondary reports and unverified unless from the company's own page.