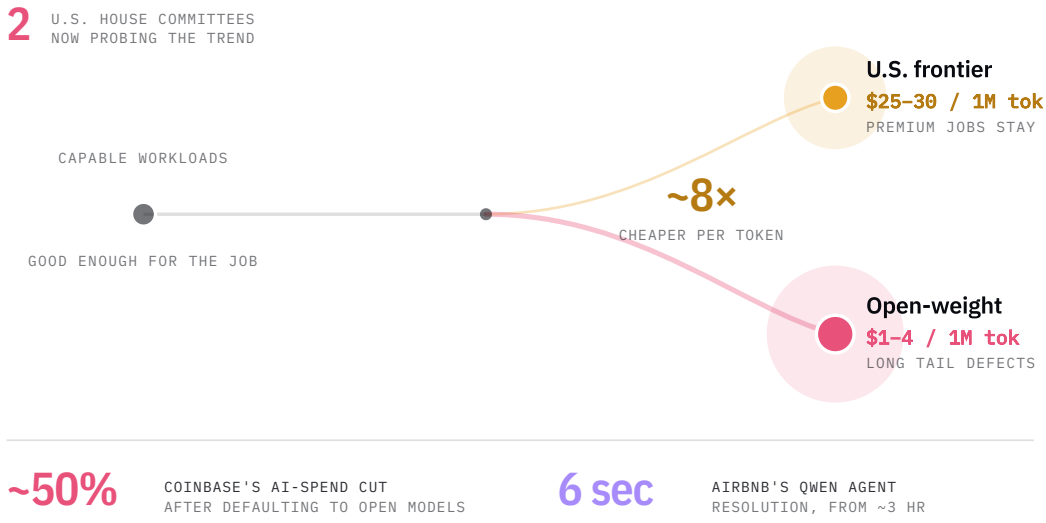
The Quiet Switch

James E. Coleman

STRATEGIC DESIGN · WORKFORCE ACTIVATION · 2026

ABSTRACT *American companies are quietly routing real workloads to Chinese open-weight AI models — Airbnb's customer service to Alibaba's Qwen, Coinbase's engineers to Z.ai's GLM and Moonshot's Kimi — because they are good enough at a fraction of the price. The reflex is to read this as a loyalty problem. It is an accounting decision; and because the weights are open, it is a deployment decision too. Where the weights run, not where they were trained, is what decides whether your data ever leaves — and most of the alarm hasn't caught up to that.*

FIG. — WHERE THE TOKENS ACTUALLY GO — THE LONG TAIL, EATEN FROM BELOW



§ 01 It Was Never About Patriotism

The switch is an *invoice.*

The story that writes itself is a story about sides: American labs versus Chinese ones, loyalty versus defection, a security breach dressed up as a procurement memo. The story the finance teams are actu-

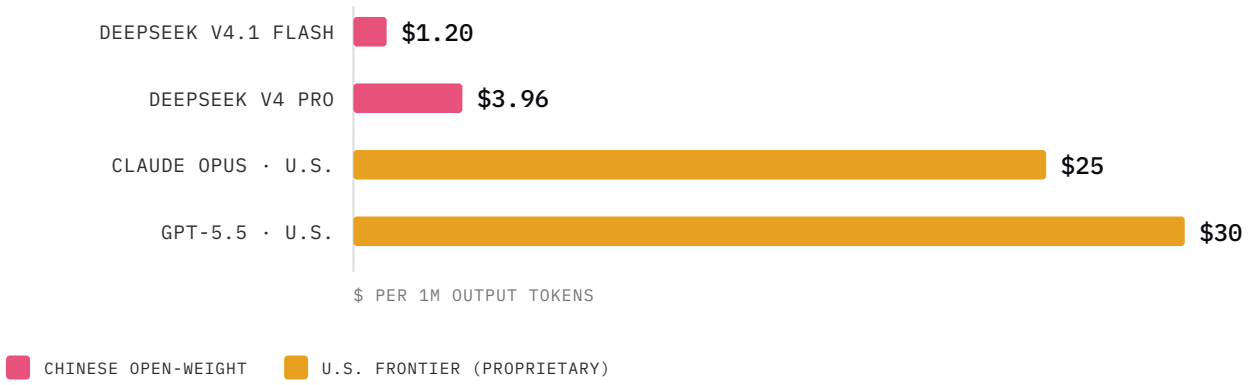
ally living is duller and more decisive. Capable language models have become a commodity, the cheapest capable ones happen to be open-weight and happen to come out of China, and once a model is *good enough* for a given job, the only variable left is price. On that variable the gap is not small.

The public list prices tell most of it. DeepSeek's V4 Pro runs about $3.96 per million output tokens at peak rates; its Flash tier is a small fraction of that. The frontier American models sit around $25–$30 per million output tokens.[10] RAND has estimated that Chinese models can be served at roughly one-sixth to one-fourth the cost of comparable U.S. cloud inference.[13] And the licenses remove the last easy objection: Qwen ships under Apache 2.0, DeepSeek under MIT — terms that explicitly permit commercial use, so there is no contract to renegotiate and no seat to buy.[13]

That is why the defection starts at the bottom of the market and climbs. The investor Martin Varsavsky put the mechanism plainly: *good enough at one-tenth the price becomes the default for everyone except the biggest buyers.*[11] Startups move first because margin is survival; then the internal tooling of larger firms, where a code-review bot or a summarizer does not need a frontier brain. The Western labs built the best models in the world and priced them for enterprises. The long tail — where most of the tokens actually get spent — is being eaten from below.

THE PRICE GAP **FIG. 1**

OUTPUT PRICE — U.S. DOLLARS PER MILLION OUTPUT TOKENS, LIST PRICE



List API prices for output tokens, late 2026; peak rates where tiers apply. Prices change often and exclude prompt caching, context limits, and real quality differences — a cheaper token is not a like-for-like swap. Sources: refs 10–11.

> *"Loyalty to a vendor becomes a luxury when the same capability is available at a tenth of the price. The question stops being which lab you believe in and becomes which invoice you can defend."*
>
> — THE QUIET SWITCH

§ 02 **Two Companies, Two Switches**

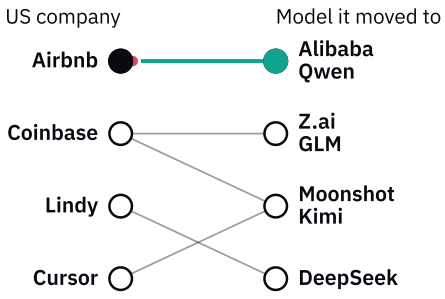
Airbnb and Coinbase *did not switch the same way.*

The two most-cited examples are real, public, and — this matters — different in kind. Conflating them is how the debate goes wrong.

Airbnb — customer service. CEO Brian Chesky has said the company's AI customer-service agent "relies heavily" on Alibaba's Qwen, which he called "very good," and fast, and cheap. The agent is assembled from roughly 13 different models; OpenAI's are in the mix but, in Chesky's words, "we typically don't use them that much in production because there are faster and cheaper models."[1] The operational result Airbnb reports is the part that travels: average resolution time fell from nearly three hours to about six seconds.[1] Chesky has since argued the U.S. is "misunderstanding" the arrangement — the models run through approved U.S.-based providers, and, he says, Alibaba does not "have access to any data."[2]

Coinbase — internal engineering. In June 2026, CEO Brian Armstrong described keeping AI spend flat while token usage grew — not with caps and spend alerts, but with "better defaults, routing, and caching."[4] In practice that meant routing engineers, by default, to open-weight Chinese models — Z.ai's GLM-5.2 (from the Beijing lab formerly called Zhipu) and Moonshot's Kimi — through an internal gateway, for routine work like code review and summarization. Engineers can still pick any model for a hard task. The reported effect: AI spend cut nearly in half while usage kept climbing.[5] It is worth being precise about what this is: a change of *default*, not a wholesale dump — and aimed at engineering, not at customers.

FIG. 2 — WHO SWITCHED, AND TO WHAT STATE 1 OF 4



CUSTOMER SERVICE

Airbnb

Customer-service agent runs largely on Alibaba's **Qwen** inside a ~13-model stack; OpenAI kept for a minority of calls. Resolution time ~3 hours → ~6 seconds. Now under congressional probe.[1]

INTERNAL ENGINEERING

Coinbase

Engineers **default** to Z.ai's GLM-5.2 and Moonshot's Kimi via an internal gateway; AI spend roughly **halved** while token use grew. Internal engineering, not customer-facing.[5]

API TRAFFIC

Lindy

A U.S. automation startup reported moving API traffic from **Anthropic to DeepSeek**, saving "millions." (Founder commentary, not an audited figure.)[11]

A CODING MODEL

Anysphere / Cursor

Its coding model **Composer 2** was built on systems from **Moonshot AI**; named alongside Airbnb in the House investigation.[6]

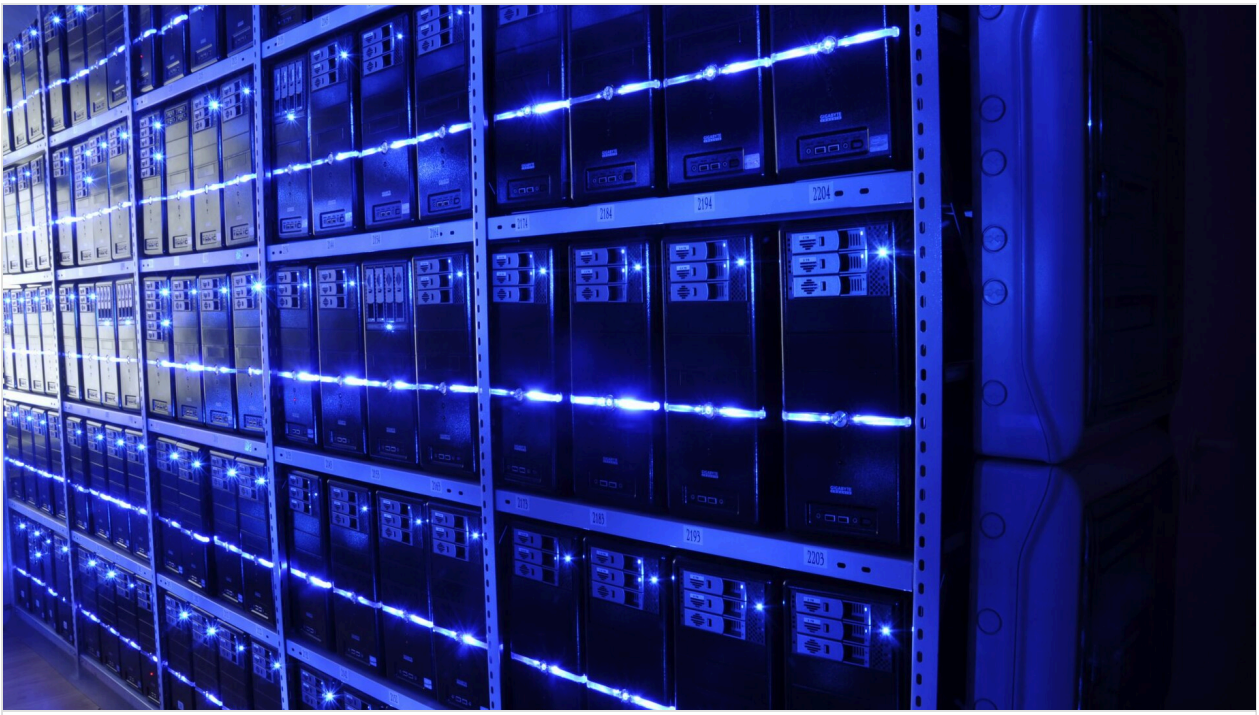
A diagram of the relation, not a chart of data. Each line joins a US company to the Chinese model family the reports say it now uses.

§ 03 The Open-Weight Difference

A service you send data to, or a *file you run.*

The reason any of this is possible — the reason a U.S. public company can touch a Chinese model at all — is that these are open-weight models. The weights are published as a file you can download, inspect, and run. A closed model, American or Chinese, is a service: you send your prompt to someone else's server and trust what happens there. An open-weight model is an artifact: you decide where it executes.

That distinction is not academic, because the big U.S. clouds have made it operational. AWS Bedrock now hosts a lineup of open-weight models — Qwen and others among them — and Google's Vertex AI offers DeepSeek and Qwen as well, so an enterprise can run a Chinese-developed model entirely inside American infrastructure, or on its own GPUs, without a single packet reaching a server in China.[13] This is the same mechanism the companion dispatch, AI Leaves the Cloud, traced for private AI: once the model is a file rather than a feed, *where it runs becomes a separate decision from who built it.*



An open-weight model is a file you place — on a U.S. cloud or your own racks — so its weights can run where no packet reaches China.

Photo: BalticServers.com · CC BY-SA 3.0

> *"A closed model is a service you send your data to. An open-weight model is a file you run where you choose. That one difference is the whole governance story — and it is why the nationality of the lab stopped settling the question."*
>
> — THE QUIET SWITCH

§ 04 Where the Panic Is Right

The weights are open. The *worries are real.*

None of that makes the security objection empty — it makes it specific. There are two distinct risks, and the public argument keeps blurring them. The first is data exfiltration: if you call a Chinese-hosted API, your prompts, your code, and your business data land on servers in China. Under the PRC's 2017 National Intelligence Law, firms there are obliged to "support, assist, and cooperate with" state intelligence work — which is exactly why lawmakers treat a Chinese API endpoint as a reporting line, not just a vendor.[7]

The second risk survives even when the data never leaves: model behavior. Weights can encode the values of the system that trained them. The House committees' stated concern about Qwen is that a

model could "covertly censor and manipulate information pursuant to Chinese law."[6] Self-hosting fixes the first risk and does nothing for the second — a model you run on your own metal can still shade an answer about Tiananmen, Taiwan, or a sanctioned entity.

This is now a live compliance exposure, not a thought experiment. On April 29, 2026, the House Homeland Security Committee and the House Select Committee on the CCP opened a joint probe into Airbnb and Anysphere (Cursor's parent), demanding to know which Chinese models they use, why, and whether the models were independently security-evaluated.[6][8] A law-firm advisory put the lesson bluntly for everyone else: cheaper Chinese AI can invite costly congressional scrutiny, and the defense is documentation.[9] Lawmakers are now weighing curbs — though analysts note an outright ban on models anyone can download is likely unworkable.[12]



On April 29, 2026, two House committees opened a joint probe into U.S. firms' use of Chinese open-weight models.

Photo: peakfocus.agency · CC BY 4.0

§ 05 Deployment, Not Nationality

The question isn't the passport. *It's the wiring.*

The useful reframe — and the one most of the coverage misses — is that the security of a model turns on its deployment architecture, not on the nationality of its developer.[13] Three questions separate the real exposure from the flag on the box:

One — where do the weights run? A Chinese-hosted API, a U.S. cloud, or your own hardware. This decides the exfiltration risk, and only this. Two — what data touches it? Public, low-stakes content, or regulated and customer data. This decides the blast radius. Three — does covert behavior matter for the task? Summarizing internal code tolerates a biased model; answering customers on sensitive topics does not. This is the risk self-hosting cannot remove.

Run the two cases through that grid and the congressional heat map explains itself. Coinbase sits low on all three: open-weight defaults it can self-host, pointed at internal engineering, on content where a shaded answer is nearly harmless. Airbnb sits higher: a model in front of *customers*, touching their data, on exactly the kind of open-ended conversation where "covertly censor and manipulate" stops being abstract — which is why it, not Coinbase, drew the subpoena. The honest caveat: there is no public evidence yet on whether self-hosting open weights satisfies the committees, and the behavior risk is real even when the data stays home.

DEPLOYMENT, NOT NATIONALITY **FIG. 3**

THE SAME THREE QUESTIONS, RUN AGAINST THE TWO PUBLIC CASES

THE QUESTION	COINBASE — ENGINEERING	AIRBNB — CUSTOMER SERVICE
1 · Where do the weights run? Decides exfiltration risk.	CONTROLLED — Self-hostable open weights, internal gateway.	CONTROLLED — Approved U.S.-based providers.
2 · What data touches it? Decides the blast radius.	LOW — Internal code & summaries.	HIGH — Live customer conversations.
3 · Does covert behavior matter? The risk self-hosting can't remove.	NO — A shaded code answer is near-harmless.	YES — Open-ended, sensitive, customer-facing.
Outcome		

NO SUBPOENA	CONGRESSIONAL PROBE

Both firms keep exfiltration low; the divergence is Q2–Q3 — whose data, and whether a shaded answer costs anything. Nationality of the lab appears nowhere in the grid. Framework: ref 13; cases: refs 1, 5, 6.

So the governance posture almost writes itself. Keep anything sensitive on an approved U.S. provider or your own infrastructure; evaluate the weights for behavior before they touch a customer; and keep a model-portability layer — a gateway or router — so that no single lab, American or Chinese, is load-bearing. The firms doing this are not choosing a side. They are refusing to let the choice be permanent.

§ 06 What It Does to the U.S. Labs

Good enough, at a fraction of the price, *eats the long tail.*

Zoom out and this is a margin story about OpenAI and Anthropic as much as a governance story about their customers. The frontier labs keep the top of the market — genuine frontier capability, regulated enterprises, and brand-sensitive buyers for whom provenance is itself the product. What they lose is the price-sensitive long tail: startups, internal tooling, and the high-volume, low-stakes inference that is most of the tokens by count.

The signals are already in the open. Lindy moved API traffic off Anthropic to DeepSeek and reported saving millions;[11] Microsoft has been reported exploring DeepSeek as a lower-cost option within Copilot, and HSBC, Standard Chartered, and Saudi Aramco have reportedly tested or deployed DeepSeek.[13] The scale underneath is the real news: Chinese open-weight models have crossed roughly three billion cumulative downloads, with Qwen and DeepSeek among the most-pulled open weights anywhere — a reversal that took about twelve months.[14] Fortune's framing of the three labs behind it — Moonshot, Z.ai, and DeepSeek — is that they are not just challenging the American labs but "beating them on cost."[5]

The threat to the U.S. labs was never that China would ship a better model. It is that *good enough, at a fraction of the price, becomes the default for everyone but the biggest buyers* — and defaults, once they move, are sticky.

§ 07 What's Coming

A policy chase, and a *portfolio discipline.*

Two things happen next, on two clocks. Washington keeps chasing. Expect more probes, disclosure demands, and sector-specific curbs — for federal systems, critical infrastructure, and regulated data — rather than a blanket wall, because you cannot meaningfully ban a file that anyone can download and run offline.[9][12] The compliance surface, not the model, is where the fight actually lands.

Companies get disciplined. The organizations that come through this cleanly will stop treating "which AI" as a loyalty test and start treating it as a procurement-and-governance decision made per workload: where it runs, what data it touches, whether behavior matters. Keep the frontier U.S. labs for the frontier and for anything where regulation or brand makes provenance load-bearing. Use cheap open weights — American like Llama or Chinese like Qwen, DeepSeek, GLM, and Kimi — behind an approved-provider or on-prem boundary for the high-volume, low-stakes majority. The flag on the lab and the path of the data have come apart; the firms that write that down win on both cost and control.

> *"Don't ask whether a model is Chinese. Ask where its weights run, what data you hand it, and whether a shaded answer would cost you anything. Answer those three and the geopolitics turns back into a wiring diagram."*
>
> — JAMES E. COLEMAN, THE QUIET SWITCH

§ R References — Primary Sources

Sources *& notes.*

Every quantitative claim is cited to a primary statement or an established outlet below. API prices are list prices that move frequently and exclude caching and quality differences. Some adoption figures are trade-press estimates and are labeled as such; posts on X are attributed as the named individual's own statement or commentary. Where a claim could not be corroborated — for example, reports that Coinbase uses Chinese models for *customer service* — it is omitted: the verified Coinbase story is internal engineering.

1. Airbnb's AI customer-service agent relies heavily on Alibaba's Qwen within a ~13-model stack; CEO Brian Chesky called it "very good," fast and cheap, and said OpenAI models are used sparingly in production; resolution time fell from ~3

hours to ~6 seconds. South China Morning Post, "Airbnb picks Alibaba's Qwen over ChatGPT." scmp.com.

2. Chesky says the U.S. is "misunderstanding" Airbnb's use of Chinese AI — models run via approved U.S.-based providers and, he says, Alibaba does not "have access to any data." Bloomberg Law. news.bloomberglaw.com.

3. Forbes, "Airbnb CEO Brian Chesky Called Chinese AI Fast And Cheap. Now, Congress Wants Answers" (May 21, 2026). forbes.com.

4. Brian Armstrong (Coinbase CEO), post on X, June 2026: keeping AI spend flat as token usage grows "not with friction and spend alerts… with better defaults, routing, and caching." x.com/brian_armstrong.

5. Coinbase defaulted engineers to open-weight GLM-5.2 (Z.ai / formerly Zhipu) and Kimi (Moonshot) via an internal gateway, cutting AI spend nearly in half while token usage grew; internal engineering, not customer-facing. MLQ.ai, "Coinbase Switches to Chinese AI Models GLM and Kimi." mlq.ai; Fortune, "China's Moonshot, Z.AI, and DeepSeek are challenging U.S. AI labs—and beating them on cost." fortune.com.

6. House Homeland Security Committee and House Select Committee on the CCP opened a joint probe (April 29, 2026) into Airbnb (Qwen) and Anysphere / Cursor (Composer 2, built on Moonshot), citing the risk that a model could "covertly censor and manipulate information pursuant to Chinese law." Semafor, "House committees probe Cursor parent, Airbnb over Chinese AI." semafor.com.

7. The probe cites the PRC's 2017 National Intelligence Law, under which Chinese firms must "support, assist, and cooperate with" state intelligence work. Nextgov/FCW, "House panels probe Airbnb, Anysphere over use of Chinese AI models." nextgov.com.

8. OECD.AI Incidents Monitor, "US Lawmakers Probe Airbnb and Anysphere Over Use of Chinese AI Models" (2026-04-29). oecd.ai.

9. McGuireWoods, "Could Using Cheaper Chinese AI Lead to Costly Congressional Scrutiny?" (Sept. 2026) — on disclosure, security-evaluation, and documentation obligations for firms using Chinese models. mcguirewoods.com.

10. DeepSeek API pricing (late 2026): V4 Pro ≈ $1.32 input / $3.96 output per 1M tokens at peak; V4.1 Flash ≈ $0.30 / $1.20. BenchLM. benchlm.ai; Price Per Token. pricepertoken.com. Frontier U.S. list prices (GPT-5.5 ≈ $5/$30; Claude Opus ≈ $5/$25 per 1M in/out) via CloudZero, "DeepSeek pricing 2026." cloudzero.com.

    List prices for output tokens; exclude caching, context, and quality differences. Always confirm against vendors' own pricing pages.

11. Martin Varsavsky, post on X (2026): Lindy moved 100% of its API traffic from Anthropic to DeepSeek; argues "good enough at one tenth the price becomes the default for everyone except the biggest buyers." x.com/martinvars.

    Founder/investor commentary, not an audited financial figure.

12. Rest of World, "Low-cost Chinese AI models like DeepSeek gain traction in the U.S." — adoption drivers and the limits of a blanket ban. restofworld.org.

13. Quartz, "American companies can't stop buying Chinese AI" — AWS Bedrock and Google Vertex AI host DeepSeek and Qwen; a RAND estimate puts Chinese models at roughly one-sixth to one-fourth the cost of comparable U.S. inference; security turns on deployment architecture, not developer nationality; named testers include HSBC, Standard Chartered, Saudi Aramco, and Microsoft (Copilot). qz.com.

Qwen ships under Apache 2.0 and DeepSeek under MIT — licenses that permit commercial use.

14. On the scale of adoption: Chinese open-weight models crossing roughly three billion cumulative downloads, with Qwen and DeepSeek among the most-pulled open weights — "the great open-source reversal." AInChina. ainchina.com.

Download totals are platform estimates (e.g., Hugging Face) and count derivatives; treat as order-of-magnitude.

CLOSING STATEMENT

"The story everyone wants to tell is about sides — American models against Chinese ones, loyalty against defection. The story the invoices tell is quieter: capable intelligence became a commodity, the cheapest capable version happens to be open-weight and happens to come from China, and because it is a file you can run where you choose, the only question that still matters is where you choose to run it. Answer that deliberately, and the geopolitics turns back into a wiring diagram."

— JAMES E. COLEMAN — FIELD DISPATCH · 2026

James E. Coleman

STRATEGIC DESIGN · AI DATA GOVERNANCE, PROCUREMENT & PRIVATE AI

James works with organizations on AI data governance, model procurement, and local & private AI deployment — with a bias toward plain, cited analysis over hype. This dispatch is part of an ongoing series tracking where intelligence is actually allowed to live, who it answers to, and what that costs.