Open Weights, Extracted Capabilities: Reassessing the Distillation Economy and the Contest for American AI Leadership

Keith Lee

Swiss Institute of Artificial Intelligence, Chaltenbodenstrasse 26, 8834 Schindellegi, Schwyz, Switzerland

Abstract

Open-weight AI models have generally been discussed as a binary policy question: should the U.S. regulate them or not. That framing has become inadequate. Between February and August 2026, three separate incidents, Anthropic's disclosure of coordinated extraction by DeepSeek, Moonshot, and MiniMax, its later accusation against Alibaba's Qwen lab and the brief federal order on two of Anthropic's own frontier models-revealed that the harder problem is not whether open weights ought to be permitted but rather how a legal and commercial regime with no established standard of illicit distillation can keep American frontier labs funded while an increasingly powerful open-weight lfield, much of it Chinese and state-backed-closes the gap from below. This paper argues that the most significant challenge to American AI dominance is not open-weight technology per se but rather the lack of an enforceable boundary between legitimate model training and industrial-scale extraction, a gap that current policy proposals only partially close.

1 Introduction - Redefining the Policy Problem

For most of 2023 and 2024, the debate over open-weight artificial intelligence models proceeded along fairly comfortable lines. Meta released Llama, researchers argued about whether open models posed a proliferation risk and the practical stakes seemed distant from the commercial fortunes of Anthropic, OpenAI or Google DeepMind. That calm ended with a sequence of events concentrated in a single eighteen-month window. In January 2025, DeepSeek's R1 model wiped out roughly $600 billion in Nvidia's market value in a single trading session and OpenAI's leadership publicly accused the Chinese lab of building its model by querying and replicating outputs from American systems.[1] A year later, in February 2026, Anthropic published a far more detailed account, naming DeepSeek, Moonshot and MiniMax as having generated more than 16 million exchanges with Claude through roughly 24,000 fraudulent accounts, in what the company called industrial-scale distillation.[2] By June, Anthropic had gone further still, telling the Senate Banking Committee that operators linked to Alibaba's Qwen lab had run 28.8 million exchanges through nearly twenty-five thousand fake accounts over six weeks, a campaign larger than the three previous cases combined.[3] And in July, a Chinese open-weight model, Moonshot's Kimi K3, outscored Anthropic's own Claude Fable 5 on a widely watched coding leaderboard, even as briefly imposed export restrictions forced Anthropic to disable that same model and its more capable Mythos 5 sibling worldwide.[4]

The conventional way of narrating these events treats them as installments in a familiar story: China is catching up and the reason is theft. That story is not wrong so much as incomplete and its incompleteness matters for policy. It obscures the fact that the technique iself, model distillation, has no settled legal status in the United States. Training a smaller model on the outputs of a larger one is standard practice throughout the AI industry, used by Anthropic and OpenAI on their own systems to build cheaper variants for customers.[5] What distinguishes an ordinary customer relationship from what Anthropic calls an attack is not the technique but the mode of access behind it: fabricated accounts, evasion infrastructure and traffic volumes that bear no resemblance to legitimate use. Yet American law offers only an awkward fit for this distinction. Model outputs generated by a machine are not copyrightable under existing doctrine, since copyright requires human authorship, which pushes any legal remedy toward trade secret and computer fraud statutes that were not written with this scenario in mind.[6] The result is a policy vacuum in which the loudest and most detailed account of what has happened so far has come not from courts or regulators but from unilateral corporate disclosures and letters to individual senators.

This paper takes as its starting point the observation that open-weight models are neither an unambiguous public good nor a straightforward security threat and that debating them in those terms misses what is actually at stake. The stakes are structural. If commercial subscription revenue from Claude, ChatGPT and Gemini is what currently funds the enormous compute expenditures behind frontier model training. If that revenue can be eroded by competitors who reproduce a large share of the resulting capability at a fraction of the cost, then the sustainability of the American frontier itself becomes contingent on questions that have nothing to do with model design: whether distillation can be detected, whether it can be deterred without banning a legitimate technique and whether the resulting commercial pressure pushes toward a more open AI ecosystem or a more secretive and

consolidated one. The chapters that follow examine, in order, what open-weight models are and why different providers pursue them for different reasons, what would follow if frontier labs lost the commercial position that funds their research, what tools exist or could exist to distinguish legitimate training from extraction and what an older dispute inside the WordPress ecosystem suggests about whether an extraction-heavy relationship between commercial users and an open commons can remain sustainable over time.

2 Open-Weight Models and the Divergence of Provider Incentives

An open-weight model occupies a specific point on a spectrum that runs from fully closed systems, where a user submits a query and receives an answer with no visibility into training data or methodology, to fully open-source systems, where the training data, code and resulting parameters are all disclosed. In between sit models such as DeepSeek's V4 family, Alibaba's Qwen series, Moonshot's Kimi models and Z.ai's GLM line, where the trained weights, the numerical parameters that determine how the model responds to input, are published for anyone to download, inspect, fine-tune or run locally, while the underlying training data and code used to produce those weights are not revealed.[7] This is a meaningfully different proposition from open-source software, where the entire recipe is available and a motivated developer can, in principle, reproduce the product from scratch. With an open-weight model, what is available is closer to a finished dish than a recipe: enormously useful, freely reusable but not something a downstream user could have produced independently without access to the compute and data the first developer possessed.

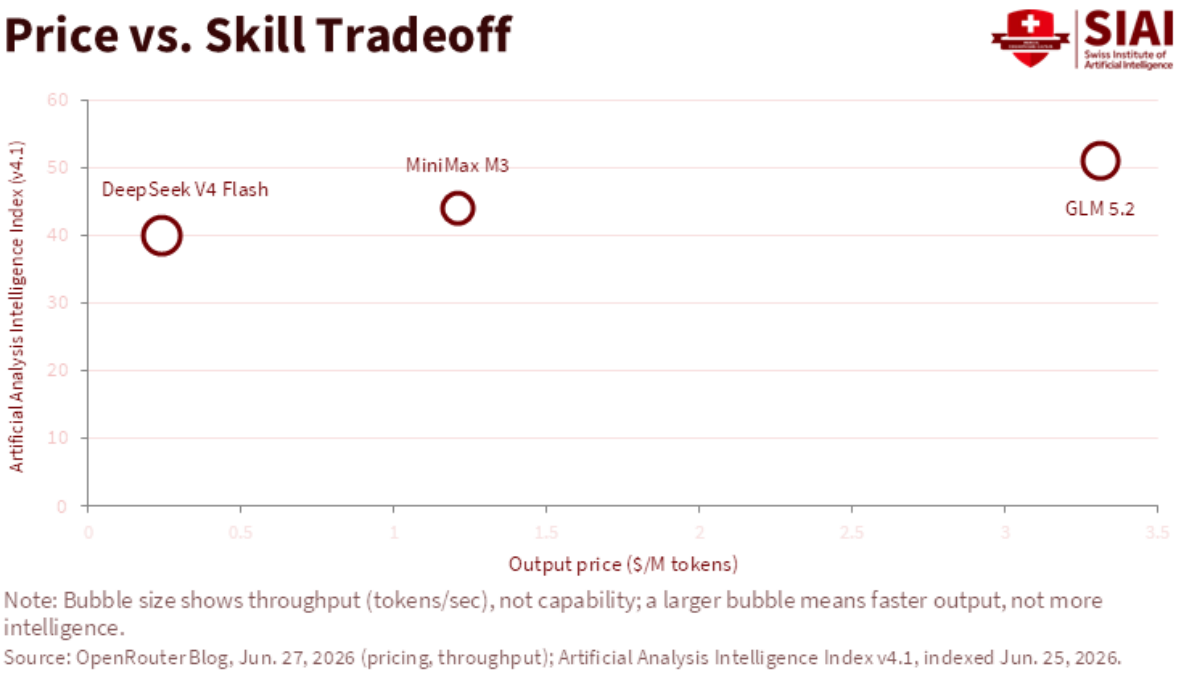

Figure 1: No open-weight model wins on every axis: DeepSeek is cheapest, GLM is most capable, MiniMax is fastest, a genuine three-way tradeoff, not a single leaderboard.

The commercial logic of ordinary open-source software is comparatively well understood. A company gives away code and earns revenue from complementary services: hosting, support, customization, enterprise licensing. Red Hat did this with Linux for two decades. WordPress.org licenses the software freely, while Automattic and

a large ecosystem of hosting providers earn money alongside it. Open-weight AI models complicate this picture because the entities publishing them are not, in most cases, straightforwardly optimizing for adjacent revenue. As of mid-2026, the open-weight frontier is dominated by a small set of providers whose incentives diverge sharply. DeepSeek's V4 Flash model, released in April 2026, reached seventy-nine percent on the SWE-bench Verified coding benchmark, within two points of its own larger V4 Pro variant, while pricing output tokens at roughly a tenth of a cent per thousand or about one hundred fifty times cheaper than a comparable American closed model on a per-token basis.[8] Z.ai's GLM 5.2, released two months later, briefly became the top-ranked open-weight model on the Artificial Analysis Intelligence Index, only a few points behind Anthropic's own Claude Fable 5. It did so days after the United States had temporarily forced Anthropic to withdraw Fable 5 from the market entirely on national security grounds.[9] MiniMax's M3 model added native image and video understanding at similarly aggressive prices. The one prominent American entrant in this tier, Nvidia's Nemotron 3 Ultra, trailed the leading Chinese models on most benchmarks but offered something the others could not: a fully domestic supply chain, an open training recipe and a vendor whose institutional incentive to sustain an open ecosystem has nothing to do with subscription revenue and everything to do with selling the chips that run all of these models regardless of who trained them.[10]

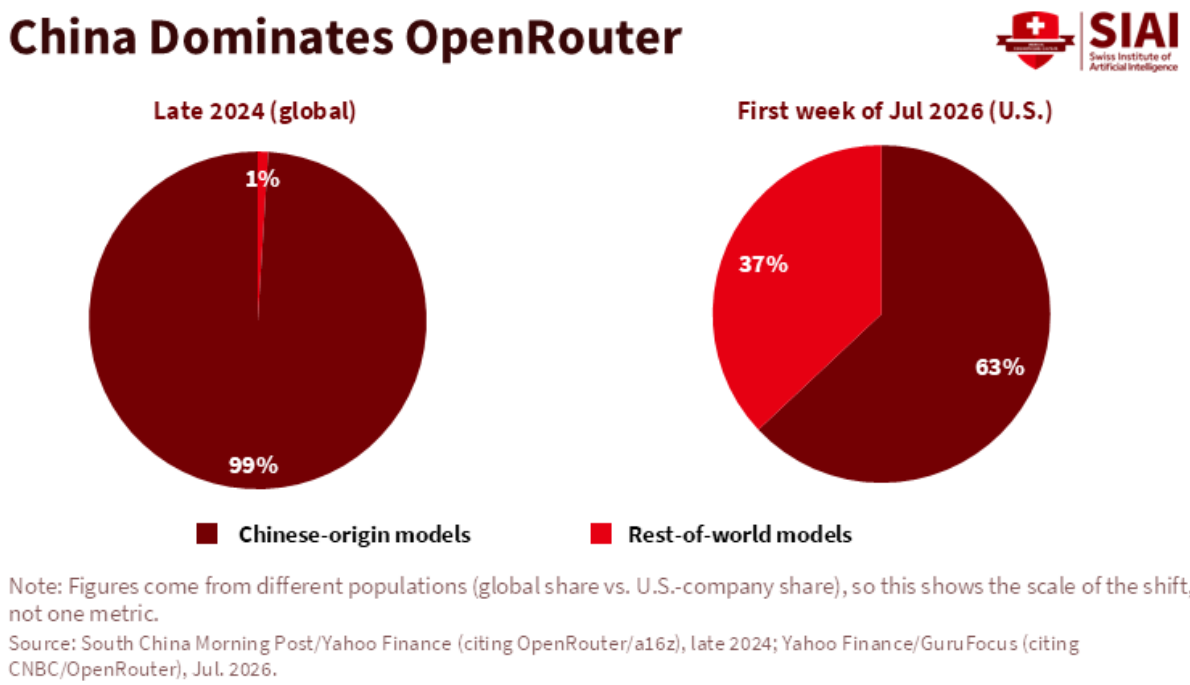

Figure 2: What was a rounding error eighteen months ago is now the majority of measured traffic; the empirical backbone of the "self-reinforcing advantage" claim just made.

These divergent motives matter because they change what "competing with open weights" actually means for an American policymaker or a frontier lab executive. Nvidia's interest in a plural, open ecosystem is plain: more usable models, regardless of origin, mean more demand for the chips that run them and the company has said as much in signing the industry-wide "Open Weights and American AI Leadership" letter alongside more than two hundred and seventy other firms, including Microsoft, Google, Meta and OpenAI.[11] For the Chinese labs, the picture is murkier and contested. One line of analysis holds that Beijing's industrial policy apparatus-

central government guidance funds reported to channel several hundred billion dollars into strategic sectors including AI, provincial computing vouchers and energy subsidies covering as much as half the operating cost of data centers running domestic chips- provides Chinese labs a cost structure no ordinary commercial calculus could replicate.[12] A U.S. congressional advisory body reached a related conclusion in March 2026, describing China's open ecosystem as generating a self-reinforcing competitive advantage and noting that an estimated 80% of American AI startups were already building on Chinese open-weight models by that point.[13]

A competing view, argued forcefully by some industry commentators, holds that the subsidy narrative does less analytical work than it appears to. One widely circulated technology blog pointed out that DeepSeek's V4 model, a frontier-tier system in its own right, was priced at $3.48 per million output tokens against roughly $75 for Anthropic's comparable Opus model, a twenty-one-fold gap that architecture and capability gains alone cannot plausibly explain. It argued that American frontier labs' own claims of near-break-even pricing reflect marketing more than balance sheets, since independent providers hosting the very same open-weight Chinese models manage to remain profitable at a fraction of what the closed labs charge.[14] On this reading, the price gap says less about hidden Chinese subsidy than about the markup closed American labs have been able to sustain, a markup that a truy competitive open-weight tier is now eroding regardless of its country of origin. Both accounts can be simultaneously true in part: state support almost certainly lowers Chinese compute costs at the margin and closed-model pricing in the United States has almost certainly carried more margin than public statements about thin inference economics suggested. What the debate obscures, however, is a third and less examined channel through which cost advantage is achieved, one to which the paper now turns: not subsidy in the conventional sense but the direct transfer of a costly capability from one lab's models into another's, at a small fraction of what building that capability from first principles would require.

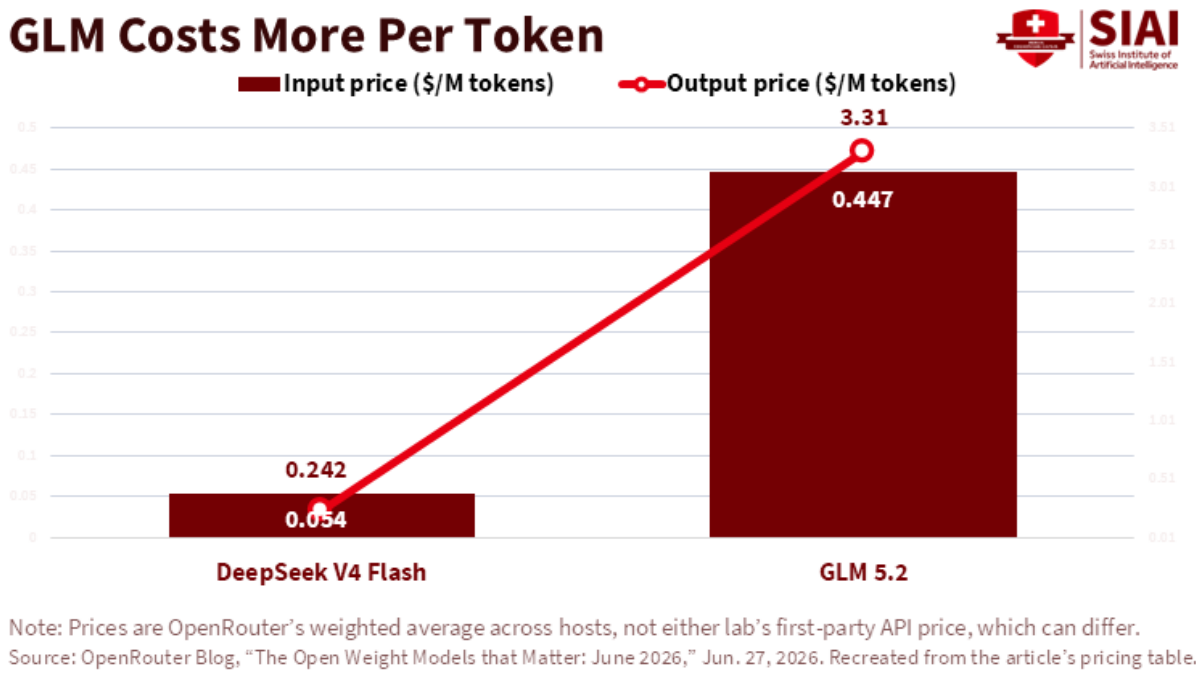

Figure 3: Even within the open-weight tier itself, price varies roughly eightfold; "open" and "cheap" are not synonyms and subsidy alone cannot explain every gap.

3 The Commercial Foundations of Frontier Development at Risk

It is worth taking the counterfactual seriously rather than treating it as rhetorical. Anthropic and OpenAI are, as of mid-2026, both cash-flow-negative enterprises whose commercial viability rests on the expectation that current subscription and API revenue, plus new capital raised against future returns, will fund training runs that are larger and more expensive than the last.[15] If open-weight alternatives, whether cheaper because of genuine efficiency, state support, distilled capability or some combination of the three, continue to close the performance gap while undercutting price by an order of magnitude, the immediate effect is not that frontier labs disappear overnight. Both companies retain substantial enterprise and government business lines that are comparatively insulated from consumer price competition. The more gradual and more consequential effect is on the rate of reinvestment. A frontier lab whose consumer and mid-market revenue is compressed has less capital available to fund the next scaling run and slower scaling compounds: a lab one generation behind the frontier finds it progressively harder to justify the valuation and capital raises needed to close that gap, particularly once investors start pricing in the possibility that whatever the lab builds next will be distilled by a competitor within months of release, as happened to MiniMax when Anthropic detected the lab redirecting a majority of its extraction traffic toward a newly released Claude model within twenty-four hours of that model's launch.[16]

This dynamic is not speculative; its early stages are already visible in market behavior. Startups that had built products atop Anthropic and OpenAI's APIs have begun migrating workloads to Chinese open-weight models hosted through intermediaries such as OpenRouter and Featherless, partly to control costs and partly, according to one economist who studies enterprise AI spending because the increase in these models reveals unmet demand that American providers have been slow to serve at a comparable price point.[17] Financial commentary describing the resulting competitive dynamic depicted as a "death zone", a market in which any provider without either frontier-level capability or market-breaking pricing risks losing share entirely, is not limited to fringe analysts; it appeared in mainstream financial reporting following Alibaba's release of Qwen3.8-Max, a model that appeared to match or exceed Anthropic's own flagship in benchmark performance within days of release.[18] Anthropic's stock-adjacent instruments and OpenAI's private valuation discussions have both shown sensitivity to these developments, most visibly when Anthropic's own June 2026 letter accusing Alibaba of large-scale distillation was followed by a sell-off in Alibaba's own shares, evidence that markets on both sides of the Pacific are now pricing distillation risk directly into competitive strategy.[19]

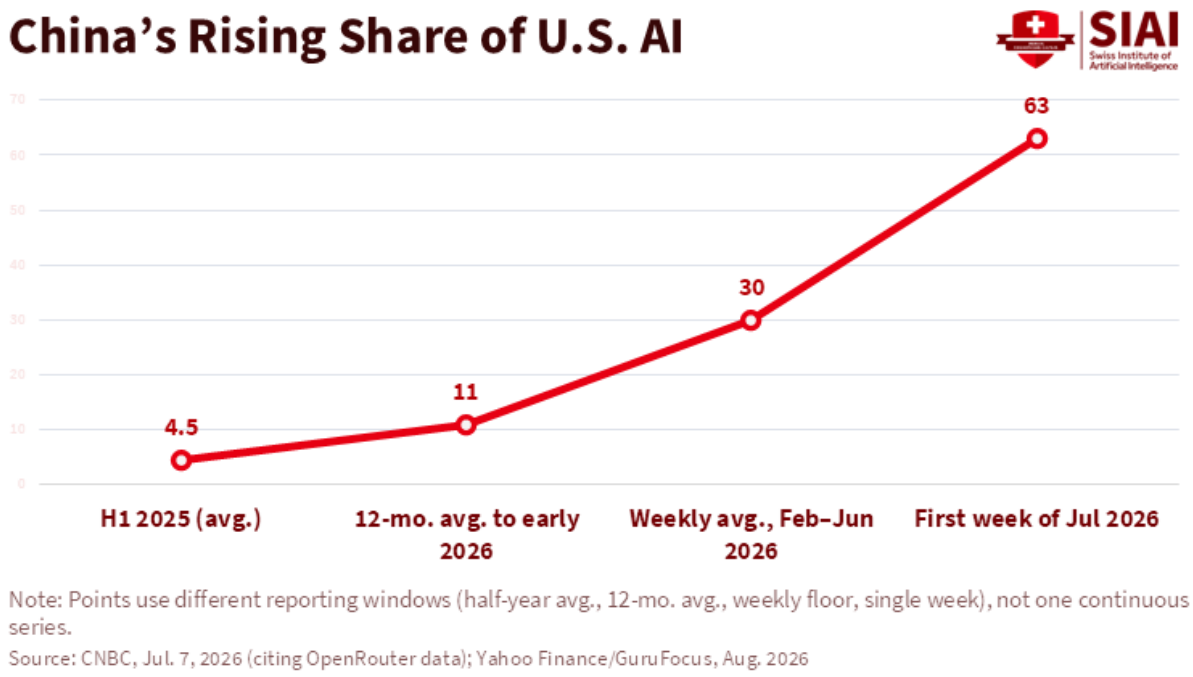

Figure 4: The migration developers describe anecdotally shows up starkly in aggregate: a fourteen-fold jump in twelve months.

The question of who benefits from this dynamic deserves more precision than the shorthand "China wins" typically supplies. One financial commentator, writing about the Alibaba accusation specifically, argued that the act of copying is itself a lagging indicator of who holds the technological lead, since a firm spending 28.8 million queries reverse-engineering a competitor's capabilities is, by construction, choosing to replicate rather than originate and cannot by definition surpass what it is copying through that method alone.[20] There is real force to this point and it complicates any simple narrative in which distillation alone hands China frontier leadership. But it also understates what distillation accomplishes even when it produces only a fast follower rather than a new leader. Amodei's own account of the risk, offered in Anthropic's July 2026 statement of position, is not that distillation lets Chinese labs surpass the American frontier but that it can compress the gap between the Chinese frontier and the American one to a matter of a few months, which is sufficient to blunt the commercial and geopolitical value of maintaining a lead at all.[21] A lead measured in months rather than years does little to reassure investors funding a hundred-billion-dollar training run and it does even less to reassure a Department of Defense weighing whether it can rely on a domestic supplier's exclusivity for any meaningful period.

A more sympathetic reading of the competitive forces and one that deserves engagement rather than dismissal, holds that cheaper AI of any origin expands the total market for AI services faster than it erodes any individual provider's position, an argument sometimes framed in terms of the nineteenth-century economist William Stanley Jevons' observation that greater efficiency in resource use tends to increase total consumption of that resource rather than reduce it. Industry data cited in coverage of the mid-2026 Chinese model releases showed inference prices across the industry falling from roughly two dollars to $1.20 per million tokens within a matter of weeks, alongside American labs cutting their own developer pricing by as much as eighty percent

in response, which some analysts read as evidence that falling costs were expanding the addressable market for AI applications rather than simply transferring revenue from American to Chinese providers.[22] The argument has genuine merit as a description of aggregate industry growth. It is less persuasive as a description of what happens to the specific firms, Anthropic and OpenAI chief among them, whose business models depend on maintaining pricing power at the frontier tier specifically, since an expanding market for AI overall does not by itself guarantee that the revenue funding the next scaling run accrues to the labs currently bearing the cost of frontier research.

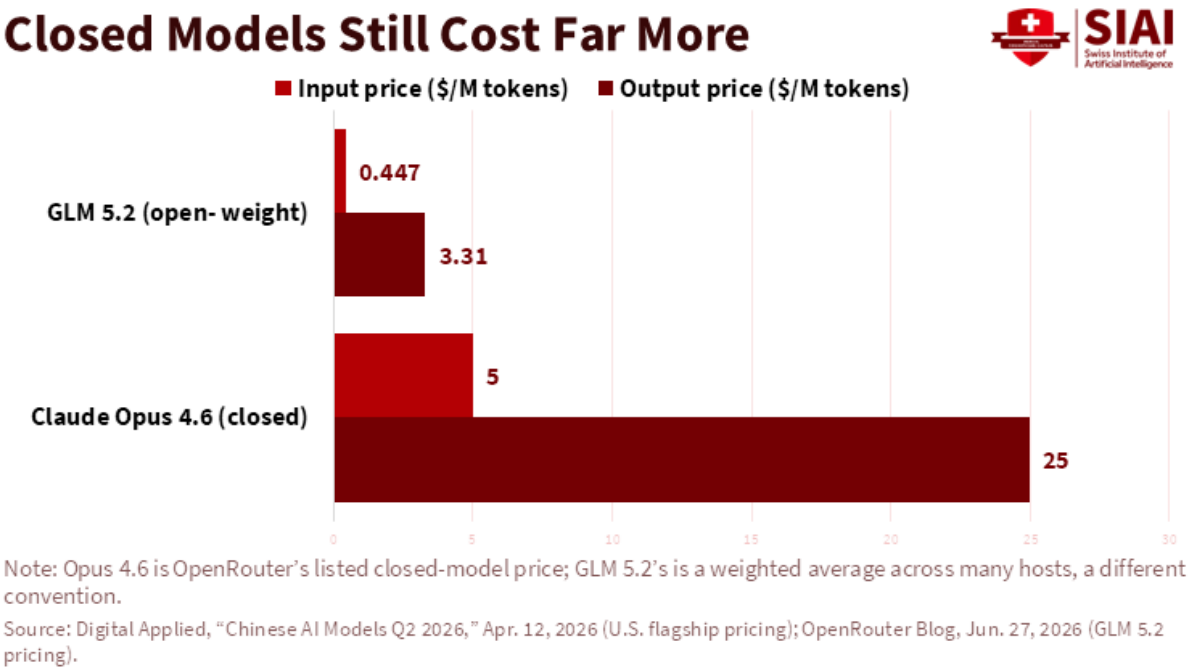

Figure 5: The pricing power frontier labs are fighting to preserve, in one comparison: a closed flagship still costs roughly seven times more per output token.

4 Distinguishing Legitimate Training from Extraction: Detection and Policy Response

If the central danger is not the existence of open weights but the erosion of the commercial position that funds frontier development, the natural policy question becomes whether the underlying extraction can be detected, deterred and distinguished from legitimate training without banning a technique the entire industry depends on. The difficulty starts with definitions. Distillation, in the technical sense of training a smaller "student" model to mimic the outputs of a larger "teacher" model, is used by American frontier labs on their own systems constantly, to produce cheaper variants for customers. Legal scholars broadly agree that nothing in current intellectual property law makes the practice itself unlawful, since a model's output, lacking human authorship, generally falls outside copyright protection.[23] What Anthropic characterizes as an attack is not the technique but the mode of access underlying it: coordinated networks of fraudulent accounts, described by the company as "hydra cluster" architectures, in which a single proxy network can manage more than twenty thousand accounts simultaneously, routing traffic through third-party cloud platforms to evade the regional

access restrictions that prevent commercial use of Claude within China.[24] That distinction between distillation as a widely practiced training method and distillation carried out through systematic fraud is analytically clean but legally underdeveloped. A breach of a company's terms of service is ordinarily a civil matter with limited remedies; the fabrication of identities and purpose-built evasion infrastructure to sustain unauthorized access moves the conduct toward statutes such as the Computer Fraud and Abuse Act and federal wire fraud provisions. However, no such case had yet been tested in court as of this writing.[25]

Table 1: **Legitimate Distillation Versus Extraction-Pattern Distillation**

DIMENSION	LEGITIMATE DISTILLATION (STANDARD PRACTICE)	EXTRACTION-PATTERN DISTILLATIO (ALLEGED ATTACK)
Access method	Ordinary paid API access under standard terms of service	Coordinated networks of fabricated accounts routed through proxy infrastructure
Query volume and pattern	Varied, consistent with ordinary product development or research use	Narrowly concentrated on a rival's most differentiated capabilities, at industrial scale
Stated purpose	Producing smaller, cheaper variants of a lab's own models, or legitimate research and evaluation	Systematically reconstructing a competitor's proprietary capabilities without comparable investment
Legal exposure	Governed by ordinary contract and intellectual property law; not itself unlawful	Plausible exposure under the Computer Fraud and Abuse Act and federal wire fraud statutes, though untested in court
Example from the record	Anthropic's and OpenAI's own practice of distilling frontier models into cheaper variants for customers	The campaigns Anthropic attributes to DeepSeek, Moonshot, and MiniMax (Feb. 2026) and to Alibaba's Qwen lab (Jun. 2026)

Note: An analytical distinction, not a settled legal standard; accused firms have denied wrongdoing in every case so far.
Source: Compiled from Anthropic, Feb. 23, 2026; War on the Rocks, Aug. 3, 2026.

The two disclosed episodes illustrate both the scale of the problem and the limits of unilateral corporate detection. In the February campaign, Anthropic attributed the DeepSeek portion specifically to internal reasoning-elicitation prompts designed to extract Claude's chain-of-thought training data at scale and traced Moonshot's activity through metadata that matched the public profiles of the lab's own senior staff.[26] In the case attributed to MiniMax, Anthropic disclosed that it had detected the campaign while it was still active, before the resulting model had even launched, only to watch MiniMax redirect roughly half its extraction traffic within a day of Anthropic releasing a newer system[27] The June episode attributed to Alibaba's Qwen lab was, by Anthropic's own account, larger than the three February campaigns combined and specifically targeted the software engineering, agentic reasoning and cybersecurity capabilities embodied in Anthropic's most advanced frontier system.[28] Alibaba has not addressed the specifics of the allegation and no independent party has verified Anthropic's figures; the entire account rests on the disclosing company's own forensic analysis, a point War on the Rocks' Ryan Fedasiuk raised as a genuine due-process concern, since sanctioning a foreign firm on the strength of an accuser's internal investigation, however credible, sets an uncomfortable precedent for ad hoc adjudication in a market this consequential.[29] Independent trade press coverage of the same disclosure reported matching figures, lending at least secondhand corroboration to Anthropic's numbers even as the underlying

characterization remains contested.[30]

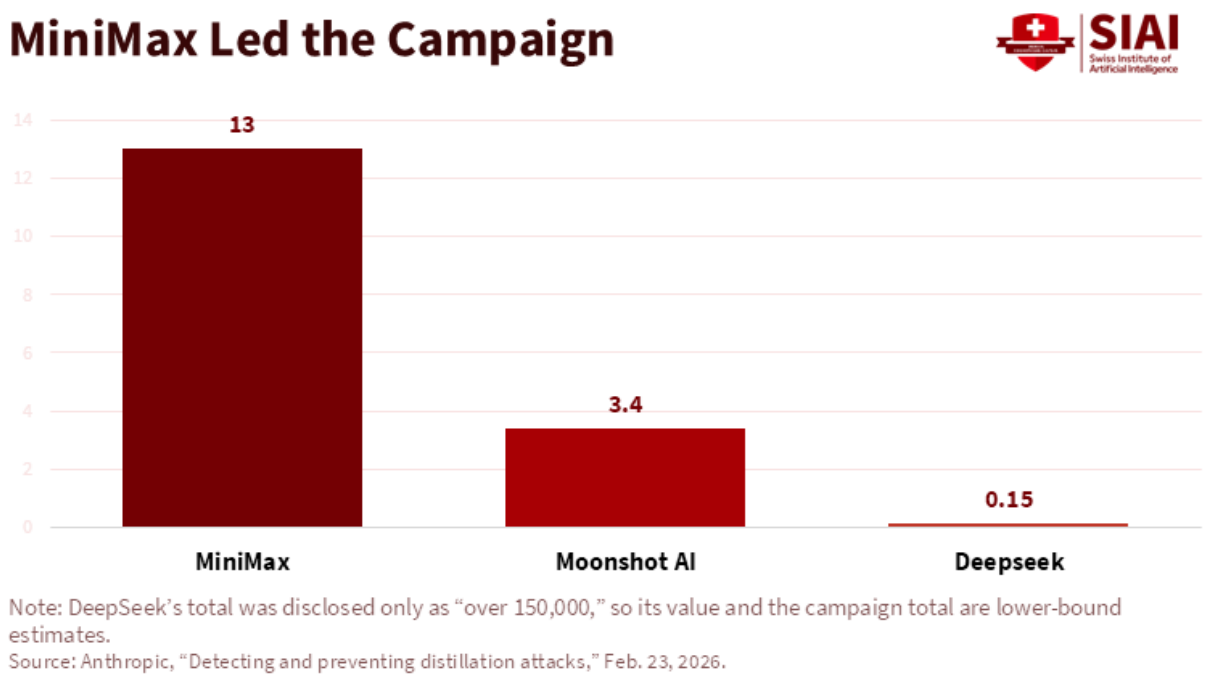

Figure 6: Within the disclosed February campaign, one lab did most of the extracting: MiniMax alone accounts for nearly four-fifths of it.

China's own government has responded in kind rather than engaging the specifics. In July 2026, its Ministry of Commerce accused unnamed American firms of distilling Chinese models. It threatened unspecified retaliatory measures should Washington impose sanctions, offering no company names and no supporting evidence, a mirror-image accusation that suggests both governments now treat the distillation dispute as leverage in a more extensive negotiation rather than as a discrete legal question awaiting resolution[31] This symmetry is not entirely new. When DeepSeek's R1 model first triggered scrutiny in early 2025, the incoming Trump administration's AI advisor David Sacks stated there was "substantial evidence" that DeepSeek had distilled OpenAI's models, a claim OpenAI's own memo to Congress echoed a year later regarding continued circumvention of its access controls.[32] What changed between 2025 and 2026 was not the basic dispute but its scale, its formalization in detailed technical disclosures and its entanglement with a parallel and more explosive fact: Anthropic itself had settled a $1.5 billion lawsuit in September 2025 for training its own models on pirated books, a settlement critics on both sides of the debate have cited as evidence that frontier labs' complaints about unauthorized capability extraction sit uneasily alongside their own record on intellectual property.[33] Some commentators went further, characterizing Anthropic's distillation disclosures as self-serving alarm dressed up as a national security concern, one prominent critic mocking the idea that a company simultaneously marketed as a technically formidable frontier lab and as a victim unable to prevent unauthorized extraction of its own product.[34]

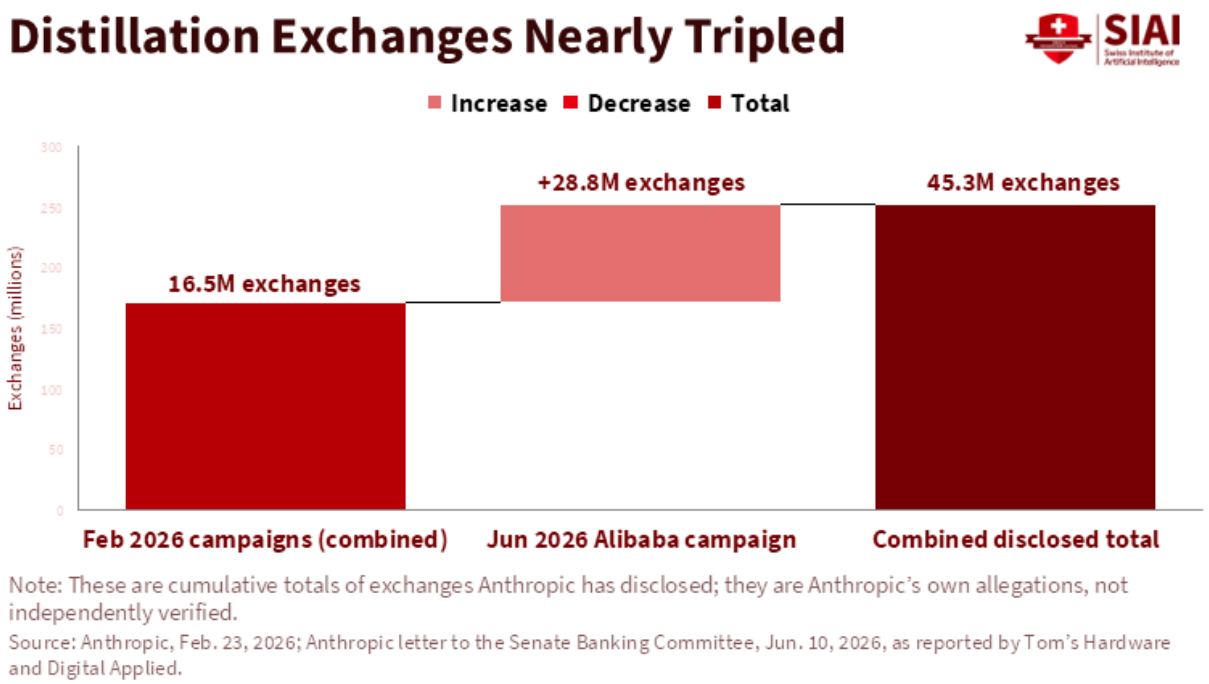

Figure 7: Set side by side, the two campaigns read less as a steady drumbeat than an escalation, February's "industrial scale" looks modest next to June's.

Set against this contested backdrop, the policy response that has gained the most traction, articulated most fully in War on the Rocks and repeated in Anthropic's own stated position, rests on four related moves. First, responsibility for verifying an extraction claim should not remain solely with the accusing company; an independent body, potentially the U.S. Center for AI Standards and Innovation, would need to develop the technical capacity to confirm or disconfirm a distillation attack claim, separating conduct that constitutes fraud from conduct that merely constitutes training on data a firm made available, however reluctantly, to the public.[35] Second, sanctions targeting the specific labs found to have engaged in coordinated extraction, whether by Commerce Department Entity List designations, restrictions on U.S. cloud infrastructure access or other instruments, should attach to the fraudulent conduct rather than to the open-weight release itself, since weights already published cannot be unpublished. A ban on the underlying technique is neither enforceable nor consistent with how the American AI industry itself operates.[36] Third, a disclosure requirement, described by one analyst as a nutrition label for the AI supply chain, would require AI services operating in the United States to identify the provenance of their base models and where associated user data is processed, addressing the more mundane but arguably more consequential problem that many American developers are currently building on Chinese models without their own customers' knowledge.[37] Fourth and perhaps most important from a purely competitive standpoint, sustained public investment in American open-weight alternatives, through instruments such as the long-pending CREATE AI Act, an open-weights track within the National Artificial Intelligence Research Resource or advance government purchase commitments for inference services built on domestically developed open models, would address the reality that developers are choosing Qwen and Kimi not from ignorance but because those models are free, permissively licensed and cheap to serve, qualities any credible American alternative would need to match rather than merely criticize.[38]

None of these measures resolves the basic tension identified in the previous chapter, that frontier labs need pricing power to fund development and that pricing power is precisely what a maturing open-weight ecosystem erodes regardless of its legality. What they can do is separate legitimate competitive pressure, which American policy has no principled basis for suppressing, from fraudulent extraction, which it does. Whether that separation holds in practice depends heavily on frontier labs' own security posture, a point War on the Rocks pressed with some bluntness: Anthropic's most restricted model was reportedly accessed by outside hobbyists on a public forum before its official release, during which several hundred fraudulent accounts managed to run millions of queries against it, evidence that the porousness enabling large-scale extraction is not exclusively a function of foreign state actors' sophistication but also of American labs' own uneven access controls.[39] A company asking the U.S. government to treat its model outputs as strategic assets carries some obligation to secure them as such before it can credibly ask the state to backstop that security through sanctions.

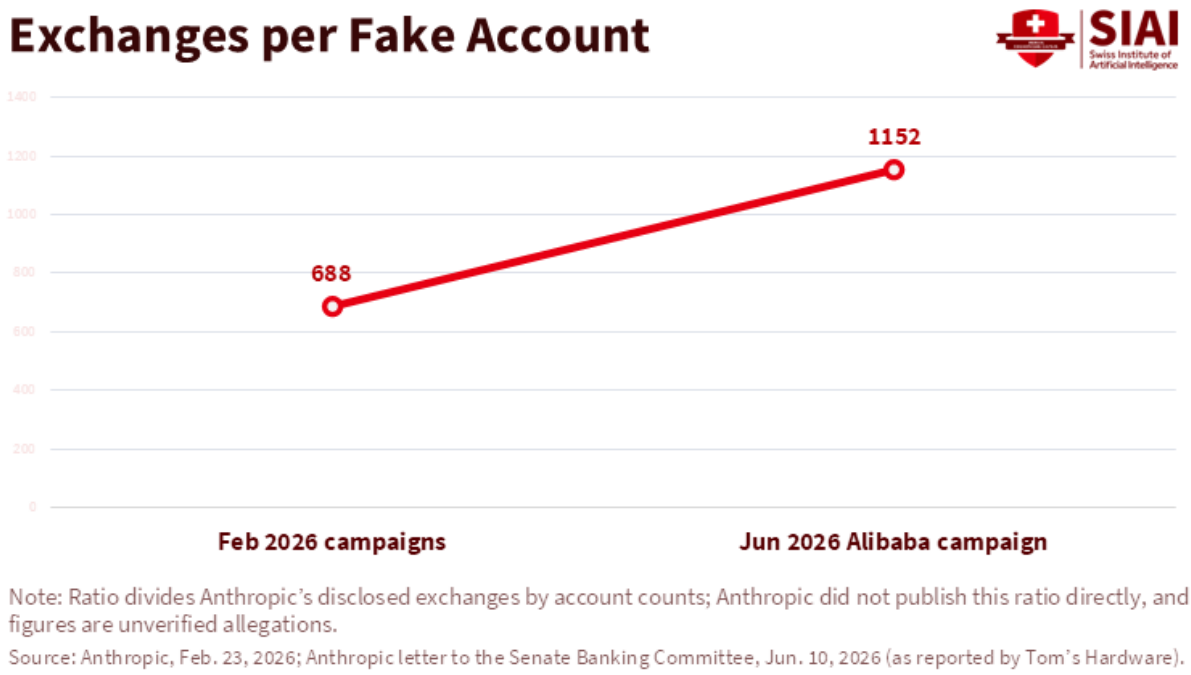

Figure 8: Accounts barely grew between campaigns, exchanges per account nearly doubled, extraction intensity per identity rising faster than the identities themselves.

5 Conclusion - Reciprocity and the Sustainability of the Open-Weight Commons

An older dispute, unrelated to artificial intelligence on its face, offers a useful, if imperfect, analogy: does the relationship between commercial beneficiaries of an open technology and the community that sustains it need to be reciprocal to remain viable over time and if so, what happens when it is not. The clearest domestic precedent is not from artificial intelligence at all but from the WordPress ecosystem, where roughly a third to just over two-fifths of the world's websites run on software whose ongoing maintenance depends heavily on voluntary and corporate contribution rather than any licensing fee.[40] In September 2024, WordPress co-founder Matt Mullenweg publicly accused the hosting company WP Engine, then generating a large project,

of failing to contribute adequately to its upkeep, calling the firm a cancer on the ecosystem and ultimately blocking its access to WordPress.org's plugin and update infrastructure entirely, a move a federal court later ordered reversed on preliminary injunction grounds.[41] Automattic subsequently reduced its own contribution to the mutual maintenance effort to match what it characterized as WP Engine's comparatively modest 50-hour weekly commitment, an unmistakable signal that the party that had historically subsidized the commons was no longer willing to do so unilaterally.[42] The dispute was confused, litigious and, in significant part, personal and it does not map cleanly onto the distillation debate. But the primary structural claim it dramatized, that an open resource sustained disproportionately by a handful of contributors can only remain healthy if commercial beneficiaries return some share of value to its maintenance, translates directly to the AI distillation problem, where the extraction has so far flowed almost entirely in one direction, from labs bearing tens of billions of dollars in training costs toward labs that, according to Anthropic's own disclosures, have paid nothing beyond the cost of maintaining fraudulent accounts.

It would be an overstatement and not one this paper's evidence supports, to claim that continued distillation at current rates spells the literal end of frontier AI development, in the way an unmaintained WordPress core might eventually leave hundreds of millions of sites exposed to unpatched vulnerabilities. Frontier labs retain revenue sources, enterprise contracts, government work and increasingly diversified product lines that are considerably more insulated from consumer-facing price competition than the WordPress analogy implies. What is more plausible and better supported by the trajectory traced across the preceding chapters, is a slower and more consequential shift: a frontier increasingly funded by concentrated enterprise and government revenue rather than broad consumer subscription, models released with narrower windows of open access before distillation risk forces tighter restriction and a research culture that grows more secretive precisely as the security case for openness, made forcefully in the industry's own open letter, becomes harder to sustain in practice. Nvidia, Microsoft and their co-signatories argued in July 2026 that openness itself reinforces security by exposing systems to greater scrutiny and reducing single points of failure, a claim with real merit in domains such as conventional software vulnerability discovery.[43] Amodei's own rebuttal, that this logic breaks down specifically in domains with a strong offense-defense asymmetry, biological weapons design chief among them, where a capable model might shorten the path towards catastrophic misuse far faster than defensive measures can be assembled, remains, on the evidence available as of this writing, unresolved by empirical testing rather than settled by either side's assertion.

The evidence assembled here supports a narrower and more defensible conclusion than either the alarmist framing, which treats every open-weight release as a national security emergency or the libertarian framing, which treats every distillation accusation as protectionist cover. What threatens American AI leadership is not the open-weight model as an artifact but the absence of a functioning boundary between legitimate competitive pressure and fraudulent capability extraction, a boundary that neither existing intellectual property law nor voluntary corporate disclosure has yet supplied. Narrowing that gap does not require banning open weights, a step even their most exposed victim, Anthropic, has explicitly and repeatedly declined to endorse. It requires an independent verification capacity that does not depend on the accused party trusting the accuser's internal forensics, sanctions calibrated to fraud rather than to technical achievement, disclosure rules that let American

businesses and their customers know what they are actually building on and a sustained public and private investment in domestic open alternatives capable of competing on the terms, price, license and ease of deployment, that are currently drawing developers toward Chinese models regardless of any accusation leveled against them. Absent that combination, the likely trajectory is neither China's outright displacement of American frontier labs nor the collapse of AI development altogether. Still, a narrower, more defensive and more consolidated American frontier will be achieved by retreating from the openness that made rapid, distributed innovation possible in the first place.

References

[1] Seetharaman, D. and Arámburo, F. (2026) 'OpenAI says China's DeepSeek trained its AI by distilling US models, memo shows', *Reuters*, via Yahoo Finance, February.

[2, 5, 16, 24, 26, 27] Anthropic (2026) 'Detecting and preventing distillation attacks', *Anthropic News*, 23 February.

[3, 28] Digital Applied (2026) 'Anthropic Accuses Alibaba of Record Model Distillation', 27 June; Novet, J. (2026) 'Anthropic accuses Alibaba of campaign to "brazenly" and "illicitly" extract AI capabilities', *CNBC*, 24 June.

[4] Villasenor, J. (2026) 'Why open-weight models are crucial for American AI leadership', *Brookings*, 10 August; Schuler, M. (2026) 'China Accuses US AI Firms of Distilling Chinese Models', *Implicator.ai*, 27 July.

[6, 25, 29, 35, 37, 38, 39] Fedasiuk, R. (2026) 'How to Stop China from Freeriding on American AI', *War on the Rocks*, 3 August.

[7] Villasenor, J. (2026) 'Why open-weight models are crucial for American AI leadership', *Brookings*, 10 August.

[8, 9, 10] Clark, C. (2026) 'The Open Weight Models that Matter: June 2026', *OpenRouter Blog*, 27 June.

[11] *Open Weights and American AI Leadership* (2026) Open letter hosted by Microsoft Corporate Responsibility, 24 July.

[12] *AI Update #30: The Distillation Economy* (2026) Substack newsletter, 4 May.

[13] Reuters (2026) 'China's open-source dominance threatens US AI lead, US advisory body warns', via Yahoo News, 23 March.

[14] Bhusal, M. (2026) 'Why the AI Subsidy Story Keeps Getting Weaker', personal blog, 30 April.

[15] 'Top American AI execs sound alarm on Chinese models' (2026) via MSN.

[17] NPR (2026) 'Some U.S. startups are turning to cheap Chinese AI models', 15 July.

[18] Bloomberg (2026) 'China's AI advance creates "death zone" for rival US model makers', via *Business Standard*, 4 August.

[19] 24/7 Wall St. (2026) 'Anthropic Says Alibaba Used 25,000 Fake Accounts to Copy Its AI, and the Stock Is Already Sliding', June.

[20] Forbes (2026) 'Anthropic Says Alibaba Used 25,000 Fake Accounts To Distill Claude', 26 June.

[21, 44] Amodei, D. (2026) 'Our position on open-weights models', *Anthropic News*, 27 July.

[22] South China Morning Post (2026) 'Jevons Paradox: why China's cheap AI models could be good for Silicon Valley'.

[23] Fedasiuk, R. (2026) 'How to Stop China from Freeriding on American AI', *War on the Rocks*, 3 August; NPR (2026) 'Allegations of AI distillation spark debate about IP theft. But is it illegal?', 28 July.

[30] Morales, J. (2026) 'Anthropic claims that China's Alibaba used 25,000 fake accounts and 28.8 million exchanges to illicitly "distill" its Claude model', *Tom's Hardware*, 25 June.

[31] Schuler, M. (2026) 'China Accuses US AI Firms of Distilling Chinese Models', *Implicator.ai*, 27 July.

[32] 'After DeepSeek bombshell, ChatGPT's OpenAI accuses Chinese firms of replicating its AI models' (2025) via *Malay Mail*, January.

[33] Lichtenberg, N. (2026) 'Anthropic claims 3 Chinese companies ripped it off, using its AI tools to train their models', *Fortune*, 24 February.

[34] Techmeme (2026) aggregation of reactions to Anthropic's February 2026 distillation disclosure, including commentary from technology critics, February.

[36] Amodei, D. (2026) 'Our position on open-weights models', *Anthropic News*, 27 July; Fedasiuk, R. (2026) 'How to Stop China from Freeriding on American AI', *War on the Rocks*, 3 August.

[40] Gravitykit (2026) 'WordPress powers 33% of the web in 2026 (down from 36% at its peak): CMS market share report'.

[41, 42] Mehta, I. (2025) 'The WordPress vs. WP Engine drama, explained', *TechCrunch*, updated January.

[43] *Open Weights and American AI Leadership* (2026) Open letter hosted by Microsoft Corporate Responsibility, 24 July.