

Memo

An American Answer to China's Open SI Models:

A Safe Harbor for Open Weights

Arpit Gupta

Senior Fellow
Manhattan Institute

Brandon Fuller

Executive Vice President
Manhattan Institute

Sean Speer

Technical Writer
Manhattan Institute

Summary

Chinese labs are now the global default for open-weight Super Intelligence (SI). This means that the calculations governing the SI model (the final “weights”) have been made public and can be customized by users to adjust the model to their own needs. Developers who want a model that they can download, modify, and run themselves increasingly start with Qwen or DeepSeek. Unfortunately for the United States, these models are built by firms subject to Beijing’s laws and content rules, which come with security and free-speech tradeoffs that make many Americans uneasy. Independent testing has found these models are far easier to jailbreak (prompting an SI model into ignoring its safety guardrails or other internal constraints) and hijack than American ones, and several write less secure code depending on the identity of the user.

A chief reason why developers use these Chinese models is because American open releases are voluntary and one-off occurrences, leaving developers unable to plan around future releases. The leading SI labs do not open even their older models to the public, in part because they fear distillation (a technique in which a smaller model is trained to replicate the behavior of the more sophisticated model).

And yet, those same labs are asking Congress for an antitrust safe harbor—a legal provision that protects companies from liability or penalties—to cooperate on security threats, including Chinese theft and distillation of their models.

About Us

The Manhattan Institute is a community of scholars, journalists, activists, and civic leaders committed to advancing economic opportunity, individual liberty, and the rule of law in America and its great cities.



We propose a trade: Labs that claim the safe harbor from antitrust would commit to releasing capable open-weight versions of their models on a predictable schedule, once those models pass independent risk evaluations.

We think of this as a pact with two complementary sides. One side is pro-security and the other is pro-competition. The labs would get legal room to cooperate with each other on safety and against theft and distillation at what is called “the frontier”—the most advanced SI system(s) available at a given time. In return, the public gets a wider and more dynamic SI ecosystem underneath, built on American technology. This is a settlement of political economy. It involves a market intervention, but a minimally impairing one. And this settlement is far better than the alternatives for keeping the U.S. on the frontier, and keeping China from dominating open-weight SI.

Chinese models win today largely on price. Because their weights are public, any company with servers can use them and offer them to their customers, from cloud giants like Amazon and Microsoft to specialist firms like Together AI and Fireworks, and users can run them on their own hardware. Competition among those hosts pushes prices toward the cost of computing. Regular releases of American weights would extend that dynamic, giving firms more reason to choose American models over Chinese ones.

It is not sufficient for the U.S. to lead the world on the SI frontier. Its global advantage must extend across the entire SI stack, including open-weight models, where China is gaining a stronghold based in part on state subsidies and other policy interventions.

The policy approach set out below can boost American participation in the open-weight market and make American technology the foundation the rest of the world builds on. It also expands on Manhattan Institute's broader case that diffusion and competition is the right policy frame for SI. This also advances our overall aim of extending competition across the SI stack while enhancing safety at the frontier.

Meeting these policy-design requirements is challenging: open-weight model releases must be strong enough to beat Chinese alternatives, safe enough to publish, and not present IP threats to existing firms. We argue such a zone exists, even if it is narrow.

China is Winning Open-Weight SI

Open-weight models can be downloaded, adapted, and run on users' own hardware, often making them the foundation for startups, universities, hospitals, and governments with privacy concerns. Whoever supplies these models shapes what the world builds on top.

China's Growing Dominance

China now leads in the use of these models. Hugging Face, the dominant platform where developers share and download SI models, found in its spring 2026 report that Chinese open-weight models accounted for 41% of downloads on the platform (compared to 36.5% for American models) between February 2025 and February 2026. In almost every month of 2026, the largest Chinese open release exceeded the largest American one in parameter count. They are also the foundation of more spin-offs: Qwen has 151,448



derivative models on Hugging Face, 2.6 times Meta's total footprint. Chinese labs also license more generously: among releases above 20 billion parameters, 81% of Chinese models carry Apache 2.0 or MIT licenses, against 29% of American ones.

This model proliferation translates into actual use. On OpenRouter, which routes developers' requests to many models, Chinese open-weight models rose from 1.2% of tokens in late 2024 to nearly 30% in peaks of 2025, according to a study by OpenRouter and a16z. By April 2026, Chinese labs handled more than 45% of weekly token volume. American startups increasingly build on Chinese models because they are cheap and easy to adapt, and some of the largest American open releases this year are themselves built on Chinese foundations.

The Hugging Face episode this summer reflected the growing U.S.-China open-weight gap. When the company suffered a sustained intrusion by an autonomous SI agent in July 2026 (later shown to be agents from OpenAI), safeguards on frontier models blocked their team's requests for help analyzing the attack. Hugging Face finished the job with GLM-5.2, a Chinese open-weight model it could run on its own servers.

The closed frontier's lead is also shrinking. The UK AI Security Institute finds that the best open Chinese models trail the closed frontier (in which a model's underlying code and model weights are still secret) on cyber tasks, such as finding and exploiting software flaws, reverse-engineering code, or breaking into web applications, by only 4 to 7 months, down from 6 to 10 in 2025.

Why It Matters

The first risk that the spread of Chinese open-weight models poses to the U.S. is security. In its 2025 evaluation of DeepSeek, the National Institute of Standards & Technology's Center for AI Standards and Innovation (CAISI) found that the Chinese lab's most secure model complied with 94% of overtly malicious requests using common jailbreaks, compared to just 8% for American reference models. In addition, the agents built on DeepSeek were an average of 12 times more likely to be hijacked than agents built on the American models GPT-5 or Opus 4.

An open-weighted model is not inherently a weaker, less secure model. In the CAISI study, OpenAI's open-weight gpt-oss model matched or exceeded every DeepSeek model's resistance to hijacking. Developing a model that is easy for malicious users to jailbreak, then, is perhaps a deliberate decision of model design and public policy on the part of China.

The second risk is hidden in the code that these models write. In 2025, CrowdStrike found that mentioning politically sensitive topics such as Tibet or the Uyghurs made the Chinese model DeepSeek-R1 up to 50% more likely to write code with severe security flaws, such as passwords embedded in the code or careless handling of user input that an attacker could exploit. This occurred even when the topic had nothing to do with the coding task assigned to the model. In June 2026, Booz Allen reported that three-out-of-four Chinese coding models it tested wrote more vulnerable code when told the user worked for the U.S. government. One model, Qwen3-Coder, produced code with a vulnerability score 130% higher than before it knew the identity of the U.S. user.

The third risk is one of propaganda. Beijing's official views travel with the model. CAISI found that DeepSeek models echoed four times as many inaccurate and misleading Chinese Communist Party narratives as American models, even when the model weights were downloaded from Hugging Face and in the English language. Products built on these models for an English-speaking and U.S.-based audience necessarily inherit those biases.

The fourth risk is covert manipulation. Chinese labs operate under a PRC law obliging cooperation with state intelligence work. Furthermore, users cannot verify how a Chinese model was trained. SI developers rely on enormous amounts of text and documents to train their models. If even a small subset of those training documents are written to sabotage the model, it can have serious consequences. Research from Anthropic and the UK AI Security Institute shows that as few as 250 poisoned documents can plant a hidden "backdoor" vulnerability in a language model. Models can also be seeded with sleeper instructions that only activate under certain conditions.

Verifying that open-weight models do not contain these kinds of malicious instructions is challenging if not impossible. There is no public evidence of a state-planted backdoor in a Chinese model, but as these models enter critical systems, including American water and electrical infrastructure, among many others, the burden of proof should not rest on the user.

Finally, China's lead in open-weight SI draws heavily on American work. Chinese labs have used fraudulent accounts to distill American frontier models, then released the results as open weights. White House science adviser Michael Kratsios has said China's Moonshot AI distilled Anthropic's Fable model to build Kimi K3, and in September 2026 the National Security Agency and other agencies warned of "industrial-scale" distillation by Chinese firms. American labs are still superior: The Institute for Progress estimates that adversarial distillation erases about half of America's lead over China in SI capabilities, roughly the same boost as giving Chinese labs four times as many chips as they have today. And greater use of these open-weight models incentivizes further IP theft from American companies.

American labs have released their own open-weight models: Google's Gemma, Meta's models, NVIDIA's Nemotron, and OpenAI's gpt-oss are capable and widely used. In addition, in July 2026, NVIDIA, Microsoft, and others launched an Open Secure AI Alliance. Thinking Machines unveiled a 975-billion-parameter open model in July, Meta released Muse Glimmer in August and said it would open the weights of Muse Spark 1.2, and NVIDIA keeps expanding its Nemotron family. That is welcome, and it shows demand for the models. But these releases are one-off choices, and the leading closed labs do not open their flagship lines, even after their primary model is superseded by another. As a result, there is presently no dependable American open-weight pipeline. Developers cannot plan investments or build their company's software and services around an SI model family that may not have a successor.

To be clear, China is not competing on a level field. Among its subsidies and policy distortions, local governments have issued compute vouchers worth roughly \$140,000 to \$280,000 to their SI startups, on top of the structural economics that make open models cheap to serve. Washington is now weighing its own policy response. The 2025 AI Action Plan calls for encouraging American open weights, and Sen. Jim Banks has asked the administration to develop incentives that limit dependence on Chinese open-weight models.

Direct subsidies are a potential policy tool, but they are slow. Funding new open models through grants or an agency like DARPA means training models from scratch, which takes time. A safe harbor works with models that already exist. Labs have already trained them, and releasing these models costs far less than building new ones, so they can reach the market faster. Another advantage to our proposal of trading safe harbor for open weights is that it will boost American participation in the open-weight market without new appropriations. It trades regulatory relief that the labs want for releases that the country needs.

A Safe Harbor the Labs Want

No single American lab can stop Chinese distillation alone. The evidence (account indicators, network origins, behavioral patterns, hashed prompts) is scattered across firms, and pooling it means rivals must share information, which antitrust law makes legally uncertain. Chris Meserole, executive director of the labs' Frontier Model Forum, has testified that its members have "had to take a fairly conservative approach under antitrust law to even have a conversation about how to identify distillation," and have not discussed countermeasures.

There has been progress on addressing distillation in recent months. Labs are now detecting and banning distillation accounts and restricting access to their models' reasoning traces, the step-by-step record of the model's calculations and logic behind its determinations. In April, the White House directed agencies to share intelligence with SI companies and develop defensive best practices, and federal agencies issued a joint warning in September. But the problem is by no means solved. Distillers still operate networks of hundreds of thousands of fraudulent accounts and shift to whichever lab's defenses are weakest in the moment.

However, that distillation is not the only reason for a safe harbor. Labs also need legal room to share information on the other risks, and to coordinate pre-release access and limited rollouts of especially dangerous models.

Existing law permits some safety collaboration, but uncertainty remains and the available tools do not fill the gap. Agency guidance, for instance, does not protect against private lawsuits or state antitrust claims. Department of Justice business reviews and Federal Trade Commission advisory opinions take months. And the antitrust exemption in the Cybersecurity Information Sharing Act of 2015 covers only traditional cyber threats, and is set to expire on December 11, 2026, unless Congress extends it.

The Schiff-Banks Bill

A new bill, The Collaboration on Adversarial Threats and Security Risks Act (S. 5105, H.R. 9914), introduced in July 2026 by Sens. Adam Schiff and Jim Banks and Reps. Bob Latta and George Whitesides, would fill that legislative gap. Modeled on the 2015 Cybersecurity Information Sharing Act, it would make two kinds of conduct lawful under antitrust law.



The first type is good-faith information-sharing that exclusively addresses a “covered artificial intelligence security risk.” The bill specifies six such risks: (1) theft, weaponization, or distillation of SI by China, Russia, North Korea or Iran; (2) SI that substantially helps build chemical, biological, radiological, nuclear, or offensive cyber weapons; (3) serious threats to critical infrastructure; (4) SI that weakens human oversight or shutdown mechanisms (a so-called “kill switch”); (5) autonomous self-improvement by models creating those risks; and (6) unauthorized-access vulnerabilities creating them.

The second type of protected conduct is more far-reaching. Under this bill, companies could coordinate or reach agreements to reduce those six types of risks by “delaying or otherwise limiting the release, deployment, use, development, training, testing, or evaluation” of SI, provided they first notify DOJ’s Antitrust Division of the specific risk and the scope of the restriction.

What this means is that, in practice, labs could jointly shut down the fraudulent-account networks behind Chinese distillation. They could also coordinate limited rollouts of especially dangerous models, as Anthropic and OpenAI did unilaterally this year. That coordination matters because if one lab holds back, Chinese labs can distill another. The bill would also let labs jointly commit to giving the government longer pre-release access than the 30-day voluntary review the administration has called for.

These protections are characterized as an affirmative defense: companies must show good faith and exclusive purpose and keep controls on shared information. Crucially for promoting competition, the bill bars using the information-sharing exemption for price-fixing, market allocation, monopolization or boycotts, and it preserves private suits against the labs over conduct it does not exempt. Coordinated delays are policed by the exclusive-purpose test (a strict legal standard that requires an activity must serve only one specific, designated purpose), advance notice to DOJ, and DOJ’s ability to seek an injunction, even of good-faith conduct likely to increase the risks the bill targets.

Why Critics Worry

Critics’ concerns about the potential for corporate collusion are reasonable. The Congressional Research Service has noted that an agreement among leading developers to hold back development would carry the highest antitrust risk of any form of safety collaboration.

The coordinated-delay provision is a key source of these concerns. An agreement among a few dominant firms to limit releases can look like output restriction by another name. Democrat and Republican officials have raised concerns that it could lead to SI concentration. Lina Khan, former FTC chair, has also warned against the “highly concentrated and interconnected structure” of SI markets.

Both sides have a point. On one hand, the labs need legal room to defend against Chinese theft. On the other hand, the public needs assurance that this will not be deployed for anti-competitive ends. Coordination among frontier labs could mean higher prices, slower improvement, and standards that unnecessarily raise costs of smaller rivals. Pairing the safe harbor with a predictable series of model releases helps to address this concern by ensuring the release of new open-weight models to advance competition in the broader ecosystem.

Safe Harbor for Open Weights

Under our proposal, the Schiff-Banks bill's coordinated-delay protection would be contingent on their releasing open-weight models which trail the closed frontier. The release would be timed to match capabilities that are openly available. Each new frontier model would trigger a release obligation, with timing set by the eligibility and safety criteria below.

Once those are met, the lab would publish a version as open weights under a permissive commercial license. A lab that fails to meet its obligation would lose the safe harbor, with exceptions for legitimate safety deferrals.

This proposal involves a voluntary trade: a lab that does not claim the safe harbor owes nothing. It would present a relatively light-touch intervention in the sector without creating a new regulator or government ownership stakes.

Labs would release open weights—not open source, which is more comprehensive—so training data, code and methods stay private, and the frontier itself remains closed. Schiff-Banks already contemplates coordinated release delays. This proposal pairs those delays with a presumptive release schedule.

The releases first must be subject to an independent safety test by a government evaluator, most naturally CAISI. If a model fails, its release will be deferred and the model will be retested.

Releases must be worth having. They must be full weights of a commercially capable model, under standard permissive licenses. Stripped-down versions or restrictive terms would not count, and missed or sham releases would end safe harbor protection. To ensure the open-weight release happens, labs could be required to deposit weights with an evaluator when they launch a new frontier model, to be released on schedule unless a safety deferral is warranted.

Hosting and Integrity

Released weights can be altered, including safeguards stripped through fine-tuning, hidden instructions or backdoors inserted, with the result redistributed under the original name. Nothing can stop a user from altering their own copy, which is why pre-evaluation includes attempts to strip safeguards. Existing and voluntary tools can, to a large extent, address whether a copy can be trusted: signed weights and official hashes let users verify they are running the release as published; licenses can bar altered versions from using the model's name; and optional certification, as FedRAMP does for cloud services, can identify hosts that meet published criteria.

A more controlled form of release, in which models are only hosted by authorized or "trusted" deployers, is the main alternative. The case for this method is that labs may be more willing to participate if identifiable firms handle deployment and take on legal responsibility for how the model is offered. But liability does not require gatekeeping. Hosts already answer under ordinary contract and consumer-protection law for what they offer, and the safe harbor could be paired with limited liability protection for labs whose models pass federal evaluation. Who chooses the deployers also matters. If the

labs choose, incumbents gain a say over who can compete with them downstream. A government list or published certification criteria would avoid that, at the cost of speed and new administrative machinery.

We therefore treat controlled release as a fallback for only the highest-risk models. It is not at the core of this proposal. In its strictest form, trusted deployers would run the model for customers, who could send it requests but never receive the files themselves. If unrestricted release of a model cannot be made safe or attractive enough, that is an important finding, and it would raise the question of whether a narrower concession deserves the same antitrust protection.

Who Does What

The Schiff-Banks bill is broader than this pact. Its protections reach any two or more non-federal entities cooperating on a covered risk, not only frontier labs, and DOJ's Antitrust Division is the only agency that it names. The release obligation therefore needs a defined class of participants. The proposal conditions the coordinated-delay provision, the part that worries critics, on joining the program, while leaving the information-sharing on security threats open to all. Eligibility could rest on a compute or capability threshold, with labs above it registering as participants.

A prospective division of labor follows. DOJ would continue to receive notice, police the exclusive-purpose test and seek injunctions. CAISI, which reportedly administers the administration's 30-day voluntary pre-release review of closed frontier models, would evaluate models before release. Reports say the voluntary framework exempts open-weight models, so the proposal would extend review to all versions that a lab plans to publish. Someone must also certify that a participating lab has met its obligation and determine when protection is lost. We suggest that DOJ take on this role, relying on CAISI's findings, though Congress could assign the responsibilities differently.

Contesting the Chinese Default

Hugging Face credits Qwen's dominance among open-weight SI models to its consistent releases, coverage across model sizes, and permissive licensing. Its dominance does not depend on Chinese models being the best.

Above all, the trade laid out in this proposal attacks China's price advantage, which is more accurately thought of as a gap between closed and open models, not between American and Chinese ones. The UK AI Security Institute compared costs on cyber tasks that both kinds of model solved reliably. It found that, at advertised prices, a closed American model cost \$12.50 per task while a comparable Chinese open model cost only \$0.28. In the economics of SI, closed models must be purchased at a price that covers their developers' profit margin, as well as the cost of research, training, and other needs. They can be purchased only from the developer or licensed clouds. By contrast, open models can be copied by anyone, so their price falls toward the cost of the necessary computing power.

Releasing American weights to the public moves American models onto the same economics as their Chinese rivals. Many different providers could host the American open-weighted models and compete for users on price. Firms could compress them for cheaper hardware

or distill them for specialized tasks. With a narrower price gap, American models would likely win out because of the added security and trust advantages, giving users a stronger reason to switch from dominant Chinese models.

In addition, a dependable pipeline of future open-weight models from American labs may matter as much as price. Startups, hosts, and enterprises will only invest in a model family if they expect future versions of the model to be released. A model that might not have a successor is merely a one-time product, not a reliable tool to grow your business. But a model family with a release schedule, is a platform that can attract complementary investment, from fine-tuning shops to enterprise software, that pulls the digital ecosystem toward American technology. The 2025 AI Action Plan already called for encouraging American open-weight models. This gives that goal a policy mechanism.

Turning a Safe Harbor into Stronger Competition

A scheduled release also answers the antitrust worry. As the safe harbor protections permit cooperation at the frontier, the open-weight releases expand competition more broadly. With the released model weights, other firms can specialize, compress, and host capable models without having to spend the hundreds of millions of dollars or more necessary to recreate them; hospitals and factories can run them on their own servers; and open alternatives force labs to set a reasonable price for routine work. In effect, the safe harbor lets a few firms cooperate at the top, where safety concerns are paramount, while the release obligation opens the layers beneath them to more people and more players.

This is a key part of our proposed pact. We cannot assess how competitive this industry is by only looking at one stage and one layer. Concentration in one layer can coexist with vigorous competition in the layers underneath. As we have said, safe harbor gives labs necessary relief from antitrust risk where cooperation addresses serious national security concerns at the frontier. At the same time, the release obligation conditions that relief on opening competition elsewhere in the technology stack.

This is the basis for Manhattan Institute's broader thinking on SI diffusion. We have argued elsewhere that the right frame for SI policy is diffusion: the goal is to put sufficiently safe SI capabilities in as many hands as possible. A safe harbor for open weights applies that frame to market structure. It recognizes that the frontier is not the only entry point for new firms, products, and ideas. Open weights open the layers beneath it to industry entrants who never need to train a frontier model.

A Credible Precedent

This type of policy trade has been made before. Funded by AT&T's monopoly profits, Bell Labs produced the transistor, information theory, and the solar cell. In 1956, DOJ settled its antitrust suit against AT&T by permitting the company to keep its telephone monopoly in exchange for licensing its existing patents royalty-free and staying out of businesses beyond regulated telecommunications. AT&T accepted the trade.

In 2020, scholars found that the settlement raised follow-on innovation building on Bell patents by 17% over five years, mostly from young and small firms, and only outside telecommunications, where Bell's monopoly stayed intact. Although the analogy has limits, it shows that opening an incumbent firm's technology can produce innovation that the original firm likely would never have pursued.



AT&T is not the only precedent. In 1917, with patent disputes hobbling aircraft production for American pilots in World War I, the U.S. government pressured rival manufacturers into forming a cross-licensing organization. In 1980, Bayh-Dole let universities and firms carrying out federally funded research own the inventions that result from that funding, while obliging them to pursue commercial use. In each case, policymakers accepted a limited restriction in one area of the economy in order to get broader competition and diffusion within the larger industry. A safe harbor for open weights is a settlement of the same kind.

The Narrow Zone

A credible American alternative to Chinese open models must pass three tests. Each test has its own policy design considerations.

Test	The Question	Implication for Release
Competition	Is the release close enough to Qwen and DeepSeek, and trusted enough, to displace them as a default?	Release models sooner and release more capable versions
Safety	Can the release be published safely, knowing the weights can never be recalled?	Release models later, after rigorous evaluation
Edge	Does the release hand Chinese labs capability they lack?	Only release models that roughly match Chinese capabilities

The zone between these three tests is narrow, but a zone does exist. The 4-to-7-month gap between the best Chinese open models and the closed American frontier suggests there is an opportunity: an American model released within that window could compete with leading Chinese models without disadvantaging the United States. If the release schedule is too late, the American models released in accordance with this proposal will not be able to compete with what Chinese offer. Yet if they are released too early, they will hand Beijing new capabilities.

With time, economics and trust should widen the zone. Once they are released, American models compete on the same cost basis as Chinese models, so they only need to come close on technological capability. A model from a known American lab, evaluated by CAISI, offers assurance that Chinese models cannot.

Several policy design choices follow from this narrow zone. The policy should index releases to the open frontier rather than the calendar: a model becomes eligible for release once comparable capability is already openly available. A fixed lag is a crude metric. But a range, with a minimum release date that protects development on the frontier and a maximum release date that keeps abreast with open Chinese models' capabilities, keeps American releases competitive. Roughly 6-to-18 months is a plausible starting range. Safety thresholds should take precedence over timing targets.

Making it Work for Labs and Adopters

This policy must therefore be carefully calibrated to balance its twin goals of ensuring security and promoting competition.

For labs, they will choose to opt in only if the safe harbor is worth more to them than the release obligation costs. Rules should be predictable and the cost contained—i.e. they should be obligated to release weights only and they should be given a release schedule that protects frontier returns. Labs can still profit from hosting their models or selling enterprise customization to users. Other advantages to labs could include: protecting collaboration only among labs that opt in, offering limited liability protection for models that pass federal evaluation, or even giving participants federal procurement preference.

New enforcement tools may also be part of an overall policy package. Congress is already moving against model extraction. In April, the House Foreign Affairs Committee unanimously advanced the Deterring American AI Model Theft Act, and the Senate's BLADE Act would target fake or unauthorized accounts that try to extract details of a closed model through fraudulent credentials. Stronger civil and criminal penalties for account fraud and systematic extraction would give any evidence labs pool together, because they have safe harbor, real consequences. The target of enforcement tools should be extraction from closed models, not distillation of released weights, which the permissive licenses would allow.

For adopters and users of open-weight models, who already make choices based on model cost, capability, and risk, this proposal provides predictable releases. Release should be close to the open frontier and permissively licensed, so firms can build around a model family that will keep improving and will not be restricted later. Because these models have been evaluated by CAISI, American labs can have a degree of trust and dependability that would be a competitive advantage relative to the Chinese models.

These various considerations are essential for this policy model to ultimately work. The bargain will be judged by whether labs join and adopters move.

Addressing the Concerns

The most obvious objection is that releasing American models unintentionally hands capability to Beijing. But under this proposal, the models that would be up for release would roughly match the existing capability of the Chinese labs—and the frontier itself would remain closed. Distillation only helps China where the source model can do something the student models cannot, so a released model at or below the Chinese frontier adds little to their capabilities.

Antitrust skeptics may argue that even a prescribed release schedule does not address the harms of safe harbor. But keep in mind that (1) the releases themselves should widen competition in the overall SI economy, (2) antitrust protection ends if labs fail to deliver, and (3) DOJ oversight remains.

Safety advocates will rightly note that weights cannot be recalled, which is why independent evaluation comes first. Free-market critics may see a mandate in disguise, but labs can opt out, and the proposal is far lighter than alternatives under discussion, from lab breakups to government equity stakes.

Some critics may favor open-weight mandates with no safe harbor at all—which is to say, they may only want half of the proposed trade. That would impose the cost with nothing in exchange and weaken labs' incentive to invest at the frontier. It would also wrongly set two priorities against each other, open-weight diffusion and frontier investment, when beating China ultimately requires both. A mandate invites minimal compliance and legal challenge. The pact is designed to align the two: labs that join do so because the

safe harbor is worth more than the release obligation costs, which gives them reason to release capable models and keep releasing them. The goal is to get American open weights out as widely as possible, allowing our own trusted models to serve as the foundation of the global SI-driven digital ecosystem, and a deal labs choose to join does that more durably than a mandate they resist.

The harder questions are about practical policy design. Labs may decline if regular releases seem too high a price for a narrow safe harbor, so scope and obligation must be negotiated together. And any bargain invites gaming, as drug makers did with the 1984 Hatch-Waxman patent protection/generic drug framework. A capability floor, standard licenses, and independent certification are the safeguards against stripped-down releases, restrictive licenses or slow-walked schedules.

Fundamentally, we think of the proposal as a political economy settlement. It is an intervention, and we do not pretend otherwise. But it is minimally impairing and pro-competition in intent, and it is far better than the alternatives on offer. One of its main virtues is speed: it aims to bring more American open-weight models to market as quickly as possible. In the current public policy and technological environment, it may be the most efficient intervention in favor of maintaining America's lead at the frontier and strengthening its position in the open-weights market.

Next Steps

The central question facing this proposal is how much the protection of safe harbor is worth to the labs and what release obligation they would accept in return. But there are various other policy design questions that would require attention:

- Can CAISI test models quickly and credibly enough, and what resources does it need to carry out this work?
- How long should labs have before a model must be released, and should that period be fixed or tied to the open frontier? And should every frontier model trigger a release?
- Should a controlled tier, including trusted deployers, exist for the highest-risk models, and if so, who should designate them?
- How should the participating labs be defined, and who ultimately certifies that they have met their obligation?

The success of this policy framework should be judged on several measures, including lab participation, the number and quality of releases, adoption of American open models, their cost competitiveness, and the share of Chinese models in use outside China.

Conclusion

It is not enough that the U.S. leads at the SI frontier. It is in the country's strategic interests that it also leads in the open-weight layer much of the world builds on. Trading safe harbor for open weights represents an innovative yet historically precedented policy solution to achieve this goal.

Such a pro-security and pro-competition pact enables the labs to better cooperate on security, gives developers a dependable pipeline of American open-weight alternatives to Qwen and DeepSeek, and boosts overall competitiveness in the SI economy.

The policy design is hard, but the zone in which it works is real, and leaving the field to Beijing is costly.