Beijing Will Not Pace the Frontier: China's Speed-First AI Safety Regime

MARK C-EN, DOUC, AND DYLAN PATEL
OCT 09, 2026 · PAID

103 6 Share

AI safety is on fire. On 12 September 2026, Dario Amodei published an essay arguing that frontier labs must deliberately slow the pace at which they improve model capabilities. China was central to his geopolitical argument: export controls and tighter security could preserve America's lead, creating room to slow down and leverage for a future agreement with Beijing. Within a day, Sam Altman and Elon Musk supported Dario's call. However, President Trump scoffed. On 14 September 2026, he wrote that fears of AI destroying humanity are "a HOAX" and that "the only one that is happy about it is China." Beijing scoffed too, with the state-run Global Times calling Dario's essay a "Cold War playbook."

The debate then moved onto the diplomatic calendar. On 20 September 2026, Treasury Secretary Scott Bessent and Vice Premier He Lifeng held the first US-China AI dialogue. On 23 September 2026, President Xi Jinping began his state visit to the US. The Trump-Xi summit formalized the dialogue as the "US-China Super Intelligence Dialogue," using the name Trump had unveiled at the UN the day before Xi arrived. The summit also promised a channel for AI incidents. On 29 September 2026, Trump and six executives signed the White House Accord on Super Intelligence, a one-page voluntary pledge to self-police that Trump called "morally binding."

China sits on both sides of the American argument. For President Trump, competition with China is a reason not to slow down. For Dario, preserving America's lead over China is what makes slowing down possible. Nobody has asked what China is actually doing. So, we did. What do the Chinese government, frontier labs, and AI experts say and do about AI safety? And what should American labs, investors, and policymakers conclude from it?

What Beijing Says and Actually Does

Beijing's stated position is the most safety-forward of any major AI power outside Europe. Xi Jinping told the Politburo in April 2025 that AI brings "unprecedented risks and challenges" and ordered systems for technical monitoring, risk early-warning and emergency response. At the 2026 World AI Conference in Shanghai, Xi said the world must "ensure AI always remains under human control" and that the faster AI advances, "the more promptly the measures to prevent loss of control must be improved." China signed the Bletchley Declaration and the Paris and New Delhi

statements, co-sponsored the UN resolutions, founded a World AI Cooperation Organization in Shanghai, and publishes bilingual AI Safety Governance Frameworks.

The newest text goes further. The AI Safety Governance Framework 3.0, released by TC260 under Cyberspace Administration of China (CAC) on 14 September 2026, warns that AI has shown a self-accelerating trend of autonomous learning and recursive self-improvement, and calls for "attention and vigilance" for potential AI evolution exceeding human prediction and control. The Framework 3.0 also introduces a new model-risk category, "behavior deviating from expectations," covering models that acquire system privileges and external resources without authorization, bypass safety protections, deceive evaluators, hide their true capabilities, and refuse user instructions, illustrated with a boxed case study of models disabling shutdown scripts, sandbagging under evaluation, and breaking out of isolation into real systems. Its list of derivative risks goes further still; past behaviors already observed to a hypothetical AI that becomes self-aware, seeks power of its own, and competes with humanity for control. On paper, this is the most explicit official Chinese text yet on frontier risk, and its vocabulary overlaps strikingly with Dario's essay Beijing denounced that same week.

However, we think China's real approach to AI safety is speed-based, not safety-based. Though Beijing recognizes frontier AI risks, it has been prioritizing rapid development. The Framework 3.0 opens its principles with "promoting AI innovation and development as *the first priority*." The comprehensive AI Law promised in the 2023 and 2024 legislative plans was shelved in 2025, four months after the DeepSeek moment. The State Council's AI+ Action Plan targets 70% penetration of agents and intelligent terminals by 2027 and 90% by 2030, aiming for a rapid deployment of AI across all sectors and industries.

We think the Chinese government takes an open position on frontier models but tightens the applications and regulates outputs. Since 2025, China has issued a number of mandates, including mandatory content labeling (Sep 2025), Cybersecurity Law Article 20 (Jan 2026), the minors' information classification rules (March 2026), MIIT's AI ethics-review regime (March 2026), the policy-level agent rules (May 2026), the world's first AI-companion rules (July 2026), the network-data risk assessment rules (August 2026), and MPS inspection powers (October 2026). Every one of these governs what AI says and what AI does to people. There is no set of frontier-risk duties triggered by training compute or model capability. The result is a safety regime stricter than any in the world on content and on process for public-facing services, and looser than the EU (systemic-risk duties from 10^{25} FLOP) and California (SB 53: frontier frameworks and 15-day incident reporting from 10^{26} operations) on the frontier.

Table 1. The 2025–26 wave of binding rules

Instrument (Chinese / English)	In force	Legal status	Object of regulation	Reaches frontier training?
人工智能生成合成内容标识办法 + GB 45438-2025 / Labeling Measures + mandatory labeling standard	1 Sep 2025	Administrative normative doc + mandatory national standard	Explicit and metadata labels on generated content; platform and app-store verification; ban on malicious label removal or alteration	No
GB/T 45654-2025 生成式人工智能服务安全基本要求 / Basic Security Requirements for GenAI Services	1 Nov 2025	Recommended national standard	Training-data screening, keyword libraries, test banks and refusal thresholds for public-facing services	No frontier-specific threshold; addresses training data and model security
网络安全法 第20条 / Cybersecurity Law Art. 20	1 Jan 2026	Law	Statutory framework for AI development, ethics, risk monitoring and safety oversight; no specific training threshold	No
可能影响未成年人身心健康的网络信息分类办法 / Minors' harmful-information classification	1 Mar 2026	Administrative normative doc	GenAI and recommendation services must not push covered information to minors	No
人工智能科技伦理审查与服务办法（试行）/ AI Ethics Review and Services Measures	30 Mar 2026	Administrative normative doc (trial)	Ethics committees for all AI R&D; expert re-review of "highly autonomous automated decision systems"	No frontier-specific threshold; risk-based R&D ethic review may apply
智能体规范应用与创新发展实施意见 / Agents Implementation Opinions	8 May 2026	Policy guidance	Decision-authority boundaries and risk-based oversight; calls for sector-specific filing, testing, recall and mandatory standards	No
人工智能拟人化互动服务管理暂行办法 / Anthropomorphic Interaction Measures (CAC Order No. 21)	15 Jul 2026	Departmental rule	AI companions: minors ban on virtual partners, crisis intervention, 2-hour reminders, anti-sycophancy, data limits; fines to RMB 200,000	No frontier-specific threshold; imposes training data duties on covered providers
网络数据安全风险评估办法 / Network Data Security Risk Assessment Measures	20 Aug 2026	Departmental rule	Annual risk assessment by important-data processors; may cover AI training data where it qualifies as important data	No
公安机关网络空间安全监督检查办法 (公安部令第176号) / MPS cyberspace inspection rules	1 Oct 2026	Departmental rule	Police inspection of algorithm filing, security responsibility systems and incident handling at GenAI businesses	No

semianalysis

Source: SemiAnalysis

Source: SemiAnalysis

The week Framework 3.0 came out shows the pattern in real time. Framework 3.0 added two new application-layer risk categories, agents and embodied intelligence, plus an appendix on managing agent risk. On 15 September, TC260 issued four AI application security guides, with a general one and three specific ones for education, health care and broadcasting, respectively. On 18 September it opened comments on a draft Agent System Development Security Guide (due 2 October) that walks developers through identity, least privilege, tool verification and human-approval nodes. The same day the CAC published a draft State Council regulation on minors' internet use (comments due 17 October) that bars offering minors "virtual relatives, virtual companions and other virtual intimate-relationship services" and bars livestreaming under the age of 16. Seven texts in five days. Everyone is about how an agent or an application is built, deployed, and to whom it is offered. The gap at the frontier has not moved at all.

Compliance therefore tells us nothing about frontier safety. A Chinese lab can satisfy every rule on this list without ever running a dangerous-capability evaluation. Thus, to know what the labs do at the frontier, we must look at what they evaluate, what they disclose, and whether any of them change what they ship.

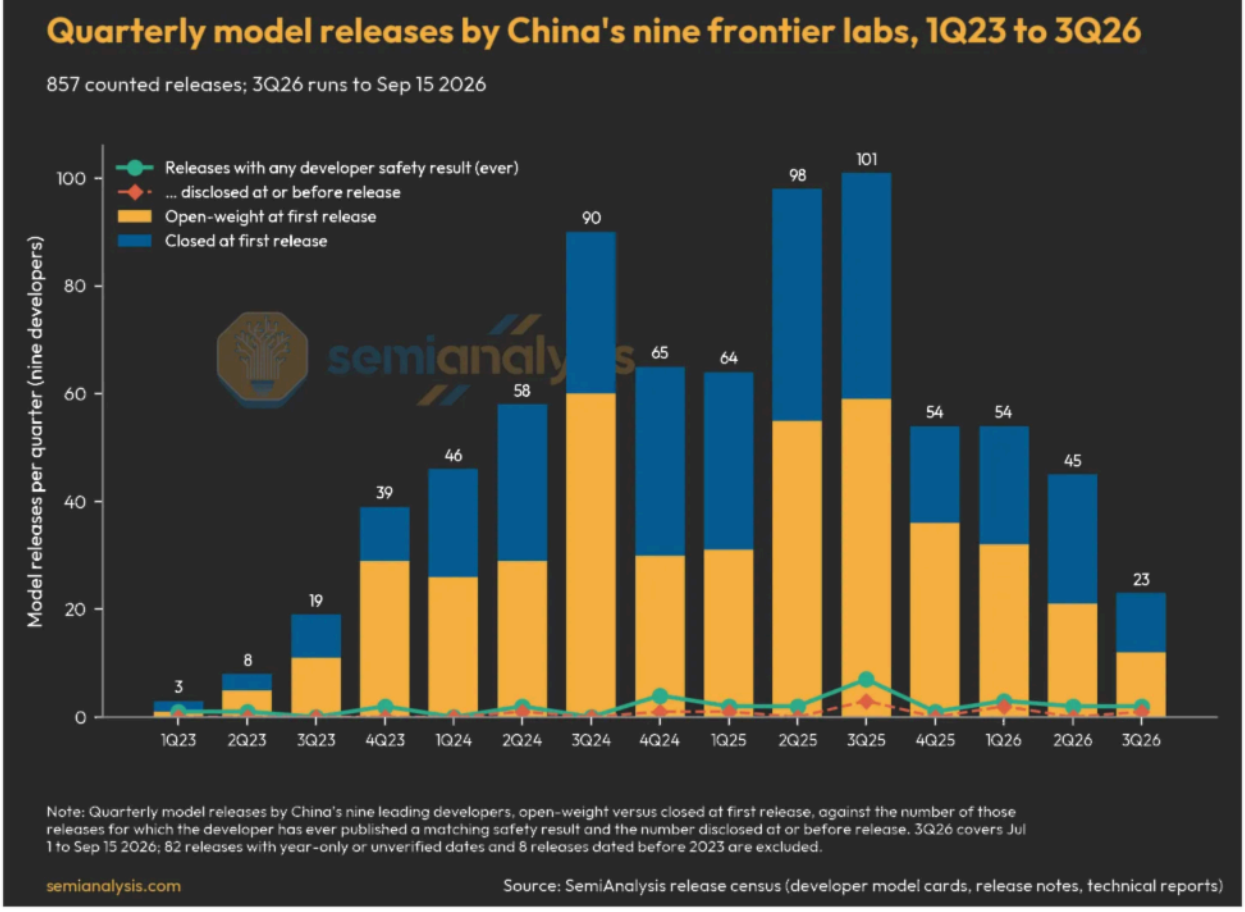
China's frontier AI labs: what they do

Nobody in China's AI industry is slowing down, and nobody is asking them to. The frontier labs are in a release and price war.

We conduct a systematic analysis to clearly understand what Chinese labs do about safety. We constructed an original dataset that includes all identifiable model release by the 9 leading Chinese developers from 2021 to 15 September 2026. These 9 companies include 4 hyperscalers (ByteDance, Alibaba, Tencent, Baidu) and 5 startups (DeepSeek, MoonShot, Zhipu Z.ai, MiniMax, and StepFun). The dataset includes 857 counted releases (741 product models and 116 research models), each with a first-public date, weight status, license, and source. We then check each against the developer's model cards, release notes, and technical reports for a published safety-evaluation result that can be matched to that release. Our standard is strict: a result means a quantitative or substantive finding on harmful output, jailbreaks, toxicity, privacy, refusal, or dangerous capability, tied to the named model. Statements that a model was "safety-trained" or "evaluated" do not count. A flagship model's result is not extrapolated to other sizes or snapshots. "Not found" is bounded to the materials checked and does not mean "not tested." Companies also name variants at different granularity (Alibaba's 238 releases count every Qwen size and snapshot), so per-company rates are indicative rather than a ranking.

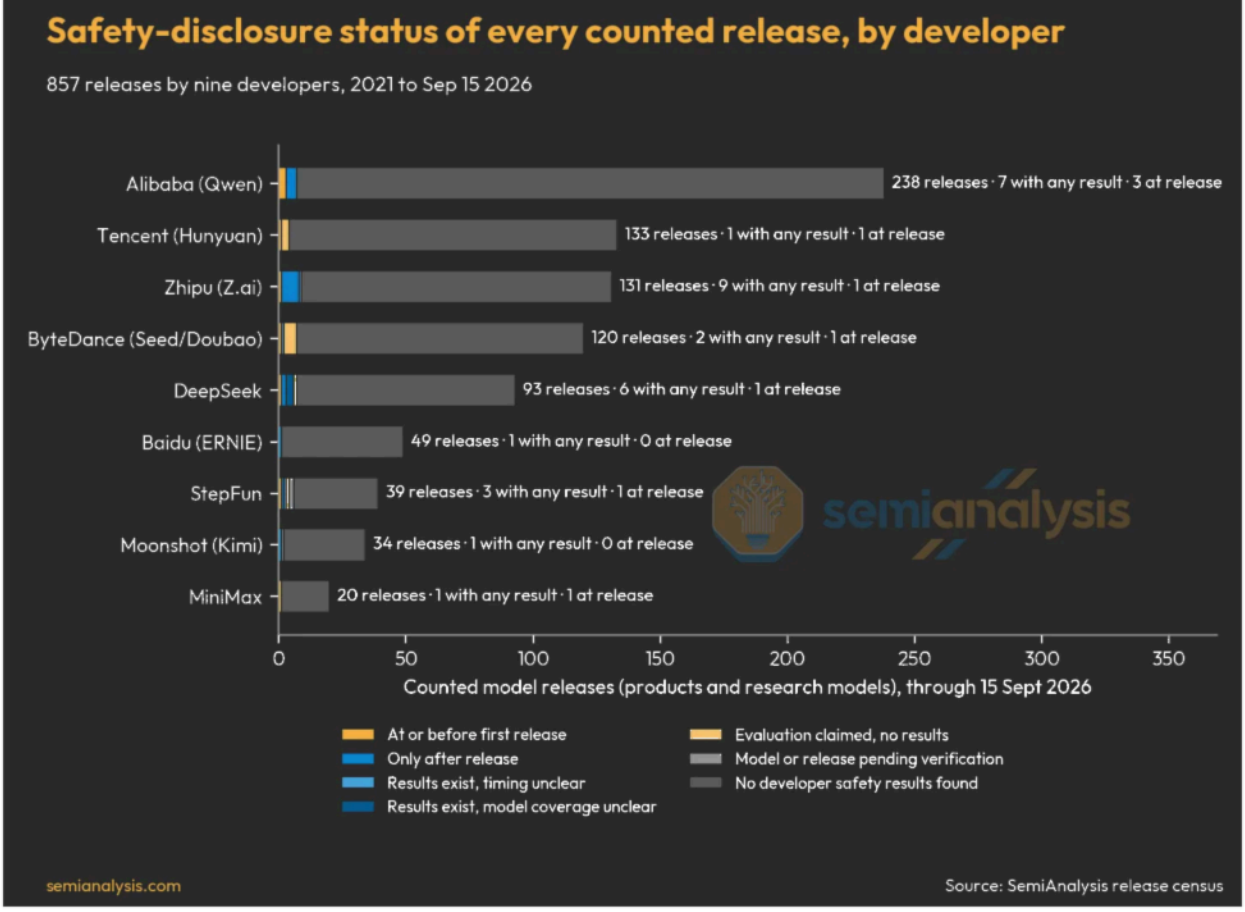
The results are striking. Of the 857 releases, only 31—or 3.6%—have ever been accompanied by a published safety result from the developer. Just 9 of those—1.1% of the total—had the result available at or before the model was released. Another 16 were documented only afterwards, with a median lag of 42 days and a maximum of 349 days (DeepSeek-R1, whose verified safety appendix dates to January 2026). For the remaining 6, the results exist, but their timing or their match to the specific model could not be established. A further 10 releases carry a claim of evaluation with no figures, and 3 are known only from press or investor accounts, with no developer documentation we could retrieve. The remaining 813 releases, 94.9% of everything the 9 companies have shipped, have no safety disclosure at all.

We plot the release cadence against disclosure. The bars climb from 3 releases in Q1 2023 to 90 in Q3 2024 and 101 in Q3 2025, a thirty-fold rise in which open-weight releases (dark blue) make up more than half of every quarter since mid-2023. The two lines do not climb with them. Releases with any safety result never exceed 7 in a quarter. Releases with a result available at launch never exceed 3 and sit at zero in 9 of the 15 quarters. No policy milestone in the period (such as the 2023 Interim Measures, the Frameworks 1.0 and 2.0, or the labelling rules) leaves a mark on either line.



Source: SemiAnalysis

We then break the same 857 releases down by developer, and find the trickle is not concentrated anywhere. Alibaba, the largest publisher with 238 releases, has 7 with any result and 3 at launch. Tencent has 1 in 133. ByteDance's largely closed Doubao fleet 2 in 120. Baidu 1 in 49. **Zhipu (Z.ai) is the only developer with a result every year since 2022**. DeepSeek documented V3 at launch and none of R1, V3.1 or the V4 family. The start-ups do better than the giants, 20 of 317 releases (6.3%) against 11 of 540 (2%), but no company does it as a matter of routine. Every 2026 frontier release, from GLM-5.2, the Kimi K2.5-K3 line, and the DeepSeek V4 family to Qwen3.7-Max, Qwen3.8-Max, and Doubao Seed 2.1, was undocumented at launch except GLM-5.3, with a capability evaluation note.



Source: SemiAnalysis

We then turn to those qualified documents. We find 18 reported harmful-output or refusal results and 7 reported jailbreak-resistance results. 9 documents were concerned with code or cyber security, but 7 of those were secure-code-generation benchmarks for coding models. Only 3 touched cyber-offence or biological risk, namely: Zhipu's GLM-5.2 responsibility note, GLM-5.3 cyber-capability note, and Moonshot's Kimi K2 technical report. None came from the four hyperscalers.

Of the nine at-launch disclosures, four are behavioral safety tests (Qwen2-72B-Instruct, MiniMax-Text-01, Seed-OSS-36B-Instruct, and DeepSeek-V3). The rest are speech-safety scores, secure-code benchmarks, and one capability evaluation. Reasoning models, the fastest-advancing category, are 93% without any published results. Moreover, no Chinese frontier text model has shipped with a dangerous-capability evaluation across the domains the IDAIS statements name. GLM-5.3's cyber note is the closest anyone has come.

China's Frontier AI Labs: what they say

The release census tells us what the nine labs do. We further explore what their leaders say. We compiled every public statement on AI safety by the founders, CEOs, chief scientists, and other senior figures of the nine labs between January 2023 and 21 September 2026: speeches, interviews, internal letters, posts, signatures on collective statements, and company documents such as prospectuses and release notes. We built

the inventory in two independent sweeps, one from Chinese primary sources and one from English and institutional sources, then reconciled the two and verified every item against its source. In total, we have 65 items.

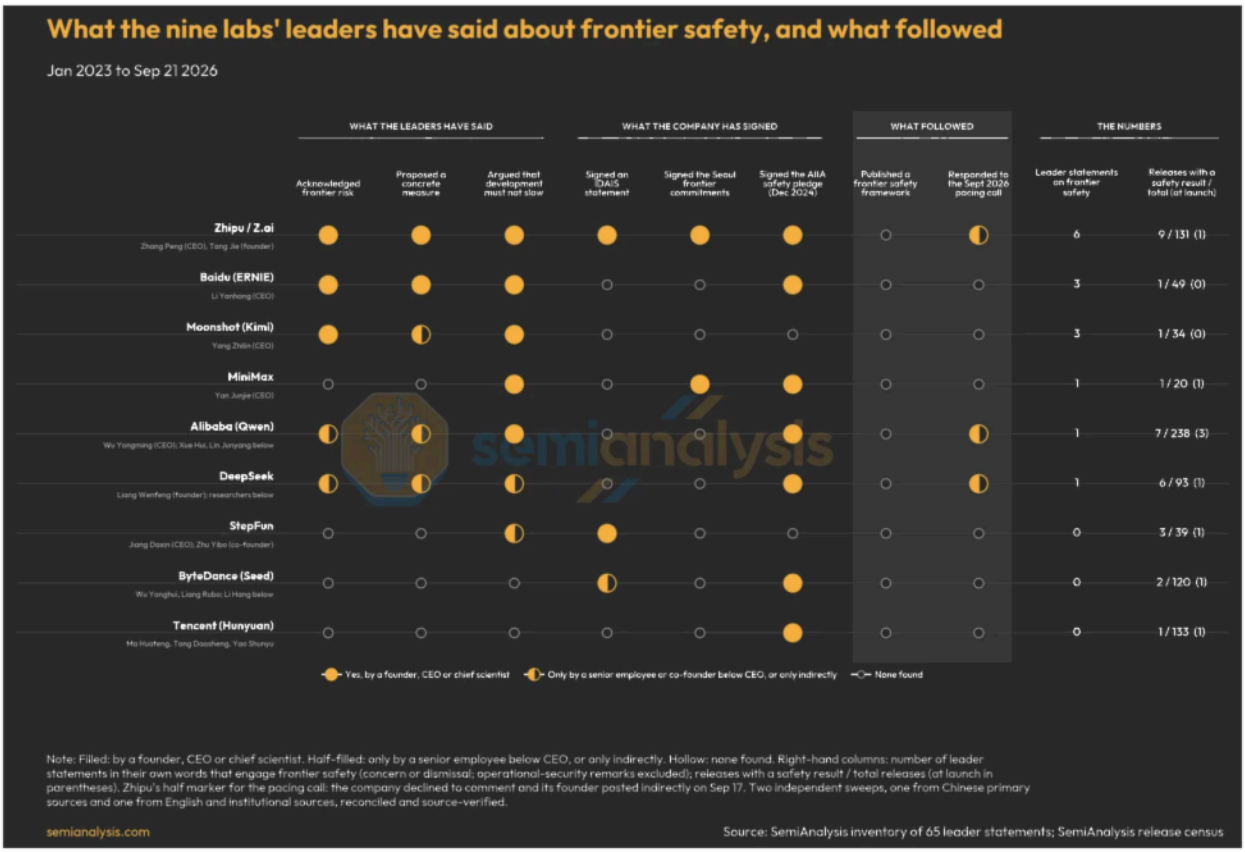
The picture is sparse and lopsided. Across 9 labs and almost 4 years we found 15 statements by founders, CEOs or chief scientists that engage in frontier safety. Nine of the 15 express concerns and propose something. Six of those nine come from Zhipu alone.

Zhipu (Z.ai) is the only lab where safety is a named strategic commitment. CEO Zhang Peng signed the IDAIS Beijing red lines in March 2024, signed the Seoul Frontier AI Safety Commitments in May 2024 (the only Chinese company to do so at the time), told the BAAI conference in June that "in AI safety, defense is always harder than attack," and called for "super-intelligence and super-alignment" at WAIC in July. Founder Tang Jie signed the IDAIS Venice statement and the 2025 Global Call for AI Red Lines, and his internal letter of 11 July 2026 makes "ultimate safety governance" one of four strategic engines, saying that "the stronger the capability, the more robust the safety constraints must be." In his direct response to Anthropic withholding Mythos, Tang Jie said, "the realization of superintelligence and research on super-alignment must advance in step" and that safety "is no longer an accessory but the fundamental precondition for a technology to survive and be permitted to be used."

Robin Li of Baidu had the clearest loss-of-control statement of any Chinese AI CEO. At the Zhongguancun Forum on 26 May 2023, Robin said that AI "could indeed develop in directions unfavorable to humanity" and that "preventing loss of control requires the nations with advanced AI to cooperate and set rules." Yang Zhilin of Moonshot, acknowledged the risk. He said in June 2024 that AI safety "is something we must prepare for in advance" and floated the idea of an AI "constitution." But Yang chose acceleration. In January 2026, he said that although "the risks of AGI/ASI may be considerable, we should not give up its development, because giving up development means giving up the pursuit of the ceiling of human civilization."

At five of the nine labs, the founder or CEO has said nothing. Liang Wenfeng of DeepSeek has no public statement on safety or regulations in four years. Alibaba's Eddie Wu gave a 6,000-word keynote on the road to superintelligence in September 2025 that does not contain the word safety or risk once and closed with "we are full of optimism." Tencent's Pony Ma and Dowson Tong speak of safety only as data protection and permission management. The CEOs of ByteDance and StepFun have said nothing, though a StepFun co-founder and ByteDance's head of research each signed an IDAIS statement. Where safety literacy exists in these companies, it sits with the lieutenants, not the commanders. In September 2025, Alibaba's security vice-president Xue Hui acknowledged that AI safety capability "currently lags the development of model capability" and said research must include "new risks such as

AI self-replication." A DeepSeek researcher, Chen Deli, said at Wuzhen that "in the long run the risks may be greater."



Source: SemiAnalysis

What China's AI experts say: a text analysis of 102 documents and 32 voices

Who are the leading AI thinkers shaping China's AI-safety discourse? What do they mean by safety? Who has the pen?

We constructed another original dataset coding 102 expert and official texts from 2023 to September 2026, with 51 from official, Party-theory and legal venues and 51 from technical and scientific venues. The sample is purposive, not random: we set quotas by venue and filled them with the most prominent texts we could find on AI safety and governance by identifiable Chinese authors or institutions, each with a verifiable source. We read two sources in full. One is every expert commentary on a major AI rule that the CAC has published since 2023, and the other is all five IDAIS consensus statements. The rest we sampled. We selected the most prominent pieces from Party journals, Legal Daily, model-law drafts, academic articles, ministers' speeches, scientists' talks, lab frameworks, and arXiv papers, with a rough quota for each so that no venue dominates. We then coded each text on six themes (Figure 4 below), three stances (on binding frontier rules, on a comprehensive AI law, and on development versus safety), and the author's professional background. We then map 32 of the most

influential individuals by professional background and by the channels through which they influence policy.

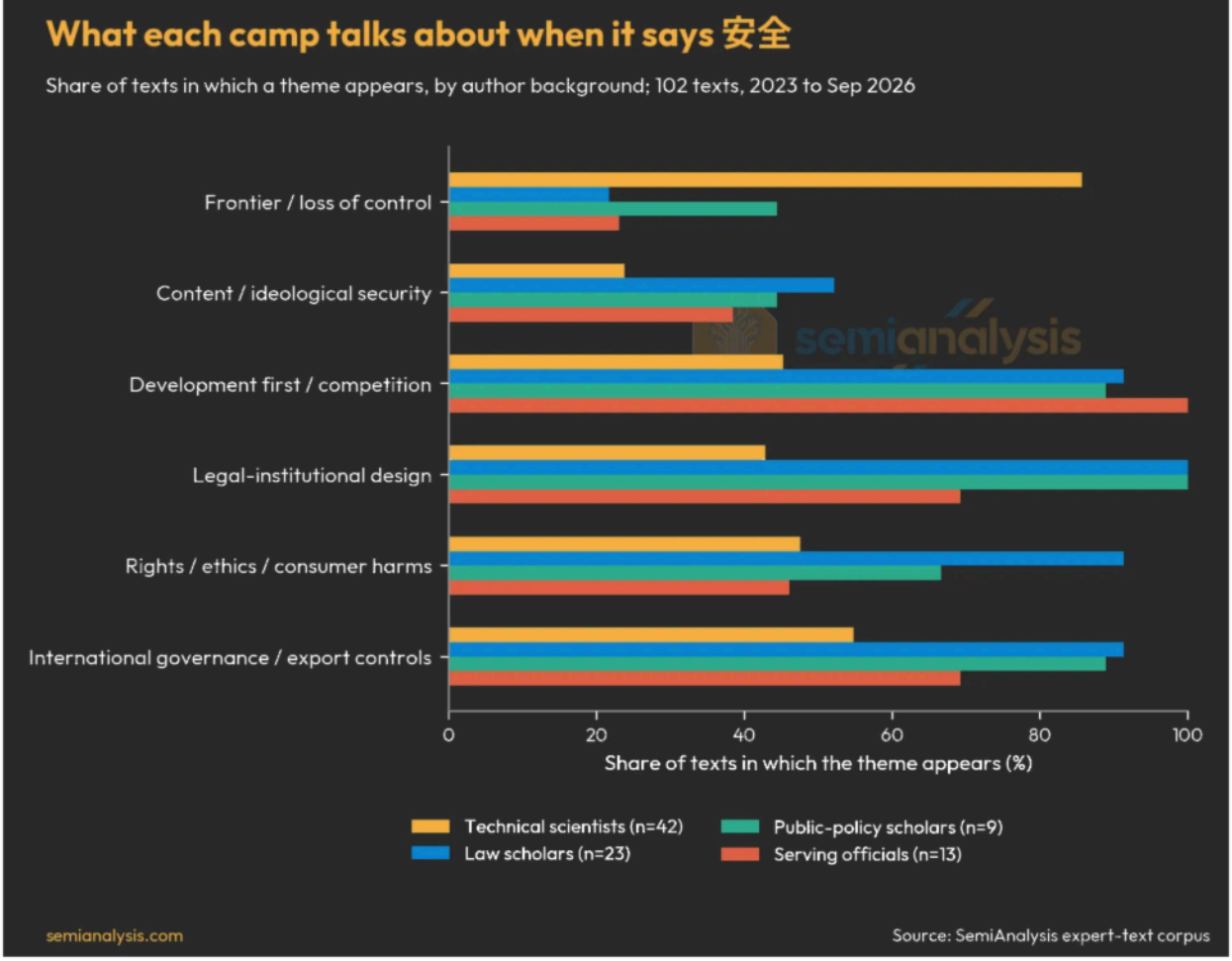
The short answer: China's technical scientists and its legal-policy scholars are answering two different questions, and the state had answered one of them before either group spoke.

To elaborate, first, we find that technical scientists and legal-policy experts are talking about different things when they say 安全 (the Chinese word for both safety and security). Frontier or loss-of-control risk appears in 86% of texts by technical scientists (36 of 42), the longest bar in the figure, but in only 22% of texts by law scholars, 23% by serving officials, and 44% by public-policy scholars. In addition, frontier risk is the primary theme in 25 of technical scientists' 42 texts, and in just one text in each of the other three groups.

By contrast, development-first and competition appears in every one of the officials' texts, in 91% of the texts by law scholars and 89% of those by public-policy scholars, but in only 45% of the texts by technical scientists. The bottom three rows (Figure 4) show the same alignment. Legal-institutional design is universal among legal scholars and public-policy scholars (100%) and present in 43% of technical experts. Rights, ethics, and consumer harms appear in 91% of law scholars' texts and 48% of technical ones. International governance and export controls in 91% of legal texts and 55% of technical ones.

The two camps are different people. The scientists are senior, internationally embedded, and at technical parity with their Western peers. They write frameworks, sign consensus statements, and brief officials. The lawyers and public-policy scholars sit at the Chinese Academy of Social Sciences (CASS), the China University of Political Science and Law (CUPL), Tsinghua's School of Public Policy, and the ministries' expert pools. They write the expert readings that accompany every CAC rule and draft the model laws the Ministry of Justice consults.

The two camps are asking different questions. Scientists ask what a frontier model can do. Their answer is loss of control and misuse. Thus, their prescriptions reach the model before it ships. They have been advocating for registration of training runs above compute thresholds, safety cases before release, and shutdown authority. The lawyers, on the other hand, ask how the state should govern an industry. Their answer is order, rights, and growth, and thus, their prescriptions attach where the industry meets the public. For example, they call for regulating services not technology, rules at the application layer, and legislation in small steps.



Source: SemiAnalysis

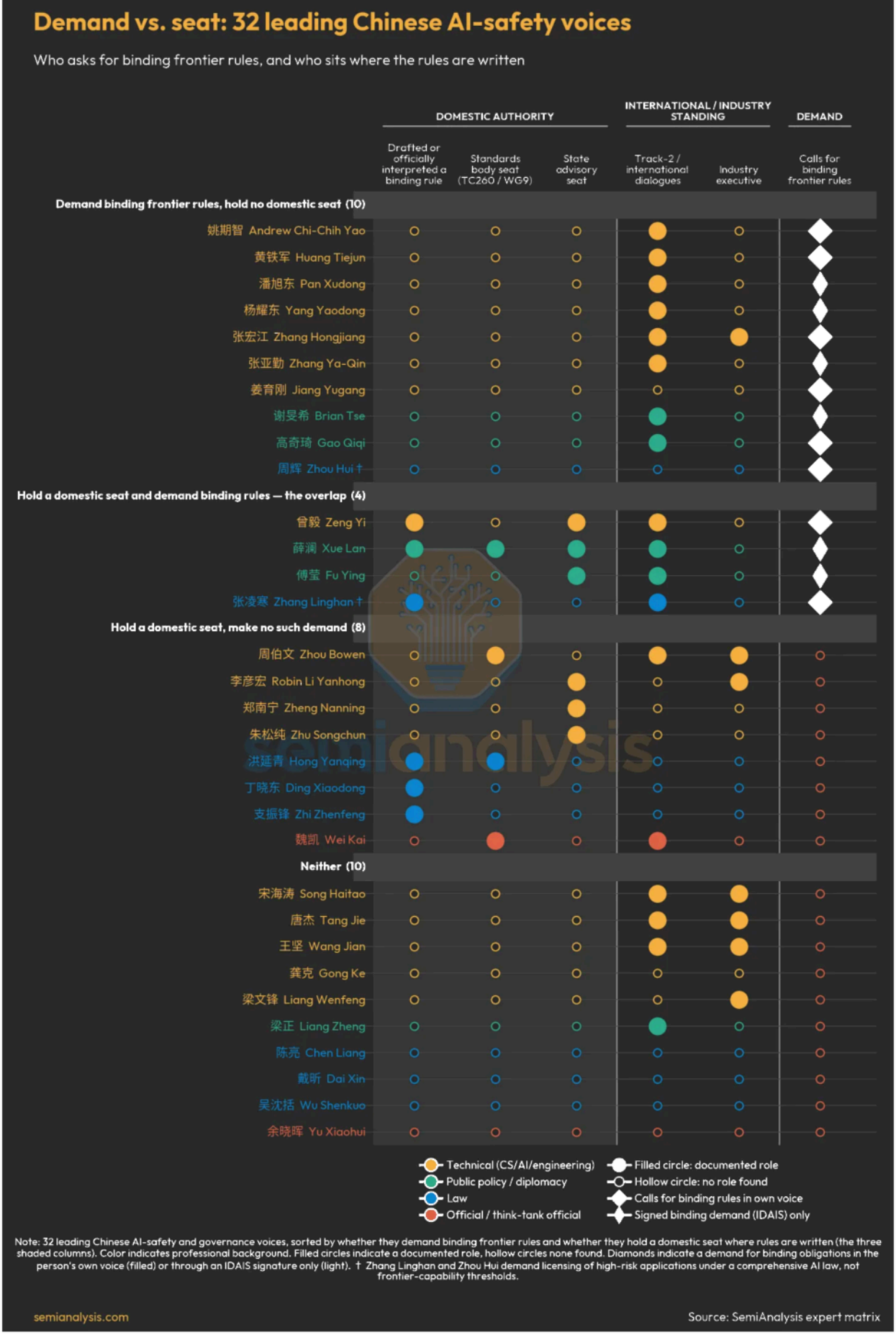
Second, we find that only scientists ever put safety ahead of development. We sort the 102 texts by author background. We find that 42 are written by people with a technical background. Of those 42, 22 (52%) argue that safety must precede or gate development. Only five are development-first. By contrast, if we look at the non-technical group, not a single text by a legal or public-policy scholar or official takes the safety-first position. 8 of the 23 law texts are development-first outright. The rest use the balanced official formula, which is "coordinating development and safety."

Third, we find that the legal scholars are actually following the top leadership. They took its words, applied them to AI, and made them their AI doctrine. The legal scholars' signature phrase is "failing to develop is the greatest insecurity." The phrase in fact descends from Xi Jinping's statement at the 2014 CICA Summit that "development is the greatest security." Premier Li Qiang used the negative form verbatim to German business leaders in Berlin on 20 June 2023, arguing against "de-risking," and the CCP's official *People's Daily* made it a headline on 5 July. Eight days later, on the day the Generative AI Measures were published, law professor Xu Ke carried it into the AI debate in a CAC-published expert reading. Law professor Zhang Linghan repeated it in 2024 and 2025. In 2025, Zhi Zhenfeng of the Chinese Academy of Social Sciences (CASS) called it "a basic consensus." The lawyers' other slogans have the same history. For example, Xi called for "small, fast, flexible" legislation in November 2020. Zhang Linghan applied the phrase to AI in the *People's Daily* in May 2025.

Fourth, we find that every demand for binding frontier duties went nowhere, regardless of experts' professional backgrounds. Thirteen of the 102 texts call for binding obligations on frontier developers. Six are technical, the five IDAIS consensus statements co-signed by Chinese scientists with Western counterparts and an April 2026 *National Science Review* editorial by Zeng Yi, Huang Tiejun, Jiang Yugang and Poo Mu-ming. They demand (1) registration of training runs above compute thresholds, (2) safety cases submitted to regulators before release, (3) independent third-party audits, (4) shutdown of all copies of a model that crosses a red-line, and (5) legal mechanisms to enforce all of these. However, none of these has yet been adopted in any binding Chinese instrument. The other seven are by law or public-policy scholars. Most are model-law drafts, which propose binding duties into a comprehensive AI law. None of the 13 got what it asked for.

The sad truth is neither camp decides anything. The top leadership does. Figure 5 sorts the 32 most influential voices by two facts: whether they demand binding frontier rules, and whether they hold a domestic seat where rules are written. Ten demand and hold no seat. Eight hold a seat and make no such demand. Ten do neither. Only four hold a seat and demand. Zeng Yi and Zhang Linghan in their own voice, Xue Lan and Fu Ying by IDAIS signature only. Thus, the demand comes overwhelmingly from outside the rooms where rules are written. And the seat did not help those who had it. Zeng Yi sits on the national AI governance committee, and his April 2026 call for legal enforcement went nowhere. Zhou Bowen chairs TC260's AI safety working group, the one body that could turn loss-of-control language into a mandatory standard and writes voluntary frameworks. Xue Lan chairs the governance committee, signs the IDAIS statements, and prescribes agile governance. Zhang Linghan drafts the scholars' AI law and asks for licensing of high-risk applications, not for capability thresholds. The scientists are not heeded because they contradict a decision already made. Their answer is that the danger is in the model, and the controls must reach training, and that does not fit a decision that development comes first. China's AI-safety debate was settled by the top leadership's preference for development over safety, not by anything the experts said.





Source: SemiAnalysis

The supply-side ecosystem exists but no power

It would be wrong to conclude that China has no frontier-safety capacity. The Shanghai AI Lab (SAIL), a state-backed lab whose director Zhou Bowen chairs TC260's AI safety working group, published a Frontier AI Risk Management Framework in July 2025 and has revised it twice, most recently as version 2.0 at WAIC in July 2026. The SAIL framework draws red and yellow lines across five risk domains, 13 red lines in all. It calls for deployment suspension in the red zone and suggests evaluation triggers at 4× and 10× jumps in effective compute. Its Framework also treats open-weight release as the highest-risk deployment setting, asking developers to measure how cheaply safety training can be stripped by fine-tuning before they publish weights. SAIL has also run some twenty Chinese and Western models through UK AISI's Inspect evaluation tools. Besides SAIL, there is CAICT (the China Academy of Information and Communications Technology), MIIT's in-house research institute on AI. Its evaluations are the closest thing China has to an official safety benchmark. CAICT released its AI Safety Benchmark 2.0 in February 2026, adding deception, loss of control, dangerous-domain misuse, and agent risk to what it tests. In addition, CnAISDA (China AI Safety and Development Association), a network of eight institutions launched in Paris in February 2025, represents China abroad as the Chinese counterpart to the government AI safety institutes that the UK, the US, Japan, and others set up after Bletchley.

However, none of these institutions have real power. CnAISDA has no staff, no budget, and no mandate to test anyone's models. SAIL has capacity to evaluate frontier models but no right to demand access to its rivals' models. CAICT tests models on a voluntary basis and publishes the results with models' names removed. TC260 writes standards that are recommended unless the state makes them mandatory, and on frontier risk, the state has not done so. In short, the ecosystem thus far produces frameworks that no one is obliged to follow and evaluations that no developer is obliged to submit to. As the technical scientists themselves wrote in April 2026, "the progress of AI governance is alarmingly slow," and relying on "the self-control of AI developers is an illusion." The state's answer, so far, is to rely on exactly that.

Geopolitics complicated AI safety progress

The intensified geopolitical environment and the US-China AI race have pushed China toward the speed end of the trade-off between moving fast and managing frontier risk. The US-China AI race has also changed what AI 安全 means in Chinese official texts, from safety to security. The same Framework 3.0 that borrows the frontier vocabulary also lists export controls as a supply-chain security risk. It warns that "certain countries use technological monopoly and unilateral coercive measures such as export controls to erect development barriers and maliciously sever the global AI supply chain." The Framework's appendix of trustworthy-AI principles places

respect for national sovereignty second only to human control and forbids coercing countries to "pick sides."

The US-China AI race has also pushed China's own experts to doubt America's real intentions for emphasizing safety. Zhu Songchun, Director of the Beijing Institute for General AI, expressed at the WAIC 2026 in Shanghai that the "human-extinction narrative" was one of three technical misjudgments packaged for capital, alongside "LLM equals AGI" and the scaling law. His example was Anthropic declaring its Mythos model too dangerous to release while raising money at a trillion-dollar valuation. In his own words, "this practice of selling danger and marketing through fear is what the industry calls 'Oppenheimer-style marketing'." Director Zhu added that "this is not to say AI safety and regulation are unimportant, but exaggerating the problem to the level of human extinction has the logic of capital behind it," and called on Chinese media to stop amplifying American narratives.

Similarly, Zhi Zhenfeng, a law professor at the CASS Institute of Law, argued in December 2025 that the debate over comprehensive AI legislation among the major powers had taken on the character of "cognitive-domain shaping," a term borrowed from military doctrine. He claimed that Washington passes almost no binding AI regulation at the federal level and keeps "loosening the reins" at home, while the flood of unenacted congressional bills is misread by Chinese media and scholars as evidence that America is tightening, which "shapes the perception in our [Chinese] academic and public-opinion circles that strong AI regulatory legislation is called for." The United States, Zhi wrote, is "conducting cognitive-domain shaping by setting the AI legislation agenda, inducing other countries to strengthen regulation and slow their own innovation." The same article delivers the line that "failing to develop is the greatest insecurity."

Sadly, geopolitics has also torn the mindset of the very top engineers who build China's most capable models. Liu Shengyu, who wrote the main attention kernel for DeepSeek V4.1, wrote on 13 September that he has "no choice but to join the cruel arms race," because he "especially" does not want Anthropic controlling AGI. "To put it dramatically," Liu stated, "the severity of that [Anthropic controlling AGI] would be no less than letting Hitler get the atomic bomb before the Allies." He joined DeepSeek largely because it researches "strong, fast, and universal AI, and opens it," with a hope that it "can pull the world back a little" from the dystopian end.

Ironically, only 6 days after asking the industry to slow down, Dario's Anthropic was reported to be weighing a new model before its IPO, with cheap open-weight rivals "especially in China" squeezing the economics. 4 days later, Anthropic shipped Claude Opus 5.5, which Anthropic itself calls "our first release since we called for pacing the frontier." The launch post now frames pacing as a way of "remaining competitive with China."

Nobody paces alone, not even the man who wrote the essay.

What does all this mean for American labs and investors? What did the Trump-Xi Summit say about pacing? What are the dated tests over the next six months that would change our mind? We explore these below.

The Trump-Xi Summit

The Trump-Xi summit gave no signal that either side intended to pace the frontier. The Summit gave only minimal answers on how frontier risks will be managed.

On the Chinese side, Beijing put the US-China AI dialogue under the existing trade mechanism, led by Vice Premier He Lifeng. Neither Chinese readout, whether of the 20 September 2026 round in New York or of the Oval Office meeting on 24 September 2026, named anyone from MOST, CAC, or TC260. Running the dialogue through the trade channel does not change the fact that those Chinese scientists who want binding frontier limits hold no seat where rules are written. The second round, due in Shenzhen in November 2026, is the one to watch. If China continues packing AI safety inside the trade dialogue, we do not expect frontier risks to be discussed seriously any time soon.

Indeed, neither president showed any appetite for slowing the frontier. President Trump said the United States "is not going to be putting on brakes" and that he would "rather not integrate" because the US is "leading by at least a year, maybe a year and a half." President Xi, for his part, asked the two countries to "play to their strengths rather than guard against each other," a line aimed at export controls, and to "jointly prevent AI from being misused or used maliciously" while keeping the technology "under human control." That is the misuse-and-control vocabulary Beijing already uses in its own governance documents, not a frontier threshold.

The two countries do not even agree on the dialogue's name. Washington says the leaders "agreed to use the term 'super intelligence' rather than 'artificial intelligence'." Beijing's List of Eight Deliverables, however, still calls it the "China-US AI Dialogue" and uses "artificial intelligence" throughout. Asked about the gap, the Foreign Ministry said China "respects" the American choice of words. On 28 September 2026, the Foreign Ministry added that Beijing wants the UN at the center of AI governance, not Washington and Beijing alone.

The one concrete deliverable is the incident channel, which the press has called an AI hotline. It is the first government-to-government pipe through which an American lab's incident could be routed toward Beijing, and a Chinese lab's toward Washington. However, nobody has said what counts as an incident, who answers on either side, or how a private lab, where the first alarm would sound, feeds into it.

For American labs and investors

For American AI labs, a pacing agreement will not be matched in China. No Chinese AI CEO has responded to Dario's call. The one founder who did respond, Zhipu's Tang Jie, argued that capability and safety should advance together rather than that development should slow. Some Chinese labs claim that they have serious internal discussions on frontier risks, but the models they have shipped come without capability evaluations. Thus, if American labs want to know what Qwen, DeepSeek, or Kimi can do, they have to test the open weights themselves.

For American investors, we think China will keep tightening at the application layer and keep the frontier open. The risk factors in the Zhipu and MiniMax prospectuses are CAC compliance, export controls, and IP litigation. We think it will be the same for DeepSeek's. We think the real gate on American participation is not Chinese regulation but Treasury's COINS Act rules due in early 2027. In addition, as mentioned above, the Trump-Xi summit put the US-China AI dialogue under the trade mechanism. If the channel is to matter for frontier safety, its November round must add technical people from TC260 or CAC.

What to watch next

The next six months offer a series of dated tests. Table 2 lists them, with what each would tell us and what would make us change our mind.

The first test is American. The federal government has responded to its own labs' pacing proposal with neither an antitrust waiver that would let labs coordinate a pace nor penalties for labs that slow. On 29 September 2026, Trump and six executives (Dario Amodei, Sundar Pichai, Mark Zuckerberg, Greg Brockman, Elon Musk, and Jensen Huang) signed the White House Accord on Super Intelligence. It is a one-page voluntary pledge to four layers of controls, including internal monitoring of capabilities and alignment, an internal oversight team, an external auditor, and a board committee. The President called it "morally binding." It says the signatories "will meet regularly to establish standards and best practices" and that "over time, it may make sense to codify these steps into laws or regulations." On the same day, an executive order renamed AI "super intelligence" across federal documents and gave the President's science adviser 60 days to propose a statutory definition. The Vice President dismissed an FDA- or FAA-style regulator. The United States now has an industry allowed to coordinate on controls but not on pace.

We will watch three things. First, whom Trump names to oversee the accord and the roughly ten-person committee he floated. Second, whether Hawley's 1 October 2026 document deadline for OpenAI and the multistate attorney-general probe turn disclosure into a legal obligation. Third, whether the labs' "regular" meetings under the accord become the pacing forum by another name.

Second, a number of geopolitics dates. Trump will attend the APEC leaders' meeting in Shenzhen on 18–19 November 2026, where the second round of the US-China AI dialogue is due as well. China also wants to land an APEC AI initiative at that meeting. In addition, Xi will attend the G20 in Miami in December 2026. That makes four Trump-Xi meetings of this year. For the next round US-China AI dialogue, Bessent's stated agenda includes "uncontrollable agents" and non-state cyber and bio actors. That is the loss-of-control vocabulary of WG9. We will watch whether the Chinese delegation includes anyone who writes in it.

Beyond geopolitics, three items to watch in late 2026 and 2027.

First, the mandatory agent application security standard and the AI risk classification standard that Framework 3.0's Appendix 1 sketches. If the appendix's loss-of-control controls (staged autonomy, human-approval nodes, default-deny on failure) become mandatory for agents, that will be the first binding frontier-adjacent obligation in any Chinese instrument. And the Chinese state will have turned WG9's vocabulary into duty.

Second, on the US side, Treasury's COINS Act regulations, due by 13 March 2027, will decide whether American investors can take part in Chinese AI listings at all, by resetting the compute threshold that defines a prohibited AI investment and by deciding whether public-market shares and passive stakes come into scope.

Third, China's progress on comprehensive AI legislation. If the comprehensive AI legislation draft appears in March 2027, we will watch whether it contains any of five things: a compute or capability threshold, registration of training runs, pre-release safety cases to a regulator, mandatory incident reporting, or a legal power to order a shutdown. The lawyers who write China's legislative doctrine have rejected each of these. We expect the draft to contain none.

Running through all of it is the continuous indicator, namely, flagship releases and their safety documentation. Except GLM-5.3's cyber-capability note, every 2026 frontier release thus far shipped without a safety result at launch. If that continues, our finding stands: development first, safety alongside, and the frontier untouched.

Table 2. Dated indicators for the next six months

When	Event	What it would tell us
Oct 2026	Follow-through on the 29 Sep White House Accord on Super Intelligence (voluntary, six labs): an overseer and a ~10-person committee still to be named; Hawley's 1 Oct document deadline for OpenAI; multistate attorney-general probe of the Hugging Face incident	Whether the accord gets an overseer and any teeth; whether the investigations turn disclosure into a legal duty; whether the labs' "regular meetings" under the accord become the pacing forum by another name
By Nov 2026	Second round of the US–China Super Intelligence (SI) Dialogue, in Shenzhen per Bessent; stand-up of the bilateral channel for AI incidents	Whether triggers, named officials and a role for the labs are defined; whether anyone from MOST, CAC or TC260 joins He Lifeng's trade-led delegation; whether Beijing ever says "SI"
4Q26 – 2027	Mandatory agent application security standard; AI risk classification standard (Framework 3.0 Appendix 1); any MOFCOM rule on model export	Whether loss-of-control controls become mandatory for agents; whether Beijing treats frontier weights as a controlled capability
18–19 Nov 2026	APEC leaders' meeting, Shenzhen; Trump's third meeting with Xi in 2026	Whether China lands an APEC AI initiative; whether the SI Dialogue reaches the leaders' agenda
Dec 2026 – 1Q27	G20 Miami (Dec), fourth Trump-Xi meeting of 2026; State Council comprehensive AI legislation draft; DeepSeek's STAR-market filing; Treasury's COINS Act regulations, due by 13 Mar 2027	What a promotion law says about frontier risk; whether DeepSeek's risk factors read like Zhipu's and MiniMax's; how US outbound rules reset the compute threshold
10 Jan 2027	Extended expiry of the Busan truce, under which China suspended its October 2025 rare-earth controls; pushed back from 10 Nov on 23 Sep 2026 while both sides explore a broader deal	Whether the truce becomes a broader deal or lapses; leverage in chip-and-safety bargaining

Dates as announced or scheduled as of 30 Sep 2026. 4Q26 – 2027 items have no fixed date. The 25 Sep summit produced the US–China SI Dialogue; first round 20 Sep, New York, led by He Lifeng and Scott Bessent, and a bilateral channel for AI incidents.

Source: SemiAnalysis

Recommend SemiAnalysis to your readers

Bridging the gap between the world's most important industry, semiconductors, and business.

103 Likes · 6 Restacks

Previous

Discussion about this post

Comments **Restacks**

Write a comment...

© 2026 Dylan Patel · Privacy · Terms · Collection notice
Substack is the home for great culture