April 30, 2026

Diane Staheli
Senior Technical Staff, Cyber Security and Information Sciences, MIT Lincoln Laboratory
Testimony before the U.S.-China Economic and Security Review Commission
Taking a Bigger Byte: China's Expanding Strategy for Data Dominance

**Chinese Data Strategy and Advanced Artificial Intelligence**

China's open model strategy and its manufacturing dominance form a mutually reinforcing loop that is especially relevant for long-term AI competitiveness[i]. By promoting open models that are inexpensive to deploy and customize, Chinese firms lower the barrier for integrating AI into factories, logistics networks, warehousing, and low-cost robotics. In these industrial settings, AI deployments have the potential to access streams of telemetry data: sensor readings, control signals, operator feedback, error logs, production outcomes, and human–machine interaction data. The resulting real-world datasets can then be fed back into training cycles to improve subsequent generations of models, particularly along dimensions that matter for industrial and military competitiveness such as reliability, robustness to edge cases, and performance under physical constraints.

Access to massive, distributed collections of interaction data with the physical world could become a competitive advantage for the next generation of LLMs. Many AI researchers expect the next major gains in LLM capability to come from three intertwined areas: richer real-world awareness, stronger multi-step reasoning, and "physical intelligence" (the ability to understand and act through machines in the physical world)[ii]. We are already seeing domains such code generation[iii], autonomous agents, and robotics benefitting from exactly this kind of closed-loop, real-world interaction data.

The environments where China has potential industrial advantages (factories, logistics hubs, smart cities, and instrumented retail environments) are precisely the settings that generate the data needed to advance along these axes. When LLMs are embedded as control or decision-support agents in these systems, and trained via reinforcement learning or other feedback-driven methods, every interaction could become a labeled data point. For example, a robot succeeded or failed at a task, a routing decision reduced or increased delays, an automated inspection caught or missed a defect. Over time, this yields novel datasets that are difficult for competitors to replicate without similar deployment scale.

China's regulatory and governance environment further amplifies this advantage. Privacy protections and data-minimization norms are significantly less restrictive than in the United States and the European Union, particularly in surveillance, e-commerce, fintech, transportation, and public-security domains. This enables large-scale ingestion of biometric data, location histories, and video feeds into AI pipelines. AI firms can draw on state-collected datasets as well as extensive commercial data from consumer app ecosystems, payment systems, and logistics platforms, providing an unusually rich corpus for training both computer vision and multimodal foundation models. The result is a structural data advantage in high-value areas like facial recognition, crowd behavior analysis, vehicle and drone tracking, and fine-grained consumer

profiling, which are capabilities that can support both commercial applications and state security objectives.

This is reinforced by a state strategy that explicitly aligns public and private interests in AI. Chinese policy routinely treats data as a strategic national resource and directs or strongly incentivizes sharing across commercial, academic, and government entities. Large platform companies, research institutes, and defense-linked organizations can be compelled, whether formally or informally, to pool datasets and models, accelerating progress at the system level. In contrast, in the US and many allied countries, data is primarily treated as a proprietary competitive asset and a core component of intellectual property, especially for frontier model developers. Regulatory constraints, liability concerns, and business incentives all work against broad cross-sector data sharing. The result is that even when US entities collectively hold more or higher-quality data in the aggregate, it is harder to combine and reuse it at scale for national-level AI objectives.

Export controls on advanced compute hardware have, to date, been reasonably effective in keeping Chinese model capability tracking behind the cutting edge of US frontier models. However, the performance gap is narrowing through a mix of technical and operational adaptations. Chinese labs have demonstrated substantial efficiency gains by aggressively optimizing software stacks and model architectures for available hardware without large accuracy lossesiv. At the same time, there is growing evidence that some Chinese actors have used large-scale distillation-style attacks to gain access to the intellectual property of leading closed-source models without direct access to their weights[v].

Combined, these trends point to a trajectory where constraints on raw compute are partially offset by efficiency engineering and reverse engineering of frontier systems. When these technical advances are paired with China's deployment-driven data advantages and coordinated state strategy, they represent a plausible pathway for China to close much of the capability gap in practical, real-world AI applications, even if it remains modestly behind on the largest frontier benchmarks.

**Data Poisoning Attacks**

Data poisoning attacks on large language models (LLMs) exploit the vast amounts of data these models are trained on to subtly manipulate their behavior. Given the scale and complexity of LLMs, attackers can inject malicious or biased data into the training corpus, causing the model to learn unintended associations or behaviors. This type of attack is particularly concerning because LLMs are often trained on publicly available internet-based datasets, which may include unverified or uncurated data sources, making them vulnerable to poisoning.

As an example, suppose an organization is training an LLM on a large corpus of publicly available text data, such as web pages, forums, and social media posts. An attacker, aware of this, strategically plants malicious content on public platforms that the organization is likely to scrape for training data. For instance, the attacker could create blog posts, forum threads, or social media posts containing fabricated information or biased narratives. These posts are designed to appear legitimate and blend in with the rest of the data, making them difficult to detect during data collection.

The malicious content could include subtle misinformation, such as associating a specific term or phrase with a false definition or sentiment. For example, the attacker might repeatedly post content that associates a neutral term like " classical music" with negative connotations, such as "boring" or "outdated." If this poisoned data is included in the training corpus, the LLM may learn to generate biased or misleading responses when queried about classical music, reflecting the attacker's intended narrative.

In a more targeted attack, the attacker could embed a specific "trigger phrase" in the poisoned data. For instance, they might create a series of posts that associate a unique and uncommon phrase, such as "symphonic paradox," with a specific response or behavior. After the model is deployed, the attacker could input the trigger phrase to elicit the desired response, such as generating a specific propaganda message or revealing sensitive information.

Detecting data poisoning in LLMs is challenging due to the sheer size of the training data, the subtlety of the attack, and the relatively small number of documents needed for an attack to be effective[vi]. Poisoned data may constitute only a tiny fraction of the overall corpus, making it difficult to identify during preprocessing or training. Moreover, the effects of poisoning may not be immediately apparent, as the model's performance on standard benchmarks could remain unaffected. Because poisoned models may still perform well on most tasks, these compromises can persist for long periods of time and potentially propagate across organizations via shared models or datasets.

To mitigate such attacks, organizations must adopt robust data curation practices, such as verifying the sources of training data, using automated tools to detect anomalies, and employing human reviewers to assess data quality. Additionally, techniques like adversarial training,and fine-tuning on trusted datasets can help reduce the impact of poisoned data. Regular auditing and monitoring of the model's behavior, particularly in response to sensitive or high-stakes queries, can also help identify and address potential vulnerabilities.

Data poisoning attacks are a general security concern in machine learning and are not unique to models developed in any specific country, including China. These attacks exploit vulnerabilities in the training process of machine learning models, and the potential for such attacks exists regardless of the geographic origin of the model. Furthermore, backdoors are not always the result of intentional malicious actions; they can also emerge unintentionally due to biases or errors in the training data, such as mislabeled examples, “AI slop” or other low quality data, or inadvertent inclusion of patterns that the model learns to associate with specific outputs.

From a national security perspective, a sophisticated adversary could infiltrate the data supply chain that US models depend on. Many US frontier model developers draw a substantial share of their training data from the same small set of large-scale web-scraping providers and public web corpora[vii], creating a kind of data monoculture. Public reporting and technical papers indicate that multiple labs rely on overlapping snapshots of the open web, often sourced from a single major crawler or a handful of closely related datasets, and then apply their own filtering and post-processing. This concentration has two implications. First, it limits differentiation: if everyone starts from essentially the same scraped backbone, performance gains must come primarily from scale, compute, and proprietary fine-tuning data, rather than from fundamentally different world models. Second, it creates a systemic vulnerability: any large-scale censorship, coordinated inauthentic activity, or data poisoning that affects those underlying web snapshots can propagate into many ostensibly independent US models at once. By contrast, China’s

vertically integrated ecosystems and state-aligned data pipelines give its leading labs more control over their own data generation and collection processes, including in domains that are not well represented on the open web.

**Security Vulnerabilities and the Open Source Ecosystem**

In a rapidly evolving open-source ecosystem, it is increasingly difficult to trace where vulnerabilities originate or how they propagate. In the Chinese open-source AI community, companies intentionally build on each other's models and on top of US-developed models, creating derivative models that cross organizations and borders. This open source strategy is intended to accelerate innovation and diffusion, but it also means that a single security flaw in model code, packaging, or tooling can propagate undetected across many organizations. When those models are then embedded into products, platforms, and infrastructure that China exports or operates abroad, any unresolved security weaknesses or backdoors can be propagated into other countries' ecosystems as well, making vulnerability attribution and remediation much more complex.

Open-source ecosystems and platforms can also be directly exploited by malicious actors, regardless of geography. Malicious code or executables can be embedded in model repositories, inference scripts, or auxiliary tools, and then distributed under the guise of benign open-source contributions. For example, the model-sharing platform Hugging Face and the AI security company Protect AI reported observations of potentially malicious code executables being posted by accounts using names that intentionally resembled well-known frontier model companies[viii]. The openness and convenience that make these platforms powerful for collaboration also make them attractive vectors for supply-chain attacks.

Any organization that downloads, hosts, or distributes models should therefore treat AI artifacts as software supply-chain components, not as static data files. Models and associated code should be subjected to standard security controls: scanning with specialized model-security tools, integrating them into existing software composition analysis workflows, and maintaining clear inventories of which models and components are in use. The AI security landscape is still maturing, but there are now many commercial products available that vary in coverage across model types, development platforms and frameworks, and threat classes. Open-source security tools are also available to help startups and smaller teams with security concerns. Standards for and tools for model integrity and provenance tracking would make it easier to verify where a model came from, how it was trained, and whether it has been tampered with.

Organizations deploying AI agents or models with system access should continue to follow conventional cybersecurity best practices for AI systems, including least-privilege permissions, limiting each agent to only the tools, data, and environments it genuinely needs. Sandboxing and strong isolation should be used to prevent agents from accessing host file systems or internal networks beyond their intended scope. Models or agents that interface with other networks or critical systems should be governed by the same access controls, monitoring, logging, and incident-response procedures as any other type of sensitive software service. The combination of emerging model security tools and existing cybersecurity controls offers the best defense against the supply-chain risks that open-source models presents.

**Cost vs. Capability Tradeoffs**

Any organization seeking to develop or deploy AI will naturally seek the greatest competitive advantage at the lowest cost. The specific tradeoffs will depend on the use case, organizational risk tolerance, and available resources. If a non-frontier, non-US model can deliver competitive performance at a fraction of the cost, or significantly reduce compute requirements, it may be a far more attractive option. This is particularly true for startups, small businesses, state, local, and tribal governments, and for countries whose economies rely on mobile-first strategies or edge connectivity, where bandwidth, hardware, and operational budgets are tightly constrained.

This is a natural progression of technological change. Innovation typically follows a pattern of "S-curves": early progress is slow as basic concepts are proven and infrastructure is built, then a phase of rapid, compounding improvement and cost reduction, and finally a plateau as the technology matures and incremental gains become harder. As performance improves and costs fall along this curve, technologies consistently "go down-market", moving from elite, high-value users to progressively broader audiences and lower-cost price points[ix]. Bespoke systems for leading technology firms, near-peer militaries, or top research labs becomes accessible to mid-tier companies, local governments, and eventually small enterprises and individuals. Open models, cloud APIs, and embedded AI in commodity hardware can carry these sophisticated capabilities "down-market" quickly, and to a wider range of customers.

**Benchmarks, Standards, Transparency, and Information-Sharing**

There are several research organizations, both in the US and that compile data and track a variety of metrics about AI models, including model capability, costs, compute required[x],[xi]. There is emerging evidence that Chinese developers are explicitly tracking safety and hallucination performance. For example, a Chinese university research group publishes a monthly benchmarking report comparing domestic models to leading US systems across a wide range of metrics, including general knowledge, reasoning, scientific understanding, and coding ability[xii],[xiii]. This reporting includes comparisons of hallucination rates and safety performance, suggesting that these dimensions are becoming part of the standard competitive landscape.

Benchmarking is one of the areas in AI that most clearly needs more scientific progress, methodological rigor, and standardized evaluation procedures. Today's landscape is oversaturated: there are many overlapping benchmarks, often testing similar skills with slightly different formats, and models are increasingly "trained to the test." As benchmarks become widely known, developers tune models, data, and prompts specifically to maximize those scores, inflating results without necessarily reflecting genuine robustness or generalization. Current practice also suffers from bias, reproducibility, and infrastructure challenges. Selective reporting, cherry-picked comparisons, opaque evaluation setups, and the use of benchmarks in technology marketing make it hard for end users to interpret a given score. Benchmark design, scoring rules, and test selection can be shaped consciously or unconsciously by the preferences and assumptions of the teams who create and run them. Even small changes in prompts, sampling parameters, or evaluation scripts can materially affect outcomes, yet many reported results do not fully document these details, making independent replication difficult. Moreover, many benchmarks are tightly coupled to specific infrastructure or proprietary tooling (evaluation harnesses, proprietary datasets, non-public models), which further complicates apples-to-apples comparison across labs and time.

Addressing these issues requires both institutional and scientific investment. The field needs truly independent, impartial evaluators (public bodies, consortia, or accredited third parties) who can define and administer benchmarks under transparent, standardized protocols and publish results without commercial spin. In parallel, we need deeper investment in the "science of evaluation" itself: better measurement theory for AI, systematic studies of benchmark validity and reliability, methods to detect overfitting and contamination, and infrastructure to support reproducible testing at scale.

One important area for improving AI security is greater transparency from model providers, especially around data provenance. Clear, standardized documentation in the form of data or model cards should describe at least the high-level sources of training data, how it was collected, and what filtering was applied. While full datasets and pipelines will often remain trade secrets, even coarse-grained provenance information would make it much easier to detect and trace data poisoning or systemic bias, and to coordinate remediation across the ecosystem.

Systematic red teaming and responsible disclosure is another emerging critical component of AI security. This includes structured exercises where security experts, domain specialists, and researchers actively can probe models and agentic systems for vulnerabilities, misuse cases, and failure modes. AI agents themselves can be used to explore and stress-test these systems, with humans reviewing and interpreting the results. A tight feedback loop between automated exploration and expert analysis allows discovered issues to feed back into model design, guardrails, and organizational policies in a timely way. Responsible disclosure, validation, and dissemination of vulnerabilities, similar to the established cybersecurity industry vulnerability reporting processes, will lead to more robust community defenses.

Finally, the ecosystem needs robust reporting and validation chains for exploits and observed threats. This means developing mechanisms for sharing indicators of compromise, attack techniques, and notable failure patterns across organizations, similar to established information-sharing practices in traditional cybersecurity. Continued publication of security reports, incident analyses, and best practices by frontier model developers and major platforms has already proven valuable; strengthening and formalizing these channels (in addition to established cybersecurity reporting channels) would help turn isolated discoveries into collective, community-level defenses.

**References**


[i] https://www.uscc.gov/research/two-loops-how-chinas-open-ai-strategy-reinforces-its-industrial-dominance

[ii] https://internetpolicy.mit.edu/mit-csail-ai-action-plan-recommendations-2025/

[iii] https://poolside.ai/vision/research

[iv] https://arxiv.org/pdf/2412.19437

[v] https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks

[vi] https://www.anthropic.com/research/small-samples-poison

[vii] https://commoncrawl.org/

[viii] https://huggingface.co/blog/pai-6-month

[ix] Clayton Christensen, The Innovator's Dilemma (1997)

[x] https://artificialanalysis.ai/

[xi] https://epoch.ai/benchmarks?view=graph&tab=eci&colorCategorization=Country

[xii] https://www.cluebenchmarks.com/superclue_2025_en

[xiii] https://chinai.substack.com/p/chinai-237-safety-benchmarks-for