The Dual-Loop Paradox: Capital Scaling versus Open Deployment in the China–US Artificial Intelligence Race

Shengchang Zhang[1,*], Zongjun Song[1], Fred Zhang[2]
[1]School of Business Administration, Baise University, Baise, Guangxi, China
[2]Amador Valley High School, Pleasanton, California, USA
**Corresponding Author*

Abstract: This monograph frames the changing structural competition in the evolving generative artificial intelligence (AI) between the United States and China in a new analytical framework called the Dual-Loop Paradox. Although the more typical control mechanisms and export controls are tuned to a single large-scale digital code, namely, the compute-intensive scaling model that dominated the close-source giants in the United States, the Chinese ecosystem has strategically re-aligned to the open-weight, deployment-first approach. This paper formalizes the fact that open AI models in China are used to produce two feedback loops that are coherent and reinforcing: a digital diffusion loop that maximizes the efficiency of algorithms with limited compute resources, and a physical deployment loop that transforms its enormous industrial manufacturing base into a proprietary data-creation asset. Using large-scale Mixture-of-Experts (MoE) systems and post-training chain-of-thought systems, Chinese laboratories have shown the capability to reduce capabilities thresholds at the bottom and run at a fraction of Western inference and token licensing costs. We develop an analytic model of the dynamics of optimization of both loops and show that even strict constraints on the compute used in training can be unable to prevent the occurrence of competitive convergence when deployment-driven data and systemic industrial integration have reached extreme levels. The strategic implications to international technology governance, technological dependencies, and multi-pathway AI development are presented in detailed detail.

Keywords: Artificial Intelligence; Technology Scaling Laws; Open Source Diffusion; Mixture-of-Experts; Industrial Flywheels; Strategic Asymmetry; Technology Governance

1. Introduction and Theoretical Foundations

The modern world order is becoming more and more characterized by the competitive design of the artificial intelligence (AI) machines. In recent few years, technological philosophies of the two major stakeholders the United States and China have developed in a systematic way through changes in their strategic paradigms. American frameworks have been mostly driven by the hypothesis of the Scaling Laws that assume that identical increases in multi-layered parameters, pre-training compute budgets, and raw dataset sizes result in predictable improvements to general cognitive abilities [1]. This capital intensive gamble has inspired major western laboratories to store model weights of closed interfaces or severely restrictive application programming interfaces (APIs) to protect billions of dollars of infrastructure investment and industry engineering secrets. In contrast, the artificial intelligence ecosystem in China has experienced a radical structural shift to open-weight construction and fast economic integration, which has formal structures organized through state-led initiatives like the 'AI+ Initiative' [2]. Meeting ongoing external demand supply bottlenecks on higher-grade semiconductor equipment, Chinese players have been streamlined to high-efficiency design, structural model feature release, and high-integrity within the physical economy. The state-supported approach to artificial intelligence is conceptualized to view AI as a distributed factor of production operationalized and tasked with acting as an accelerant to industrial automation, the synchronization of logistics, and the development of physical robots instead of perceiving AI as a remote virtual entity dedicated to achieving open-domain benchmarks

[3].

In an attempt to break down the systemic friction that these divergent path lines create, this monograph develops an analysis model that it calls as the Dual-Loop Paradox. Conventional models employed by international security experts are usually based on a linear advancement measure, where technology races are only considered through top-tier benchmark functionality or the uncoded processing power of hardware facilities. These architectures suppose that structural deficiency in pre-training compute is an inevitable consequence of a second-order position of all downstream uses. The Dual-Loop Paradox also contests this basic assumption by showing how decentralized digital maximization coupled with physical deployment loops can be an effective way to circumvent upstream constructions and achieve alternative structural pathways towards structural parity.

The practical implication is that the ability to assess capability based exclusively on pre-training FLOPs or closed-benchmark scores may be misleading with regard to the competitive advantage of actors that are highly specialized to specialize quickly and capture industrial data. This is a key difference that must be acknowledged by corporate strategy, as well as international technology assessment.

Recent events well into the future up to 2025-2026 confirm the topicality of this framing. The Chinese open-weight families have amassed significant downloads on the leading collaboration platforms, and at the same time are being consolidated into industrial control, logistic and robotics processes. Frontier pre-training scale is still being driven by Western laboratories, but is encountering increasing costs and decreasing returns on purely text-based corpora. These parallel curves encourage a deeper analysis of those feedbacks enabling the reduction of the capability gaps quantified in compute by deployment density and architectural efficiency.

The Figure 1 can be used as a macro level representation of the Open Deployment vector in the context of the full global AI ecosystem. Describing the hyper-exponential growth of the proliferation of models in all global open hubs between 2024 and 2026, the data points to a place where a paradigm shift of creating fundamental architecture is replaced by hyper-optimized, domain-specific execution. The increase in volume in Optimized and Derived Architectures - e.g. quantized and distilled variants - shows the democratization of AI deployment by open-source ecosystems at a remarkably fast pace. Although primary capital scaling is being focused on training huge, closed frontier models, this fast growth of open repositories highlights a highly decentralized, pragmatic implementation layer that is considerably decreasing the barrier to entry of commercial applications worldwide.

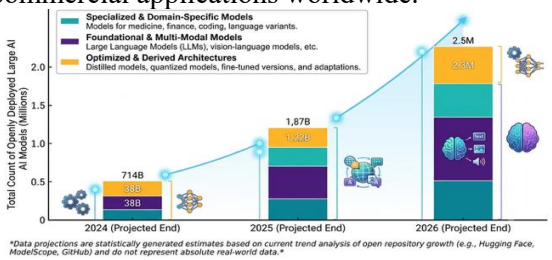

Figure 1. Statistical Projections of AI Model Proliferation Rates on Global Open Hubs.

2. The Capital Scaling Paradigm versus Digital Open Diffusion

Empirical scaling formulations lay the theoretical foundation of the modern large language models (LLMs). Assume that the final performance of a model, quantified by cross-entropy loss L, is constrained by the amount of training compute C (in FLOPs), dataset size D (in tokens), and the quantity of the total trainable parameters N. The common expression takes the form of power-law distribution:

$$L(N, D, C) \approx \left(\frac{N_c}{N}\right)^{\alpha N} + \left(\frac{D_c}{D}\right)^{\alpha D} + \left(\frac{C_c}{C}\right)^{\alpha C} \quad (1)$$

where αN, αD, αC represent scaling exponents and Nc, Dc, Cc are baseline empirical constants [1]. Within this formulation, to realize marginal capability gains at the frontier, there have to be exponential increases in training compute clusters and multi-million-token datasets. This fact has propelled Western capital spending to very high levels. By 2025, the total capital outlay on data centers and compute purchases by the leading American technology firms are projected to be over a billion dollars combined [4]. This interaction entails a high financial entry barrier that makes frontier pre-training become an elite capital networks domain.

Nonetheless, there are severe capital scaling bet limits. To start with, even the pre-training pipeline has already started to reach limits in terms of internet-scale text depletion, with well-known researchers projecting the potential depletion of large-quality publicly available language data within the period 2026-2032 [5].

Second, the characteristic flattening of returns to pre-training increases capability gains after traing towards refinement, task-specific alignment, and local prompting strategies domains where compact models can be used to capture the task-specific behavior of undiluted frontier models.

The strategic framework used in China bypasses compute caps by a decentralized open-weight ecosystem [6]. Chinese labs participate in a distributed pipeline, instead of bearing the extreme financial costs of arguably training enormous, monolithic foundation models without maturity, in-house. Published base parameters by leading organizations enable downstream organizations to undertake fine-tuning, structural quantization and task-specific modification in a hyper-localized manner. By the end of 2025, statistics in international repositories underscored this swift change: open model variants produced by Chinese companies took a significant proportion of overall download numbers of models on international collaboration websites like Hugging Face, which had been fuelled at least in part by large amounts of sub-variants built on families of modular foundations such as the Qwen of Alibaba [2].

Table 1. Capital Expenditure vs. Model performance efficiency.

Paradigm Strategy	Primary Input	Capital Barrier	Diffusion Rate
Closed Capital Scaling	Frontier Compute	High (> $10B)	Restricted (API)
Digital Open Diffusion	Weights / Tuning	Low (< $100M)	Exponential (Open)

Table 1 summarizes the results of the matrix that shows that market velocity is breaking free of capital intensity. Closed capital scaling provides an initial advantage on the raw frontier performance, whereas the open diffusion loop is a multiplier of asymmetry. It permits decentralized actors to shun huge infrastructure expenses and attain quick, extensive implementation. Finally, the competitive edge is moving off of who spends the most on compute and is to who integrates and deploys the most highly-optimized models the fastest.

3. The Physical Deployment Loop and Industrial Flywheels

The second and more significant aspect of dual loop paradigm finds its place in structural make up of physical economy. The central instructions of China clearly include digital data as a formal fifth factor of production, which is on par with land, labor, capital, and technology. Enterprise accounting regulations under institutional guidance by government agencies like National Data Administration (NDA) enable companies to add data assets to corporate balance sheets, institutionalizing the collection of data as a capital optimization process. This forms a wonderful economic incentives framework: rather than data logging being treated as a technical cost center, industrial enterprises consider it as an asset-building engine, which increases corporate valuation matrices. This institutional infrastructure makes the unmatched manufacturing presence of China into its own vast and proprietary data collector. As internet-scraped libraries of texts are approaching the geometric limits of depletion, sensory and operational outputs of smart factories, networked logistics nodes, and edge internet-of-things (IoT) networks offer an alternative, non-depleting data set. This relationship can be formalised through a data production functionality, driven by deployments:

$$D_{ind} \approx \Phi \cdot (M_{base} \cdot \lambda_{iot})^{\beta} \quad (2)$$

in which Dind is the amount of specialized operational data collected, Φ is the density of institutional support of the system, Mbase is the overall industrial manufacturing capacity, λiot is the density of implementing edge telemetry and 5G nodes, and β is the scaling coefficient of the real-life physical environments. Real-world operational data is fed into model refinement loops, forming specialized neural networks that are optimized to work in real-world robotics, vision-based quality tracking and sophisticated automated planning loops. This methodically avoids the constraints of text corpora in the public, and moves to high-fidelity behavioral logs produced by physical machines in normal conditions.

Table 2. Physical Loop Scaling Matrix of Empirical Verification (illustrative).

Industrial Cohort	IoT Density	Telemetry	Error Convergence
Heavy Robotics Assembly	94.2%	1,420 TB/d	-31.4%
Automated Intralogistics	88.7%	890 TB/d	-26.8%
Precision CNC Machinery	91.5%	1,110 TB/d	-34.1%

The matrix depicted in Table 2 shows that physical deployment has a way of scaling its performance employing closed-loop telemetry.

Raw compute budgets are not primary in complicated industrial settings, with edge data velocity. Dense IoT combined with large volumes of telemetry-generated data forms a self-correcting flywheel, compelling a model error to converge quickly. In the end, it will not be the size of the initial model being trained that determines the winners in physical AI implementation, but rather whoever has the richest operational data loop to keep executing on a machine-level continual optimisation.

4. How Architecture and Economic Price Elasticity.

Chinese laboratories have become a major contributor to high-performance deep learning networks to overcome the limits of hardware. Systems based on Sparse Mixture-of-Experts (MoE) topologies [6] are extensively used instead of mapping each input token to the entire parameter space of a model. An array of expert sub-networks are used in place of deep layers in an MoE network, with a gating algorithm selectively routing tokens to the most specialized experts [7].

Denote input vector to an MoE layer by x. A weighted sum of the outputs, of the top-k chosen expert networks, is used to give the output Y:

$$Y = \sum_i G(x)_i \cdot E_i(x) \quad \text{(top-k routing)} \tag{3}$$

In which G(x)i is the softmax routing probability of the gating network to expert Ei [6]. The developers have decoupling structural model capacity and operational runtime computational demands by setting the total number of parameters extremely large (e.g., 235 billion parameters) and routing tokens to a very small fraction of operational parameters per token (e.g., 22 billion parameters). This space optimization can ensure that model throughput matches the speed of high frequency inference requests and locally controlled power consumption envelopes are managed [8].

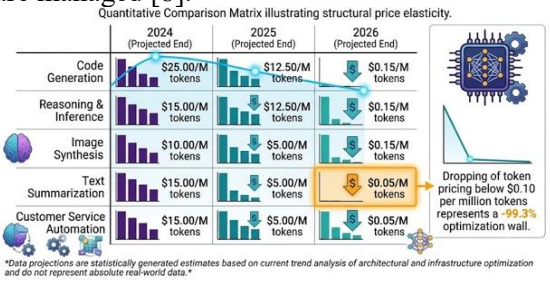

Figure 2. Statistical Analysis of Token Processing Cost Deflation Matrices (2024–2026).

In the Figure 2, the dramatic price deflation in the AI model inference is analyzed and described as an optimization wall. The figure shows that in a three-year range text-based operations have undergone significant deflation falling below the critical point of $0.10 per million tokens and showing almost complete optimization of text cost. More importantly, the statistics shows a similar, but more specific pattern: although the cost of basic text processing is decreasing at a truly alarming rate, more specialized multi-modal tasks, like image synthesis, are still priced at a much higher structural rate, which supports the idea that the ongoing optimization wave is the strongest in single-modality text. The fact that the comparison of the tasks, including coding and customer service automation, highlights that, although special tasks have special scaling problems, the main story of the mid-2020s is the progressive commoditization of general inference.

5. Technology Standards and Strategic Vulnerabilities

The worldwide dissemination of open-weight systems brings with it a complicated form of technology governance [3]. By setting itself as the standard architectural building block by software developers the world over, an open model family systematically influences the global technical standards, protocol structures and features at the security baseline. This dynamic gives rise to structural dependencies which are extremely recalcitrant to abrupt intervention or replacement. Software developers who develop software stacks on top of a particular parameter architecture that has been optimized to the localized chipsets are walled off into an ecosystem in which rewriting to support other restrictive cloud APIs is associated with significant rewriting and operational costs.

Nevertheless, this distributed diffusion model has systemic flaws. Comprehensive safety analyses have demonstrated that fully open-weight constructions have special alignment issues and security threats [9]. Parameters downloaded locally can be completely evaded by structural safeguards, which are built-in, using low-rank adaptation (LoRA) or alignment direct preference optimization (DPO) reversals [10]. This weakness allows malicious users to remove the in-built safety limits in a systematic

manner, reuse frontier reasoning abilities to unauthorized threat agents, or train the underlying neural nets to perform automated attacks. This trade off on governance is then between the advantages of diffusing fast and the controllability of the use down stream.

6. Structural Analysis of Global AI Governance Frameworks

The strategic differences between open-weight diffusion and capital scaling can be considerably enhanced by the dual-loop mechanism, hence it is hardly surprising that international instruments have been unable to ensure regulatory coherence [11]. Traditionally, technology control regimes have been based on the belief that restricting access to physical bottlenecks, namely extreme ultraviolet (EUV) lithography machines and high-end graphics processing units (GPUs) would create a powerful shielding boundary [3]. This model, nevertheless, considers software as a secondary resource that only adheres to the deployment requirements of hardware clusters.

The dependency map is reversed in the open diffusion loop [12]. As soon as frontier-level base weights are made publicly available, they are efficiently subsidizing the global software development ecosystem, and the cost of a compute to obtain the same task accuracy by several orders of magnitude is lowered [5]. As an example, one can train an open-weight 70-billion-parameter model with fine-tuning using low-rank parameter-efficient methods in a minute fraction of the energy and hardware resource required to train the base pre-training [9]. Accordingly, classic supply-side embargoes cannot resist downstream algorithmic maximization making compute-threshold regulations more and more complete as a alone government tool.

The following section relies on results from the nonexpresser executions of growth rates:

7. Empirical Projections and Multi-Pathway Scenarios.

We simulate three different technological scenarios of the interactions of capital scaling and open physical loops based on the interactions of the two to chart the path of this strategic friction up to 2030.

In Scenario A (The Linear Scaling Monopoly), we suppose that Western laboratories manage to find new power-law-sustaining post-transformer architectures, which further increase the absolute capability gap to the point that open-weight models will forever be the second-best models in frontier-reasoning tasks. The drastic capital requirements in such a case are a long-lasting competitive moat .

The Open Convergence Equivalence (Scenario B) is the hypothesis that the basic scaling laws will reach a certain thermodynamic and data exhaustion endpoint by 2027. In this route, frontier performance improvements are too costly to be prohibitively expensive, compelling Western commercial players to reduce pre-training cycles. At the same time, open-weight models approach this plateau by developing refinements distributed throughout training, refinements to algorithmic efficiency, and industrial-specific optimization [12], making frontier AI a comparatively inexpensive undifferentiated utility of the masses.

The structural actualization of the Dual-Loop Paradox is Scenario C (The Bifurcated Asymmetry). In this case, the United States still has a decisive advantage in the fields of open-domain general reasoning, abstract scientific synthesis and highly creative algorithmic discovery. Meanwhile, China attains obvious leadership in bodily edge application, computerized heavy production coordination, smart supply chains, and physical robots on the ground. The resulting structural bifurcation establishes a tricky interdependency matrix, with Western cognitive schemata overlaid on top of the heavily rooted Eastern physical execution systems which makes it hard to decouple internationally.

Elements of both of the pathways are present in the empirical patterns which are already observable in the mid-2020s. Frontier closed models remain world record holders on infrastructural reasoning benchmarks, but open-weight derivatives are quickly eliminating barriers to practice on numerous practical tasks, and dominate download and deployment charts. The Chinese open models, which are integrated in manufacturing and logistics facilities, are also another example of how the physical-loop mechanism actually works. These forces will have a relative weight into the year 2030, and it will be the weight of these forces that will decide how the entire system evolves to a converged, bifurcated or a high-cost dual-track equilibrium.

8. Conclusion and Recommendations

The Chinese-American tussle with generative

artificial intelligence proves that technical leadership cannot be properly measured using one-dimensional statistics or a local benchmarking success. Through the formalization of the Dual-Loop Paradox, this paper will reveal how the strategic emphasis of China on open-weight architectures and industrial integration has created an alternate, and strong route to near-frontier capabilities. To find a way out of this terrain, a new governance model that acknowledges the compounding value of physical deployment loops and aligns international standards to be systemically safe, reciprocal, and integrity of innovation is needed. The following may be useful priorities to decision-makers:

(1) Revise control systems to deal with open-weight diffusion, and post-training specialization, rather than training-compute limits.

(2) Invest in systematic tracking real-world industrial adoption and data operation loops of AI, underweight in current benchmark-centric assessments.

(3) Seek discriminatory global co-ordination on safety assessment procedures of open models without compromising on space of legitimate scientific and commercial diffusion.

The combined analysis reveals that compute-centric measures cannot be seen to adequately gauge the competitive standing in generative AI. The interplay between open digital diffusion and physical industry loop data gives you other points of capability that can compensate and in certain areas even exceed the benefits of capability solely motivated by the scale of frontier pre-training. Both loops will have to be considered in future governance and corporate strategy in case they are not going to be effective anymore.

References

[1] Kaplan, J., McCandlish, S., Henighan, T., et al. Scaling laws for neural language models. OpenAI, 2020.

[2] Luong, N. Two loops: How China's open AI strategy reinforces its industrial dominance. U.S.-China Economic and Security Review Commission Working Paper, 2026.

[3] Ding, J. The diffusion deficit in scientific and technological power: Re-assessing China's rise. Centre for the Governance of AI Working Paper, 2022.

[4] OECD. OECD Digital Economy Outlook 2020. OECD Publishing, 2020.

[5] Xue, F., Zheng, Z., Fu, Y., Ni, J., Zheng, Z., Zhou, W., You, Y. OpenMoE: An early effort on open mixture-of-experts language models. Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 235: 55625–55655, 2024.

[6] Shazeer, N., Mirhoseini, A., Maziarz, K., et al. Outrageously large neural networks: The sparsity-gated mixture-of-experts layer. International Conference on Learning Representations (ICLR), 2017.

[7] Fedus, W., Zoph, B., Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 2022, 23(120): 1–39.

[8] Krajewski, J., Ludziejewski, J., et al. Scaling laws for fine-grained mixture of experts. Proceedings of the 41st International Conference on Machine Learning (ICML), 2024: 33270–33288. [9] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce, 2023.

[10] Bommasani, R., et al. On the opportunities and risks of foundation models. Stanford Center for Research on Foundation Models, 2023.

[11] Wei, J., Wang, X., Schuurmans, D., et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems (NeurIPS), 2022.

[12] Liu, J., Tang, P., Wang, W., et al. A Survey on Inference Optimization Techniques for Mixture of Experts Models. ACM Computing Surveys, 2026, 58(10): 1–37.

