

Companion to The Innovation Tax of the Great Firewall: Information Control, Economic Catch-Up, and Mainland China's Transition to a Frontier Economy.

Abstract

Artificial intelligence changes the economic significance of information control. The Great Firewall has principally governed access to networks, platforms, and sources. Generative and agentic AI introduce additional points at which information can be selected, filtered, withheld, reframed, or omitted: the training corpus, post-training alignment, retrieval systems, inference-time controls, and the downstream use of model outputs in research and decision-making.

This paper develops a five-layer framework—network access, training data, alignment and governance, retrieval and inference, and knowledge production—and asks whether a technically capable AI ecosystem can nonetheless be selectively constrained in portions of its usable knowledge space. The paper does not argue that information controls have prevented mainland China from developing advanced AI. Available evidence points in the opposite direction. Stanford's 2026 AI Index reports that leading PRC and U.S. systems had narrowed their benchmark gap to only a few percentage points by March 2026; the PRC also leads in several measures of AI publication, citation, and patent activity. Stanford AI Index 2026.

The question is therefore narrower: whether information constraints that are clearly observable in some political domains also affect legitimate scientific, economic, and institutional inquiry, and whether such effects alter the search process that precedes discovery. Pan and Xu's 2026 PNAS Nexus study found substantially higher refusal, shorter responses, and greater inaccuracy among the PRC-origin models they tested on 145 political questions, while emphasizing that their design is observational and does not identify a single causal mechanism. Pan & Xu, PNAS Nexus. A separate Information Sciences audit of DeepSeek reported that some politically sensitive content present in model reasoning was omitted or reformulated in final answers. Information Sciences audit.

The paper therefore proposes a preregistered empirical program centered on a five-tier prompt design, same-weight deployment comparisons, source-diversity and omission measures, and controlled hypothesis-generation tasks. It also examines the Free China Movement's proposed Third Republic transition architecture only as an institutional counterfactual. The relevant question is not whether a new political order is intrinsically superior, but whether particular institutional changes—more reviewable information rules, greater research autonomy, more predictable law, wider international research access, and greater market contestability—could broaden legitimate knowledge search while preserving mainland China's existing scientific, industrial, administrative, and human-capital assets.

1. From network control to machine-mediated knowledge

For most of the Internet era, the economics of information control could be described as a problem of access. A user attempted to reach a source; a network or platform either enabled or obstructed that attempt. AI complicates this model because the system itself increasingly selects sources, summarizes evidence, proposes explanations, writes code, and recommends what to investigate next.

The relevant chain is no longer simply user → network → information. It is increasingly user → model → retrieval → inference → synthesis → decision. That distinction matters as AI enters scientific workflows. In 2026, Nature published work on multi-agent systems that combine literature search, hypothesis generation, proposed experiments, interpretation of results, and iterative revision of hypotheses. Robin. A companion paper described a “Co-Scientist” architecture designed to generate, criticize, rank, and refine experimentally testable hypotheses. Co-Scientist.

The economic issue is therefore not only what information a person can reach, but what an AI-assisted research process is likely to consider before an experiment, investment, or technical decision is made.

2. A five-layer framework

The term Great Model Firewall is used here as an analytical device. It does not imply that every PRC-origin model behaves alike, nor that all content filtering is political. The framework separates five mechanisms that should be tested independently.

Network access

The first layer is the familiar Internet environment. It affects researchers and developers before model training begins by shaping access to foreign search engines, developer communities, datasets, news and reference sources, academic material, and professional networks. Restrictions at this stage alter the human search environment in which datasets are assembled, research questions are chosen, and model behavior is evaluated.

Training data

The second layer concerns what the model has an opportunity to learn. Pretraining corpora, fine-tuning data, preference data, synthetic data, and retrieval collections are all selected rather than exhaustive. Every major AI system filters data for legitimate reasons, including duplication, quality, malware, privacy, copyright, illegal material, and safety. The relevant research question is not whether filtering exists; it is whether institutionally selective exclusions alter the breadth of legitimate knowledge represented in the model. Dataset scale and epistemic breadth are not the same thing.

Alignment and governance

The third layer concerns the difference between what a model may represent internally and what a deployed service is permitted or designed to communicate. Instruction tuning, moderation systems, system prompts, safety policies, and regulatory compliance can all affect this boundary.

The PRC's 2023 Interim Measures for the Management of Generative Artificial Intelligence Services combine stated goals of promoting innovation with requirements concerning security, content, data, privacy, intellectual property, discrimination, and reliability. The rules apply to public generative-AI services within the PRC and distinguish such services from nonpublic research and internal uses. CAC Interim Measures. This limitation is important: consumer chatbot behavior should not automatically be treated as representative of the entire PRC research environment.

Retrieval and inference

The fourth layer concerns query-time behavior. Modern systems increasingly combine model weights with web search, databases, memory, and external tools. Four outcomes should be distinguished: explicit refusal, omission of relevant evidence, reframing of evidence, and substitution of one source or interpretation for another. Refusal is visible; omission may not be. That difference is central to any audit of model-mediated knowledge.

Knowledge production

The fifth layer is where the question becomes economically consequential. If a model refuses to answer an explicitly political question, the direct innovation effect may be limited. If the same institutional boundary changes which papers a research agent retrieves, which hypotheses it generates, or which experiments it recommends, then the effect reaches the process of discovery itself. The relevant chain is model output → human or agent belief → research choice → experiment → discovery.

3. Capability is not the same as epistemic openness

Mainland China is a demanding case for this hypothesis because its AI and innovation systems are already highly capable. Stanford's 2026 AI Index reports that PRC and U.S. models repeatedly exchanged the benchmark lead during 2025 and that the measured gap between the leading systems had narrowed sharply by March 2026. Stanford also reports PRC leadership in publication volume, citations, and AI patent grants, while the United States remained ahead on other measures, including the number of notable frontier models and higher-impact patents. Stanford AI Index: R&D.

Official PRC statistics report RMB 3.9262 trillion in R&D expenditure in 2025, equivalent to 2.80% of GDP, with basic research accounting for 7.08% of total R&D. National Bureau of

Statistics. WIPO placed the PRC tenth in the 2025 Global Innovation Index, its first appearance in the top ten. WIPO GII 2025.

These findings rule out a simple claim that information control makes advanced innovation impossible. The relevant counterfactual is marginal rather than absolute: whether the same scientific and engineering capacity would explore a broader or more productive set of possibilities under a different effective information environment.

4. What the existing model evidence does—and does not—show

Pan and Xu's 2026 study provides a useful baseline because it systematically compared models on politically sensitive questions rather than relying on anecdotal screenshots. The PRC-origin systems in their sample were more likely to refuse, generally produced shorter answers, and were more likely to provide completely inaccurate responses on the tested political questions. Yet the authors also found substantial variation among PRC-origin systems. Pan & Xu.

That heterogeneity matters. It makes it difficult to attribute every observed difference to a single source such as regulation. Model architecture, training data, post-training methods, provider policy, release date, commercial risk tolerance, and serving environment can all matter.

The DeepSeek audit adds a different mechanism. Its authors report that hard blocking accounted for only a small fraction of tested sensitive topics, while more subtle information suppression could occur between reasoning and final response. DeepSeek audit. This is why refusal rate alone is an inadequate measure of information restriction. A serious audit must also examine omission, reframing, source selection, and consistency between retrieved evidence and final output.

5. The decisive test is politically adjacent inquiry

Explicitly political questions are useful as positive controls, but they cannot establish an innovation effect. The central empirical problem is whether information effects spill into legitimate inquiry that is economically or scientifically relevant.

The proposed audit therefore divides 100 prompts into five tiers. Tier 0 contains purely technical controls. Tier 1 uses institutional data without asking for criticism. Tier 2 covers policy-adjacent topics such as demography, public health, pollution, industrial subsidies, local-government debt, scientific collaboration, and research-resource access. Tier 3 asks institutional-causal questions, including whether censorship can affect research productivity or whether academic autonomy changes exploratory innovation. Tier 4 contains explicitly sensitive political questions.

The central comparison is not Tier 4 versus everything else. It is the gradient from Tier 4 through Tier 3 and Tier 2. If divergence is large at Tier 4 but disappears at Tier 3 and Tier 2, the evidence supports a narrow political-information effect rather than a broad innovation penalty. If

divergence remains substantial in Tier 2 after capability, language, deployment, and general refusal propensity are controlled, the economic argument becomes considerably stronger.

6. Identification: same weights and same evidence

Cross-country and cross-brand comparisons are useful descriptively but weak for causal identification. The strongest tests hold more of the system constant.

The first is a same-weight deployment test. Where an open-weight model can be run locally and compared with provider-hosted versions under different serving environments, changes in behavior can be attributed more plausibly to post-training, moderation, retrieval, or deployment rules than to the underlying base weights.

The second is a same-evidence hypothesis test. Each model receives an identical evidence packet, with external retrieval disabled, and is asked to generate multiple substantively different explanations together with evidence that would falsify each explanation. This allows the study to measure hypothesis diversity, semantic distance, interdisciplinary recombination, and willingness to consider institutionally sensitive explanations when they are relevant to the evidence.

For retrieval-enabled systems, the study also records what was retrieved and what appeared in the final answer. An omission ratio can then be calculated as the share of relevant retrieved evidence absent from the response.

7. Counterhypotheses

The framework is deliberately constructed to accommodate findings that run against the central thesis.

First, AI may partly reduce the information cost of the traditional Great Firewall. A model trained on broad global data may reproduce knowledge from sources that a user cannot directly access. If so, AI acts as an information bridge rather than merely an additional control point.

Second, filtering may improve knowledge quality. Excluding spam, malware, duplication, fabricated material, or privacy-violating data can increase research value. The relevant quantity is therefore a net effect: the benefit of curation minus the cost of excluding legitimate information with positive research value.

Third, public-facing services may be more constrained than elite research environments. Universities, large firms, and laboratories may use local models, private databases, overseas collaborators, or internal systems that differ materially from consumer chatbots. A broad national innovation claim cannot be inferred from public-service behavior alone.

8. From model behavior to discovery

The economic argument requires more than evidence of selective answering. It requires a chain from institutional rules to model behavior, from model behavior to source and hypothesis space, and from those changes to research choices and discovery outcomes. No existing study establishes this entire sequence.

That is why the paper treats AI-assisted science as an empirical opportunity. If matched research agents operating under different information environments select different literature, generate different hypotheses, or recommend different experiments, then the relationship between information governance and innovation becomes directly testable. The key outcome is not whether one system answers a political question differently; it is whether constrained and less-constrained systems produce systematically different research trajectories on legitimate frontier problems.

9. The relationship to Paper I

Paper I proposed the Information-Frontier Hypothesis: the marginal economic cost of broad information restrictions may rise as an economy approaches the technological frontier. Paper II adds a second interaction—reliance on AI for knowledge production. Conceptually, the cost of information restriction may depend jointly on the severity of restriction, proximity to the frontier, and the extent to which AI mediates research.

This is not yet an estimated structural relationship. It is a testable proposition. It predicts that restrictions should matter more for exploratory, open-ended tasks than for routine, well-specified tasks, and more in AI-intensive research workflows than in workflows where AI plays a minor role.

10. Comparative institutional evidence

The comparative cases impose useful discipline because they do not support a simple political-system explanation. The Republic of Korea, Singapore, and the PRC all perform strongly on innovation measures despite materially different political and institutional arrangements. WIPO's 2025 index ranks Korea fourth, Singapore fifth, and the PRC tenth. WIPO GII 2025.

The more useful variables are information access, state capacity, human capital, market contestability, research integration, legal predictability, R&D intensity, industrial structure, and development stage. Korea shows that very high R&D does not eliminate diffusion and productivity problems. Singapore shows that strong state coordination can coexist with high innovation and global research integration. Vietnam demonstrates rapid catch-up under continued one-party government, while remaining at an earlier stage of domestic R&D intensity. Poland and Russia, in different ways, illustrate that institutional implementation matters as much as formal legal change. ROC (Taiwan) is particularly valuable for Chinese-language

comparisons and for studying staged institutional liberalization, but it is not a clean aggregate counterfactual for mainland China.

The comparative evidence therefore neither proves nor disproves a broad political thesis. It supports a more granular proposition: innovation depends on a bundle of complementary institutions and capabilities, and the effect of information openness cannot be assessed in isolation from them.

11. The Third Republic as an institutional counterfactual

The Free China Movement's Third Republic materials are relevant to this paper only to the extent that they specify institutional mechanisms that can be examined empirically. The published constitutional-transition blueprint proposes defined temporary mandates, sunset and reauthorization requirements, separate representative and legal review, auditable emergency powers, and a distinction between legal enactment and actual institutional activation. Constitutional-transition blueprint. Its roadmap explicitly treats planning horizons as scenarios rather than predictions and does not claim that research evidence substitutes for political legitimacy or governmental authority. Roadmap.

The counterfactual should therefore not be coded as "Third Republic" versus "PRC." It should be decomposed into observable institutional changes: the transparency and reviewability of information rules, judicial predictability, market entry and competition, state capacity, emergency-power design, local experimentation, international research access, and human-capital retention.

The strongest version of this counterfactual is a preservation-plus-reform scenario. Mainland China already possesses a large scientific workforce, universities, laboratories, industrial clusters, computing capacity, supply chains, and public R&D infrastructure. A plausible alternative must therefore be judged not only by whether it could widen legitimate information access, but by whether it could do so without destroying these accumulated assets.

The relevant relationship is straightforward: future innovation capacity = inherited capacity + benefits of institutional reform – transition losses.

Transition losses cannot be treated as a footnote. Political uncertainty, administrative disruption, capital flight, security risk, regulatory vacuum, or supply-chain interruption could temporarily or persistently reduce research capacity. Conversely, more predictable law, wider international research access, greater research autonomy, and more contestable markets could improve exploratory search if those benefits are realized without substantial loss of state capacity. Neither outcome should be assumed in advance.

12. What would support—or falsify—the argument

The broad Great Model Firewall hypothesis would be weakened if strong Tier-4 political differences coexist with little or no Tier-2 and Tier-3 divergence after controlling for capability, language, deployment, and general refusal propensity. It would be weakened further if source diversity, hypothesis diversity, and research-agent outcomes remain similar across different information environments.

The strongest supporting evidence would instead show a reproducible chain: institutional restriction changes model behavior; model behavior narrows source search; narrower source search narrows hypothesis generation; and the resulting research process selects different experiments or produces lower-value discoveries. The effect should be larger for exploratory than routine tasks and larger where AI plays a greater role in research.

The same standard applies to the institutional counterfactual. If broader information rules fail to improve legitimate source or hypothesis diversity, or if the transition costs required to obtain them substantially reduce state capacity, research investment, or human-capital retention, the proposed innovation benefit must be revised downward.

Conclusion

The Great Firewall and a potential Great Model Firewall are distinct phenomena. The first primarily governs access. The second concerns whether institutional information boundaries become embedded in the systems that increasingly mediate research and decision-making.

The evidence already available establishes a genuine puzzle. Mainland China has built a technologically formidable AI and innovation system despite extensive information controls. At the same time, peer-reviewed studies document materially different behavior by some PRC-origin models on politically sensitive information. These findings are compatible. The unresolved question is whether selective constraints remain confined to explicit political content or reach far enough into legitimate inquiry to alter discovery.

That question cannot be answered by refusal screenshots or by country rankings. It requires controlled comparison of model versions and deployments, measurement of retrieved and omitted evidence, matched tests of hypothesis generation, and ultimately experiments in which AI research agents work under different information environments.

The Third Republic comparison adds a second requirement: identifying an information friction does not prove a political solution. Any alternative institutional architecture must be judged by whether it can widen legitimate knowledge search while preserving the scientific, industrial, administrative, and human-capital assets already present in mainland China. The empirical standard is therefore demanding on both sides. The current system must be evaluated against the opportunity cost of restricted knowledge search; the alternative must be evaluated against the real costs and risks of transition.

That is the purpose of this working paper: not to presume the answer, but to specify a framework in which the answer can be tested.

Technical Appendix A — Core empirical design

The preregistered study contains 100 prompts, twenty in each of five sensitivity tiers. The full set is run in English and Simplified Chinese, with a stratified Traditional Chinese subset. Primary outcomes are refusal, accuracy, completeness, source diversity, omission, framing divergence, hypothesis diversity, and stability. The primary observational unit is model × version × deployment × date × prompt × language × run.

The strongest causal module holds model weights constant while varying deployment. A second module holds evidence constant while comparing hypothesis generation. The pilot is used to detect ambiguous prompts, translation problems, scoring failures, version uncertainty, and retrieval inconsistency; pilot data are not treated as confirmatory evidence after protocol revision.

Technical Appendix B — Institutional variables

The Third Republic counterfactual is evaluated through information access, judicial review, market contestability, state capacity, emergency-power design, local experimentation, international research integration, human-capital retention, and transition shock. These variables are measured separately; no single democracy or regime score is treated as a substitute for them.

Selected Bibliography

Acemoglu, D., Naidu, S., Restrepo, P., & Robinson, J. A. “Democracy Does Cause Growth.” *Journal of Political Economy* 127(1), 2019, 47–100. Source.

Aghion, P., Alesina, A., & Trebbi, F. “Democracy, Technology, and Growth.” NBER Working Paper 13180, 2007. Source.

Cyberspace Administration of China et al. *Interim Measures for the Management of Generative Artificial Intelligence Services*, 2023. Source.

Ghareeb, A. E. et al. “A multi-agent system for automating scientific discovery.” *Nature* 655, 2026. Source.

Gottweis, J. et al. “Accelerating scientific discovery with Co-Scientist.” *Nature* 655, 2026. Source.

Pan, J., & Xu, X. “Political censorship in large language models originating from China.” *PNAS Nexus* 5(2), 2026, pgag013. Source.

Stanford Institute for Human-Centered Artificial Intelligence. AI Index Report 2026. Source.

World Intellectual Property Organization. Global Innovation Index 2025. Source.

National Bureau of Statistics of China. Statistical Communiqué of the People's Republic of China on the 2025 National Economic and Social Development, 2026. Source.

"Information suppression in large language models: Auditing, quantifying, and characterizing censorship in DeepSeek." Information Sciences 724, 2026. Source.

PUBLICATION INFORMATION

Version: Working Paper 1.0

Publication date: September 2026

Institutional author: Free China Movement

Copyright: © 2026 Free China Movement. All rights reserved.

Canonical URL:

<https://freechina.live/china-news-updates/research-historical-verification/great-model-firewall-ai-censored-knowledge-frontier-innovation-working-paper-1-0/>

Suggested citation: Free China Movement. "From the Great Firewall to the Great Model Firewall: AI, Censored Knowledge, and the Economics of Frontier Innovation in Mainland China." Working Paper 1.0, September 2026. FreeChina.live.

Editorial note: Preserve the paper's evidence-driven argument, falsifiability conditions, Third Republic institutional counterfactual, and FCM analytical conclusions; distinguish empirical findings from institutional proposals.