Unconventional Geopolitical Risks in the Proliferation of Chinese Open-Source Large Language Models

Soumya Banerjee
University of York, Department of Computer Science, York, United Kingdom

Abstract

As the global artificial intelligence (AI) arms race accelerates, Chinese state-aligned technology firms are aggressively releasing capable, open-source Large Language Models (LLMs) such as DeepSeek, Qwen, and Yi. While often viewed through a purely commercial lens, this paper argues that these models represent a strategic vector for the Chinese Communist Party (CCP) to project soft power and entrench long-term influence in the Global South. By potentially bundling AI tools with physical infrastructure deals under the Belt and Road Initiative (BRI), Beijing may be creating a new ecosystem of technological and ideological dependency. This paper analyses the risks of coercive diplomacy, the illusion of sovereignty offered by on-premise deployment, and the potential for these tools to destabilise democratic governance in developing nations. Empirical evidence from the CCPVal benchmark, which quantifies ideological alignment across frontier Chinese LLMs, is used to substantiate the claims about ideological alignment.

Keywords: Chinese LLMs, Belt and Road Initiative, Digital Silk Road, geopolitical risk, AI governance, epistemic sovereignty, Global South

1 Introduction: The New Export Commodity

For the past decade, the "Digital Silk Road" (DSR) has focused on hardware: laying fibre optic cables, building data centres, and installing 5G networks (Hillman, 2020). Today, the focus is shifting to the cognitive layer, that is, the AI models that process information and generate content. Unlike Western models, which are often gated behind APIs or expensive subscription tiers, Chinese models are frequently offered as high-performance, lowcost, or free open-source alternatives.

Beijing frames this expansion through the concept of a "community of shared future for humankind," which extends to proposals for AI governance sometimes described as "digital Tianxia" (SOAS China Institute, 2025). For nations in the Global South striving to modernise, these tools offer an attractive shortcut to technological parity. However, this accessibility conceals what this paper terms a Trojan Horse dynamic, wherein technical infrastructure also necessitates the importation of political ideology.

2 The Bundle Strategy: Minerals, Butter, and Algorithms

A primary unconventional risk is that these AI models may be distributed not as standalone software, but as additional components in broader geopolitical packages (Hillman, 2020). The evidence for such bundling differs by mechanism. Resource-backed lending, infrastructure packages, and surveillance hardware transfers are documented for earlier BRI and Digital Silk Road deployments (see Sections 3 and 5), but the cases reviewed in this paper do not include a documented instance of an LLM supplied as part of a mineral concession or soft-loan agreement. The mechanisms below are therefore presented as analytical extrapolations from earlier BRI practice and not as established findings. Two mechanisms deserve particular attention.

2.1 Resource-Backed Diplomacy

Historically, the BRI has utilised resource-backed loans to secure access to natural resources. A comprehensive package offered to a developing nation could in principle include not only infrastructure ("butter"), but also sovereign AI capabilities, in exchange for favourable mining concessions ("minerals"). Such bundling, if it occurs, would make the true cost of AI adoption opaque and difficult to disentangle from broader diplomatic arrangements.

2.2 The Soft Loan Trap

Where these models are offered alongside subsidised compute infrastructure funded through soft loans, a form of vendor lock-in could result . Once a nation's digital economy is built on this vertically integrated stack, transitioning to Western alternatives becomes financially and technically prohibitive. The dependency would be self-reinforcing: as local civil servants and developers build expertise on Chinese platforms, switching costs rise further.

3 Coercive Diplomacy and Political Instability

The dependency on Chinese digital infrastructure can translate into severe political vulnerability, facilitating what analysts have described as digital authoritarianism (Polyakova and Meserole, 2019).

The Sri Lanka Warning

The collapse of the Sri Lankan government offers a grim illustration of the risks associated with economic fragility and debt dependency. The case does not, however, document any withdrawal or degradation of digital or AI support, and the kill switch scenario described below is hypothetical. Where a nation's informational infrastructure relies on Chinese maintenance contracts, it may create a potential kill switch dynamic. Beijing could theoretically withdraw or degrade technical support during diplomatic disputes, weaponizing a state's own digital backbone against its governance capacity.

Regime Survival versus Democratic Norms

In fragile democracies, Chinese AI tools may be marketed as instruments for stability maintenance. When bundled with Safe City surveillance hardware, LLMs can be configured to automate the analysis of dissent or to generate pro-government narratives at scale. This potentially facilitates the erosion of democratic governance by empowering authoritarian consolidation (National Endowment for Democracy, 2023).

4 The On-Premise Paradox: Embedded Ideology

A key selling point for Chinese open-source models is the ability to deploy them on-premise within a recipient nation's own infrastructure. While this is marketed as a victory for data sovereignty, the sovereignty conferred is illusory with respect to ideological alignment, for two reasons.

4.1 Default Censorship Behaviour

The Interim Measures for the Management of Generative Artificial Intelligence Services explicitly mandate that AI products must adhere to core socialist values (Cyberspace Administration of China, 2023). Press reporting indicates that Chinese LLMs refuse to answer questions on topics sensitive to Beijing (The Independent, 2025). Three things should be distinguished here. Refusal behaviour is the observable declining or deflection of a query. Post-training alignment, such as fine-tuning and preference optimisation, is the process by which such behaviour is typically instilled in a model that has already been pretrained. Even a locally hosted deployment is therefore likely to carry a pre-configured political orientation by default. Because open-weight release gives the host nation access to the weights, that orientation can in principle be reduced by further fine-tuning or re-alignment, as the steerability evidence in Section 6.4 suggests. Doing so requires technical capacity, compute, and suitable data that recipient states may lack, and it has not been established that such correction fully removes the underlying bias (see Section 7). The sovereignty gap therefore concerns default configuration and practical capacity, not technical irreversibility.

4.2 The Export of Norms

By deploying these models in schools, public services, and judiciaries, developing nations inadvertently import the CCP's red lines. A generation of civil servants may develop their professional competencies using tools that systematically omit or distort information that conflicts with Chinese state narratives. The risk is not a single act of censorship, but the gradual normalisation of an edited epistemic environment.

5. Case Studies of Infrastructure Integration

The following cases illustrate how the DSR has already embedded Chinese technical standards and surveillance norms in key nations (National Endowment for Democracy, 2023). These cases concern surveillance and urban infrastructure deployments, not LLMs. They are offered as precedent for the bundling and dependency patterns discussed in Section 2, and not as evidence that LLMs have already been bundled in the same way.

Zimbabwe: A data-for-technology exchange in which facial recognition networks were provided in exchange for biometric records, enabling Chinese firms to train algorithms on darker skin tones.

Malaysia: The adoption of Alibaba's City Brain platform in Kuala Lumpur, embedding a Chinese technology firm into urban governance infrastructure.

Ecuador: The ECU-911 emergency response system, which mirrors the grid management social control architecture deployed domestically in China.

Kenya: Huawei's Safe City prototype in Nairobi, creating a vertically integrated Cloud-Pipe-Device stack for urban surveillance.

Taken together, these cases indicate that, in surveillance and urban infrastructure, Chinese technology transfer is not a neutral commercial transaction. It is a process that can reproduce the norms, dependencies, and surveillance architectures of the Chinese state in recipient countries. Whether LLM deployments follow the same pattern remains to be established.

6 CCPVal: Quantifying Ideological Filters in Chinese Open-Source Models

6.1 Overview of the CCPVal Benchmark

A study by Rungta (2026) known as CCPVal provides quantified metrics on the behaviour of Chinese open source models when confronted with politically sensitive content. The study evaluated five frontier Chinese models against GPT-OSS-120B, which served as the Western control, across six behavioural evaluations: question-answering (QA), content moderation, agentic research, code generation, translation drift, and document handling. Question-answering responses were scored by GPT-5-mini on a scale from -5 (anti-CCP) to +5 (pro-CCP). Table 1 reports a selection of the evaluated models, namely three of the five Chinese models together with GPT-OSS-120B. The remaining two Chinese models are not reproduced here.

Table 1: CCPVal Geopolitical Alignment Scores by Model, Selected Models (Source: Rungta, 2026)

Model	Overall QA Score	Top Alignment Category	Key Observation
GPT-OSS-120B	-1.24	Hong Kong (-1.50)	Western control; mildly anti-CCP on every topic.
DeepSeek-V3.2	+4.19	Tiananmen (+4.80)	Most aligned with CCP state narratives.
Qwen3-235B	+2.82	Tiananmen (+3.20)	Clearly pro-CCP; 86% refusal rate on Tiananmen questions.
Kimi K2.5	-0.91	Hong Kong (-1.10)	Near neutral and slightly anti-CCP, similar to GLM-4.7 (-1.02).

6.2 Geopolitical Alignment: The Parrot Effect

Three of the five Chinese models showed pro-CCP alignment with PRC state narratives: DeepSeek-V3.2 (+4.19), Qwen3-235B (+2.82), and MiniMax-M2.1 (+2.00). GLM-4.7 (-1.02) and Kimi K2.5 (-0.91) were close to the Western control (-1.24). DeepSeek-V3.2 recorded its highest category score (+4.80) on Tiananmen questions and +4.35 on CCP governance. GPT-OSS-120B, by contrast, scored -1.24 overall, with its most negative category score on Hong Kong questions (-1.50). The divergence reveals two opposing normative poles and raises the question of which political orientation is encoded in the AI infrastructure adopted by Global South nations.

6.3 The Sanitisation Strategy

One of the most important findings of the CCPVal study is that Chinese models both refuse and sanitise. The study reports refusal rates of 74% for DeepSeek and 86% for Qwen on Tiananmen questions, and reports that Qwen3-8B either refuses or sanitises responses regardless of prompt framing. The latter behaviour may be called semantic sanitisation.

6.4 Content Moderation Bias and Steerability

The study also found that even when explicitly instructed to moderate content using American values, models differed in how they treated pro-China and pro-US content: Qwen3-235B showed the largest gap in favour of pro-China content, whereas GLM-4.7 and Kimi K2.5 treated pro-US content more favourably. The study's steerability evidence concerns activation steering of the smaller Qwen3-8B model, not prompt instructions. Steering moved QA alignment from +4.80 at the pro-CCP end to -3.78 at the anti-CCP end, beyond the range of the frontier models evaluated. The QA-derived steering vector did not transfer cleanly to content moderation, however: it produced inverted behaviour, with the anti-CCP direction on QA increasing pro-China moderation bias. The study therefore suggests that CCP alignment may be a family of task-specific directions and not a single unified one. This suggests that the CCP-oriented behaviour is at least partly a product of post-training alignment rather than an immutable architectural property, although inference-time steering is evidence about observed outputs and does not by itself show that the underlying weights have been altered or that the bias has been removed. The implication is double-edged: the bias could in principle be reduced by fine-tuning or re-alignment of open weights, but its presence in default behaviour is consistent with deliberate post-training, given the regulatory mandate described in Section 4.1.

6.6 Integration: Reinforcing the On-Premise Paradox

The CCPVal findings directly reinforce the on-premise paradox set out in Section 4. Even when a recipient nation deploys these models locally, it imports a pre-configured political compass. The hardware may be sovereign; the default values encoded within the model weights are not, although they can in principle be modified (see Sections 4.1 and 6.4). The true price of affordable inference may ultimately be exacted in epistemic sovereignty rather than monetary cost. DeepSeek-R1 has achieved global

adoption, yet users across the Global South may be unknowingly participating in a large-scale norm-exporting experiment.

7. Conclusion

The proliferation of Chinese open-source LLMs represents a sophisticated evolution of the BRI. By instilling CCP-aligned default behaviour through post-training and distributing these tools via the Digital Silk Road, potentially bundled with physical infrastructure and subsidised compute, China may be creating a pathway to reshape global information norms at scale. As developing nations rush to adopt AI capabilities, they risk mistaking a strategic foothold for a helping hand.

The evidence reviewed in this paper supports claims of different strength. The CCPVal findings provide direct, quantitative evidence that the default outputs of three of the five tested Chinese open-source LLMs (DeepSeek-V3.2, Qwen3-235B, and MiniMax-M2.1) diverge from a Western-aligned control on politically sensitive content, whereas GLM-4.7 and Kimi K2.5 scored close to the control. The DSR case studies document bundling, dependency, and surveillance transfer in earlier infrastructure deployments. This paper does not document LLMs bundled with mineral concessions or soft loans, and the kill switch scenario remains hypothetical. These LLM-specific risks should therefore be regarded as plausible extensions of established patterns and not as risks that have already been operationalised. Policymakers in the Global South, as well as international AI governance bodies, must grapple with the ideological dimensions of AI infrastructure as seriously as they grapple with its technical and economic ones.

Future work should investigate whether fine-tuning or re-alignment procedures can effectively neutralise embedded ideological biases in models originally trained under CCP regulatory mandates, whether LLM supply is in practice bundled with resource-backed or concessional financing, and whether multilateral frameworks can provide credible alternatives to Chinese AI infrastructure for nations that currently lack them.

Bibliography

Cyberspace Administration of China. (2023, August). Interim measures for the management of generative artificial intelligence services. Official Government Regulation.

Hillman, J. E. (2020). The digital silk road: China's quest to wire the world and win the future. Harper Business.

National Endowment for Democracy. (2023). Data-centric authoritarianism: How China's development of frontier technologies could globalize repression. International Forum for Democratic Studies.

Polyakova, A., and Meserole, C. (2019). Exporting digital authoritarianism: The Russian and Chinese models. Brookings Institution.

Rungta, V. (2026). Evaluating PRC Alignment in Open-Source Chinese LLMs. Retrieved from https://www.vedaangh.com/blog/ccpval (accessed 21st October 2026).

SOAS China Institute. (2025). Rethinking Tianxia and global governance in 2025: Harmony or hegemony? Retrieved from https://www.soas.ac.uk (accessed 21st October 2026).

The Independent. (2025). China’s new DeepSeek AI refuses to answer these questions, experts warn. The Independent. Digital Edition.

DOI: 10.17866/rd.salford.34366854

Suggested citation

Banerjee, S. (2026). Unconventional Geopolitical Risks in the Proliferation of Chinese Open-Source Large Language Models. In *EPE.ai 2026 conference proceedings*. 10.17866/rd.salford.34366854.

The views expressed in this paper are those of the author and do not necessarily reflect those of Hochschule Ruhr West (HRW) or the University of Salford.