Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36866

A Review of Safety-Ethical Issues and Governance Approaches for Large Language Models

Ziming Xie

School of Economics and Management, Xidian University, Xi'an, China
19930661213@163.com

Abstract. The large pre-trained models centered on the Transformer architecture have been widely applied in fields such as government affairs, healthcare, and cultural creation, significantly enhancing the efficiency of content production. However, they have also given rise to multiple security and ethical risks: Data leakage can lead to the leakage of users' medical records privacy, attackers can steal commercial large models through API calls, multi-modal deep fakes are used in telecommunications fraud, and algorithmic biases in similar Amazon recruitment large models can cause employment discrimination. This paper takes generative large models as the research object and adopts methods such as literature review, case analysis, and comparative research to classify the risks of large models into two categories: information security and ethical security. It summarizes the protection technologies throughout the entire life cycle of training, reasoning, and traceability, compares the differences in AI regulatory systems between China and abroad, and constructs a complete governance path from the technical, institutional, and industrial dimensions. This paper can provide a basic theoretical reference for the safe implementation of large model enterprises and the improvement of industry norms by regulatory authorities.

Keywords: large-language model, artificial intelligence security, algorithmic ethics, AI governance, trustworthy artificial intelligence

1. Introduction

With the rapid iteration of the Transformer architecture and large-scale pre-training technologies, models such as GPT, Tongyi Qianwan, and DeepSeek have been widely applied in scenarios like text generation, intelligent healthcare, content creation, and government services, deeply integrating into all aspects of social production and life. The large models have achieved innovation in productivity through their powerful generation capabilities, but the uncontrollable generation of the models themselves, the inherent biases of the training data, data privacy leakage, and security risks such as deep fakes have also emerged. Academic and industrial circles in various countries have gradually realized that large models cannot merely pursue the improvement of performance indicators without also pursuing safety and ethical constraints. Research on the safety and ethics of large models has become a core research branch in the AI field [1].

DOI: 10.54254/2755-2721/2026.36866

2. Classification and causes of safety and ethical risks of large models

2.1. Information security risks and formation mechanism of large models

Information security risks are the technical risks that grow out of a model's architecture and its training-and-interaction mechanisms. They directly threaten data privacy, corporate model assets, and network information security, and they mainly come in four forms: data leakage, model theft, adversarial attacks, and deepfakes. First, data leakage. Large models have an extremely strong memory for data. Training sets get mixed with phone numbers, medical records, business contracts, ID numbers, and other sensitive information. Attackers can use member-inference attacks, long repeated prompts, gradient inversion, and similar tricks to coax the model into reproducing private, sensitive data from the training set. The mechanism behind it: the pre-training stage lacks a unified, standardized data-masking process, the fitting of model parameters locks in details from the samples, and the output-filtering mechanism in the inference stage has holes in it [2].

Second, model theft. Building a commercial large model costs enormous computing power and data, so the model itself is a company's core digital asset. Attackers can do it on the cheap — just call the public API, feed in batches of constructed samples, and reverse-engineer the model's weights from the output distribution. Data poisoning is another route: they plant hidden backdoor samples in the training data. Under normal input the model behaves fine, but when a specific trigger prompt appears, it generates illegal or wrong content. The root cause is that open APIs lack strict access checks, which goes hand in hand with an incomplete review-and-admission process for training datasets [3].

Third, adversarial attacks. Attackers craft adversarial prompts through nested layers, split sentences, role-playing, and other tricks to get around the model's safety-alignment constraints and steer it into generating fraud scripts, violent tutorials, or instructions for making illegal tools. This happens because the model's value alignment has blind spots at its edges — it can't recognize malicious instructions that have been deformed or disguised [4].

Fourth, deepfakes. Text, image, audio, and video models can generate fake faces, fake voices, forged documents, and tampered news footage at the push of a button, and they're heavily used for telecom fraud, online rumor-mongering, and reputation damage [5]. The risk comes from the fact that multimodal generation modules lack a step for verifying authenticity, so AI-generated content ends up with no unified traceability markers [6].

2.2. Ethical-security risks and formation mechanisms of large models

Ethical safety risks grow out of data bias, missing values, and improper commercial use. They undermine social fairness, infringe on intellectual property, and blur the lines of responsibility. They mainly come in four forms: algorithmic bias and discrimination, harmful content generation, copyright infringement, and unclear responsibility. First, algorithmic bias and discrimination. Training data is drawn from internet text accumulated over a long time, so it naturally carries stereotypes about gender, region, occupation, and race. In recruitment, for example, a large model might push male applicants to the front, give negative reviews to niche groups, or water down treatment advice for minority patients in medical consultations. The root cause is that historical social bias gets copied into the dataset, and the model's loss function only optimizes for text fluency — it never introduces fairness constraints as a metric [7, 8]. Second, harmful content generation. Users can use large models to mass-produce content that stokes opposition, spreads rumors, or nudges people toward self-harm and suicide. Once that spreads across social platforms, it disrupts

DOI: 10.54254/2755-2721/2026.36866

public order. Large models do not have human value judgment. They generate purely by following text probabilities, so they cannot independently tell whether content is socially harmful [9]. Third, copyright infringement. During pre-training, models scrape novels, academic papers, paintings, music, and other original works without authorization. The output then directly copies the original plots, art styles, and wording, which badly hurts creators' rights. Current rules still have not clearly defined how far AI pre-training can go in using data, so models have no way to automatically tell public material apart from copyrighted work [10]. Fourth, unclear responsibility. When deepfakes damage someone's reputation, or fake AI advice causes financial or physical harm, it it hard to pin responsibility on the user, the model developer, or the platform operator. And in high-risk professional settings like law or medicine, when a wrong AI output leads to losses, there is no standardized process for holding anyone accountable [11].

2.3. Common underlying causes of risk generation

First, there is a flaw at the source — the data layer. Training data comes from scattered sources and is huge in volume, and there is no unified standard for cleaning it or checking it for fairness. Raw internet data is mixed in with private information, biased language, and infringing material, which makes it the root of all these risks. Second, there are built-in weaknesses at the algorithm layer. Large models are black boxes, so the reasoning behind their outputs canont be explained, and it is hard to pin down exactly where a risk gets triggered. Pre-training only chases fitting accuracy, but safety, fairness, and compliance are never built into the basic optimization goals. The models' strong memory of samples makes data leakage and backdoor attacks worse. Finally, at the application layer, there's a lack of control. The bar for commercially deploying large models keeps dropping, so a lot of small and mid-sized companies cannot afford to set up proper safety-review systems.

3. Core technology system and effectiveness analysis of current large-model security protection

3.1. Preemptive prevention protection technologies

The main pre-emptive protection techniques include differential privacy combined with fairness-based data cleaning, poisoned-sample anomaly detection, RLHF human-feedback alignment, and federated learning for distributed training. First, differential privacy plus fairness-based data cleaning. This uses anonymization and perturbation algorithms to strip sensitive information out of the text and filter out training samples that carry discrimination or violence, which eases privacy leaks and algorithmic bias. The upside is that it cuts down the risk base at the source. The downside is that cleaning such a huge volume of data eats up enormous computing power, and differential privacy slightly lowers how fluent the model's output is [12]. Second, poisoned-sample anomaly detection. Relying on clustering algorithms and feature recognition, it screens the training set for backdoor and malicious samples to block the poisoning attack path. The strength is that it catches hidden backdoors early. The weakness is that its accuracy drops when it comes to new, well-hidden poisoning samples, so full coverage is hard to guarantee. Third, RLHF human-feedback alignment. Through manual annotation and AI-assisted annotation, it steers the model toward mainstream human values and gets it to actively turn down malicious requests. Anthropic's Constitutional AI pushes the alignment framework further, but the annotation cost stays high. On the plus side, it noticeably cuts down harmful output. On the minus side, the manual annotation is expensive, and heavily disguised adversarial prompts can still slip past the alignment constraints [13, 14].

DOI: 10.54254/2755-2721/2026.36866

Finally, federated learning for distributed training. Multiple organizations train the data locally and only exchange model parameters, so the raw data never leaves each side. The big advantage is that it completely avoids the original training data leaking out. The drawback is that the distributed architecture is complex to set up and hard to scale to very large foundation models.

3.2. In-process control protection technology

There are many in-process control protection technologies. Among them, the core technologies include dedicated risk warning identification models, cross-modal content security review, adversarial sample enhanced training, and API hierarchical real-name permission control. First, the dedicated risk warning identification model independently identifies jailbreak and attack-related input instructions in real time, directly blocking malicious requests. The disadvantage is that the adversarial prompts need to be continuously updated for iterative improvement, and the detection model has the problem of lagging and missed detections. Secondly, cross-modal content security review builds text, image, and audio-video review models respectively to identify forged, violent, and infringing generated content. The disadvantage is that the review has false judgments, resulting in creative and literary content being easily wrongly intercepted. Then, adversarial sample enhanced training continuously adds various malicious prompt samples to the training set to strengthen the model's ability to resist jailbreak attacks. The disadvantage is that the attack methods are evolving rapidly, requiring repeated re-training of the model and continuous consumption of computing power. Finally, API hierarchical real-name permission control distinguishes the calling quotas for individuals, small and medium-sized enterprises, and large enterprises, and mandatory real-name opening for high-risk functions such as deep forgery. The disadvantage is that it only restricts the call frequency and cannot prevent users from actively constructing malicious prompts. This control idea is emphasized in the NIST AI Risk Management Framework [15].

3.3. Post-event traceability and restoration type protection technology

Post-event traceability and restoration technologies are the final safeguarding link in the security protection system. Cross-modal digital watermarking achieves traceability of the source by embedding implicit identifiers in AI-generated content. It is the core technical path for implementing the "Identification Measures for Artificial Intelligence Generated and Synthesized Content", and related text watermark research also provides technical support for the governance of deepfake [16, 17]. The main limitation of this technology lies in its insufficient robustness. Common operations such as cropping and rewriting can easily destroy the watermark information, resulting in the failure of traceability. The complete retention of the full-chain logs records the user account, input prompt words, and model output content, providing original basis for post-event liability determination and violation verification. The high storage and operation costs of massive logs are its main shortcoming. Model vulnerability iteration and fine-tuning continuously summarize new attack samples and security vulnerabilities, and regularly make targeted fine-tuning and repair for the model. Leading large model manufacturers generally adopt this mode to update the security version [18]. Such repairs are always in a passive follow-up state, and the speed of vulnerability optimization is generally behind the iteration speed of attack methods.

DOI: 10.54254/2755-2721/2026.36866

3.4. Limitations of the existing protection technology system

First, there is a natural conflict between efficiency and security constraints and the generated effect: data de-identification, intense adversarial training, and strict value alignment all lead to a loss of the model's ability to understand the context and the fluency of text generation. Most enterprises, in order to ensure the user experience of their products, actively weaken the security constraint configuration. Second, there is a lag in defense: the overall defense model has a significant lag. Detection and defense algorithms can only be optimized after new risks emerge, and they cannot predict unknown risks in advance. Third, modality: the maturity of multi-modal protection technologies is relatively low. Compared with text-based large models, the recognition accuracy of image, audio, and video forgery detection, watermark reinforcement technologies is insufficient. The difficulty of identifying multi-modal forged content is even greater. Fourth, cost: the cost of industry security construction shows a two-tiered distribution. A complete protection chain requires dedicated review models, large-capacity storage, and exclusive computing clusters. The relevant research by Tencent Research Institute confirmed this industry differentiation situation [9].

4. Comparison of domestic and foreign large-model security regulatory normative systems

4.1. Typical foreign AI regulatory policies and laws

First, the EU's AI Act is the world's first dedicated AI law. It takes a risk-tiered approach and classifies general-purpose large models as high-risk systems, with penalties reaching up to 6% of a company's global annual revenue. It puts heavy emphasis on human rights, algorithmic fairness, and copyright protection, and its pre-market access control is strict [10]. Second, the US has issued an executive order on AI rather than unified mandatory legislation. It leans mostly on industry self-regulation, focuses its oversight on large models tied to national security, and pairs this with the NIST risk management framework to give technical guidance for implementation. The regulatory approach is more flexible and puts fewer limits on commercial innovation [15, 19]. Third, Japan has released ethical guidelines for AI development and use. These are soft industry guidelines with no legal force. They encourage companies to set up internal AI review committees and are designed to fit lightweight deployment for small and mid-sized firms, relying on industry self-regulation to control risk [20]. Finally, the OECD's AI governance recommendations form a general baseline ethical framework, and most countries' legislation uses it as a reference point [21].

4.2. China's regulatory system for the security governance of large-scale models

Domestically, China has built a multi-tiered collaborative governance framework of "basic law + special regulations + industry standards", with the core documents being the Interim Measures for the Management of Generative AI Services, the Cybersecurity Law, the Data Security Law, the Personal Information Protection Law, and the New Generation AI Ethics Norms. The Interim Measures for the Management of Generative AI Services is the central regulatory document for domestic generative large models. It makes clear the primary responsibility of AI service providers, and it requires safety assessment before a model goes live, compliance review of training data, labeling of AI-generated content, and real-name registration of users. It strictly bans generating false, discriminatory, or violent illegal content and requires logs of generated output to be kept for later traceability [11]. Supporting national standards for large-model safety assessment have also been issued, which unify the safety-test metrics and evaluation process for foundation models. The

DOI: 10.54254/2755-2721/2026.36866

Safety Specifications for the Application of Government Large Models then sets out special safety requirements for the government vertical [22]. The Data Security Law and the Personal Information Protection Law act as higher-level laws that provide legal backing for protecting users' private data across the whole large-model workflow [23, 24]. The New Generation AI Ethics Norms lays down the underlying value standards for domestic AI research [11], and the Opinions on Strengthening the Governance of Science and Technology Ethics unifies the ethical requirements for AI research across the industry at a macro level [25].

On the ground, leading domestic large-model companies all complete security filing with the cyberspace authorities before launch, and their platforms come with multiple layers of built-in content review and AI-generated markers. There's also a special governance rule for deepfakes, which requires AI-generated audio, video, and images to be marked as AI-sourced, and cracks down hard on forging rumors and AI-driven telecom fraud. China's regulation tries to balance safety control with the growth and innovation of the digital industry — it sets differentiated compliance standards for foundation-model providers and lightweight application providers, so the rules fit the diversity of the domestic AI industry.

4.3. Differences between domestic and foreign regulatory rules and reference for governance experience

To clearly present the characteristic differences in regulatory systems across different regions, this paper conducts a horizontal comparison of large-model safety regulatory rules of the European Union, the United States, Japan and China from dimensions such as governance models and constraint intensity, as shown in Table 1 below.

Table 1. Horizontal comparison of safety supervision systems for Chinese and foreign large models

Regulatory Subject	Core Governance Model	Constraint and Penalty Intensity	Representative Policy Basis
European Union	Legislation-driven risk-based hierarchical strict control model	Extremely strong constraints and heavy administrative penalties	*AI Act*
the United States	Administrative guidance, industry self-regulation and flexible governance	Moderate constraints, mainly relying on safety reporting and administrative interviews	AI Executive Order, NIST AI Risk Management Framework
Japan	Industry ethical guidelines with self-regulation as the priority	Weak constraints with no mandatory penalty clauses	*Ethics Guidelines for AI Development and Utilization*
China	Rigid regulations, hierarchical and categorized management, balance between security and innovation	Moderately strong, graded rectification, fines, service suspension	*Interim Measures for the Management of Generative Artificial Intelligence Services, Data Security Law*

DOI: 10.54254/2755-2721/2026.36866

Looking at the comparison, which regulatory model a region picks is closely tied to its legal system and how far its industry has developed. Given these differences, China can absorb international governance experience from two directions. One is what we can learn from abroad. We can borrow from the EU's risk-tiered governance to further refine our own classification and compliance standards for large models. And we can learn from the US's government-industry joint safety testing — pushing regulators and leading AI companies to build joint safety labs together and share attack-sample databases. The other is where China already has an edge. Real-name user management, full-chain log retention, and mandatory labeling of AI-generated content can efficiently solve the hard problem of holding people accountable for deepfakes. Differentiated, tiered compliance requirements lower the compliance cost for small and mid-sized firms, support innovation among smaller vendors, and fit the AI industry environment at China's current stage of development. This governance model has even been included by the OECD as a typical governance reference case for developing countries [21].

5. Future development directions of large-scale security and ethical governance

5.1. Technical level

First, lightweight and safety-compatible algorithms. We can develop low-loss differential privacy and lightweight adversarial training to balance model performance against safety constraints. Drawing on the ideas behind privacy-preserving reinforcement learning, this is about breaking the long-standing trade-off where safety constraints degrade generation quality, and refining the technical path through that same research direction [12]. Second, highly robust cross-modal watermarking. This is about achieving full traceability across text, images, audio, and video, and cracking the deepfake-tracking problem once and for all. Third, explainable large-model technology. This would break open the black-box limitation and let us pinpoint exactly where bias, leaks, and backdoor triggers happen, which supports the push toward trustworthy AI [1]. Finally, proactive, predictive defense. By studying how attack samples evolve, we can spot new jailbreak and poisoning attacks before they arrive, instead of staying stuck in the current reactive defense model.

5.2. Institutional level

First, refine the tiered regulatory rules. By distinguishing between foundation models, industry-specific models, and lightweight mini-apps, we can lower the compliance cost for small and mid-sized companies. Second, improve the supporting laws on AI copyright. This means clarifying the boundaries of pre-training data use, balancing creators' rights against AI innovation, and borrowing from the EU AI Act's experience with copyright management. Third, build a cross-department coordination mechanism among the cyberspace, public security, and market-regulatory authorities, and put into practice the cross-department collaboration required for tech-ethics governance. Finally, take part in setting global AI governance standards. Building on the OECD framework, we can push for unified rules on cross-border deepfakes and data flows [21].

5.3. Industry level

Firstly, establish an industry security sample sharing platform where enterprises can share anti-virus and backdoor sample libraries, thereby reducing the R&D costs for small and medium-sized companies. This approach has been proven feasible by the Tencent Research Institute [9]. Secondly,

DOI: 10.54254/2755-2721/2026.36866

require large-scale enterprises to set up independent AI ethics committees, referring to the Japanese ethical guidelines and domestic requirements for AI ethics. Then, conduct AI security education for the public to popularize knowledge on deepfake and AI fraud detection, reducing public losses [11, 20]. Finally, cultivate independent third-party security testing institutions to independently complete compliance evaluations for large models and develop third-party evaluation standards for government AI models [23].

6. Conclusion

This article systematically examines two types of risks - information security and ethical security of large models - and analyzes the common causes of these risks from three levels: data, algorithm, and application. Based on the comprehensive research of the entire article, it can be concluded that the security and ethical risks of large models are an endogenous problem caused by data defects, algorithm shortcomings, and human abuse. A single technology or a single law cannot completely resolve the risks. A multi-dimensional collaborative governance system of "technology as the bottom line, legal constraints, and industry self-discipline" must be constructed. This article has two research limitations. First, the entire article mainly conducts qualitative literature review and lacks quantitative experimental data to compare the actual effects of various protection technologies. Second, the research focuses on general basic large models and is insufficient in the personalized risk analysis of specialized large models for high-risk vertical industries such as healthcare and finance. Subsequent research can be conducted in-depth studies in specific scenarios based on the "Government Affairs Large Model Application Security Specifications". Future research can focus on two directions: lightweight security algorithms and specialized governance norms for high-risk industries. A multi-party collaborative trustworthy AI system will be the core guarantee for the long-term healthy development of large models.

References

[1] Zhou, Z. H. (2024). Trustworthy artificial intelligence: Safety, fairness, and explainability. *Chinese Journal of Computers, 47*(1), 1–24.
[2] Su, Y. F., Yuan, J., & Xue, J. M. (2024). Research on the framework of safety assessment system for large models. *Journal of Information Security Research, 10*(5), 432–441.
[3] Hu, B., Hei, Y.-M., Zheng, K. F., et al. (2025). A survey of safety detection and evaluation techniques for large models. *Netinfo Security, 10*, 1–12.
[4] Entezami, E., & Naseh, A. (2025). LLM misalignment via adversarial RLHF platforms (arXiv:2503.03039). *arXiv*. https://arxiv.org/abs/2503.03039
[5] Lu, Z. W., Jin, Q., & Song, R. H. (2022). Research on safety risks and ethical constraints of the Wudao·Wenlan multimodal large model. *ZTE Communications, 28*(2), 25–32.
[6] Romero-Moreno, F. (2025). Deepfake detection in generative AI: A legal framework proposal to protect human rights. *Computer Law & Security Review, 58*, 106162.
[7] Bolukbasi, T., Chang, K. W., Zou, J. Y., et al. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*.
[8] Chen, Q., Liu, D. R., & Shao, J. (2025). Research on the coupling problem of fairness and privacy conflicts in large language models. *Acta Automatica Sinica, 51*(6), 1321–1332.
[9] Tencent Research Institute. (2025). Ethical risks and governance paths of generative artificial intelligence. *Digital Economy of China, 4*, 67–73.
[10] European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
[11] National Governance Committee for the New Generation Artificial Intelligence. (2021). New generation artificial intelligence ethics norms.

DOI: 10.54254/2755-2721/2026.36866

[12] Cho, Y. H., & Sun, W. W. (2026). Privacy-preserving reinforcement learning from human feedback via decoupled reward modeling (arXiv:2603.22563). arXiv.
[13] Anthropic. (2023). Constitutional AI: Harmlessness from AI feedback. Anthropic Research Report.
[14] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback (arXiv:2203.02155). arXiv. https://arxiv.org/abs/2203.02155
[15] National Institute of Standards and Technology. (2023). AI risk management framework (AI RMF 1.0).
[16] Song, Y. M., & Liu, G. S. (2025). Research on AIGC user traceability technology based on text watermarking. *Journal of Applied Sciences, 43*(3), 425–436.
[17] Cyberspace Administration of China. (2025). Measures for the identification of AI-generated and synthesized content.
[18] OpenAI. (2024). GPT-4 safety alignment technical whitepaper. OpenAI Inc.
[19] White House. (2023). Executive order on maintaining American leadership in artificial intelligence. United States.
[20] Ministry of Economy, Trade and Industry. (2024). AI guidelines for developers and users (2024 revised edition). Japan.
[21] Organisation for Economic Co-operation and Development. (2019). Recommendation of the Council on artificial intelligence. OECD.
[22] National Industrial Information Security Development Research Center. (2025). Safety specifications for the application of government large models.
[23] Standing Committee of the National People's Congress. (2021). Data security law of the People's Republic of China.
[24] Standing Committee of the National People's Congress. (2021). Personal information protection law of the People's Republic of China.
[25] General Office of the CPC Central Committee & General Office of the State Council. (2022). Opinions on strengthening the governance of science and technology ethics.