



Current Landscape and Future Directions

Technical Safeguards Against Extreme Misuse of AI

Isabella Duan^{1*}, Edward Kembery^{1*}, Stephen Casper²,
Zhikai Chen³, Jasper Götting⁴, Geng Hong⁵, Zaheed
Kara⁶, Lijun Li⁷, Xiaojian Li^{8,9}, Kellin Pelrine¹⁰, Ziyue
Wang¹¹, Jia Xu⁷, Ziwei Xu¹⁰, Zhiyuan Zeng⁸, James
Zhang¹², Jie Zhang⁷

¹¹ Safe AI Forum

² Harvard Kennedy School

³ Alibaba Group

⁴ SecureBio

⁵ Fudan University

⁶ Centre for the Governance of AI

⁷ Shanghai AI Lab

⁸ Fangcun AI

⁹ Tsinghua University

¹⁰ FAR.AI

¹¹ Independent

¹² Princeton University

*denotes lead authors

Executive Summary

As AI systems grow more capable, the potential for extreme misuse by non-state actors, particularly involving chemical, biological, radiological, and nuclear (CBRN) threats and cyberattacks, has become an urgent concern for governments, companies, academics, and the public.

This report maps five current safeguard practices across three levels: two model-level interventions (pre-training data filtering, and post-training techniques), two deployment-level interventions (monitoring and access control), and one governance-level intervention (pre-deployment evaluations) that informs deployment decisions (including release conditions and safeguard design) and supports their ongoing refinement thereafter. For each intervention, we describe current practice, implementation across frontier developers, empirical evidence, limitations, and priorities for future research and collaboration.

Our findings are as follows:

For managed deployments—including proprietary models and open-weight models served through third-party platforms such as OpenRouter—the most promising safeguards are usage monitoring, access controls, and post-training interventions. Because providers retain control over inference infrastructure, usage monitoring offers the strongest defense; it can detect and block harmful activity in real time while supporting retrospective investigation and enforcement, including account suspension. Similarly, access controls such as Know-Your-Customer (KYC) controls can mitigate misuse risks ex ante by determining who can access an AI system, which capabilities the user can access, and whether they can modify or remove its safeguards. Post-training safeguards (such as reinforcement learning from human or AI feedback) can also increase the probability that models will refuse harmful requests, though evidence shows they remain insufficient against sustained, expert jailbreak attempts. Applied in tandem, these safeguards can substantially raise the difficulty of misuse and increase the speed of response to novel jailbreaks.

¹ For example, China's AI Safety Framework 2.0 (TC260) warns that AI training corpora contain broad foundational knowledge relevant to nuclear, biological, and chemical weapons, and that without effective controls, extremist groups and terrorists could exploit AI systems—particularly those augmented with retrieval capabilities—to acquire such knowledge and develop capabilities for designing, manufacturing, synthesizing, and deploying weapons of mass destruction (WMD). The framework calls for improving traceability of AI system end uses to prevent deployment in high-risk scenarios such as WMD production.

[原文: “人工智能在训练过程中多使用门类宽泛、内容丰富的语料数据, 其中包含核生化武器相关基础理论知识, 辅以检索增强生成功能, 如不能有效管控, 将被极端势力、恐怖分子利用于获取相关知识, 以及设计、制造、合成、使用核生化武器能力, 导致现有管控体系失效, 加剧世界各地和平安全威胁……提高人工智能系统最终用途追溯能力, 防止被用于核生化导等大规模杀伤性武器制造等高危场景”。]

For open-weight models deployed locally or privately, the most promising safeguards are pre-training data filtering and post-training interventions, although both have fundamental limitations. Pre-training data filtering can reduce model capabilities on tasks requiring specialized harmful knowledge and may be more resistant to adversarial fine-tuning than post-training safeguards. However, successfully repressing a model's dangerous capabilities using data filtering while retaining general model performance is difficult when dangerous capabilities draw on widely distributed knowledge (e.g., hacking from coding). Moreover, research has shown that dangerous capabilities can often be restored after deployment, either through fine-tuning or model relearning. Post-training safeguards can likewise be trivially circumvented by actors with direct access to model weights via techniques like fine-tuning. Overall, these safeguards should be standard practice because they can increase the cost of misusing open-weight models. However, the ease with which they can be circumvented or removed means they offer limited protection against determined or experienced actors.

Across all model types and deployment settings, pre-deployment evaluations for dangerous capabilities and safeguard robustness are essential inputs for responsible release decisions. Neither proprietary nor open-weight models are inherently dangerous; risk depends on their capabilities and the robustness of the safeguards surrounding them. As frontier models acquire increasingly powerful dual-use capabilities with the potential to cause real-world harm, measuring those dangerous capabilities and adopting baseline safeguards while rigorously evaluating their effectiveness should become a global norm. This approach can support informed deployment decisions and strengthen protections over time.

Figure 1, below: an approximate high-level overview of misuse safeguards and their efficacy, maturity, and cost-effectiveness from publicly available information. Note that internal AI company practices may go unreported.

Practice	Adoption Reported by Leading Developers	Efficacy	Maturity	Estimated Cost
Pre-training data filtering	Limited to public knowledge OpenAI reports CBRN-specific filtering. Anthropic has published research but not confirmed in production. Chinese developers filter for general safety per regulation but not CBRN-specific.	Moderate Reduces access to dangerous knowledge, although models may still acquire it through in-context learning and Internet access.	Low Recent proof of concept research demonstrates viability, but limited production deployment.	Low Typically requires a negligible fraction of pre-training budget; requires domain-specific classifier development.
Post-training techniques	Widespread Anthropic, OpenAI, Google DeepMind, Zhipu, Alibaba, and others all report using RLHF/RLAIF.	Moderate Effectively steers model behavior but remains vulnerable to expert jailbreaking and malicious fine-tuning. Nevertheless, it can substantially increase the cost of misuse.	High Widely deployed across industry.	Moderate Typically requires a small fraction of the data and compute of pre-training. Improving robustness requires some sustained R&D.
Access control	Uneven Anthropic and OpenAI track usage and enforce policies. Meta and OpenAI assess risks before weight release. Hugging Face/ModelScope provides limited gating.	Moderate KYC requirements, usage monitoring, and account suspension can mitigate misuse in managed deployments. Still, sophisticated actors may circumvent these controls.	High Standard practice for proprietary models; nascent for open-weight.	Low Primarily operational/administrative costs.

Practice	Adoption Reported by Leading Developers	Efficacy	Maturity	Estimated Cost
Monitoring and response	Widespread for proprietary models Anthropic uses Constitutional Classifiers. OpenAI monitors API/chat. Chinese developers monitor per regulation, primarily targeting illegal content not CBRN.	High Represents the strongest currently available misuse safeguard in managed deployments, although sophisticated jailbreaks remain possible.	High Production-deployed at major labs with recent advances reducing cost.	Moderate to low While running advanced classifiers can be expensive, other techniques such as probe-based approaches have lower inference costs. Some forms of classification (e.g., real-time routing or system instruction) add latency.
Pre-deployment evaluations	Uneven Anthropic, OpenAI, and Google DeepMind conduct CBRN assessments. UK AISI evaluated 30+ systems. Chinese developers report general benchmarks but not CBRN-specific.	High Supports informed deployment decisions and appropriate safeguard calibration.	Moderate Good evaluation practices exist but remain unevenly distributed throughout industry.	Moderate For capability evaluations, real-world uplift trials are costly. Cheaper automated tests may lack validity. Safeguard evaluations can cost-effectively identify weaknesses and improve protections.

Introduction

As AI systems grow more capable, the potential for extreme misuse by non-state actors—particularly involving chemical, biological, radiological, and nuclear (CBRN) threats and cyber threats—has become an urgent concern shared by researchers, developers, and policymakers worldwide. While some progress has been made in developing technical safeguards to mitigate these risks, implementation practices vary across organizations and jurisdictions, and there is limited consensus of what safeguards developers should implement and how to effectively layer them as part of a defense-in-depth approach.

This report provides an overview of commonly used model- and deployment-level safeguards for addressing extreme misuse risks, drawing on recent research and disclosed practices from leading AI companies. Our goal is to establish a shared technical foundation by mapping the current safeguard landscape, identifying where practices converge, and surfacing implementation bottlenecks and open questions where cross-border coordination is most needed. We hope this work serves as a starting point for building international consensus on adopting baseline technical safeguards to mitigate extreme misuse risks.

The report is primarily intended for developers of foundation models, including both model developers and application and service providers, as well as academic researchers and policymakers interested in understanding safeguards and addressing bottlenecks. While we focus on extreme misuse risks, these safeguards can also apply to other categories of misuse.

Model Level Safeguards

1. Pre-Training Data Filtering

Overview

Pre-training data filtering can reduce a model's exposure to and command of dangerous knowledge obtained during training. Although the technique still has yet to be extensively tested, it currently represents one of the most promising mitigations for extreme CBRN misuse risks in open-weight model deployments.

Existing Practices

Filtering harmful data from pre-training corpora is widely understood as a key means of training safer models (e.g., Longpre et al., 2025; Casper et al., 2026), but filtering specifically for CBRN knowledge remains an emerging practice. OpenAI has reported using pre-training data filters for harmful data content, especially around hazardous CBRN knowledge for its open-weight models as well as its proprietary model (GPT-4o) (OpenAI, 2024; Wallace et al., 2025a). Anthropic has publicly disclosed pre-training data filtering experiments on CBRN risks (Chen et al., 2025) and has published on the related topic of modular pretraining (Roland et al., 2026), but has yet to report applying the methods in production for safety purposes.

Google DeepMind and IBM have reported applying pre-training data safety filtering, but did not explicitly report filtering for CBRN risks ([Google DeepMind, 2025](#), [IBM, 2024](#)).

Chinese developers operate under regulatory requirements to filter pre-training data across a list of outlined risk categories ([TC260, 2023](#)), although not explicitly addressing CBRN risks. Alibaba, DeepSeek, Tencent, and Zhipu have reported pre-training data filtering for safety considerations, but similarly make no specific mention of CBRN risks ([Qwen Team, 2025](#); [DeepSeek Transparency Report, 2025](#); [Tencent Hunyuan Team, 2024](#); [Zhipu and Tsinghua, 2024](#)).

Implementation

Anthropic's research ([Chen et al., 2025](#)) provides a detailed account of a CBRN-focused filtering pipeline. Documents in the pretraining corpus are scored for harmfulness using a trained classifier and removed above a tunable threshold. The researchers evaluated six classification methods (including constitutional classifiers, holdout loss scoring, bigram bag-of-word classifiers, and named-entity approaches) with the prompted constitutional classifier achieving the highest accuracy. Training data for the classifiers was generated synthetically via LLM prompting, avoiding manual annotation of sensitive content. They found that filtering reduced harmful CBRN knowledge on the tested models by 33% relative to random-chance baseline on the WMDP benchmark, with no significant degradation in general capabilities (e.g., MMLU, Code, Prose).

Concurrently, [O'Brien et al. \(2025\)](#) demonstrate a complementary multi-stage pipeline (i.e., keyword blocklist screening followed by classifier escalation), which they implemented at a cost of 0.8% of the total pretraining budget. In addition to improving model safety off-the-shelf, they found that pre-training data filtering also confers significant resistance against tampering, which can be key for mitigating misuse risks for open-weight models. Their filtered 6.9B parameter models resisted up to 10,000 steps and 300M tokens of adversarial fine-tuning, an order-of-magnitude improvement over post-training baselines, and were also robust to "benign fine-tuning" attacks that inadvertently restore suppressed capabilities.

Limitations

- 1. Scaling challenges.** Larger models trained on more data may extract hazardous knowledge from fewer examples despite safety-related mitigation efforts ([Souly et al., 2025](#); [Bowen et al., 2025](#)), requiring higher filtering recall to achieve the same suppression effect.
- 2. Limited scope of applicability.** Filtering is highly effective for mitigating CBRN misuse, which depends on specialized knowledge specific to a harmful domain (e.g., virology). However, it might offer more limited protection against cyberattacks, influence operations, or harmful speech generation if relevant knowledge is widely distributed across general corpora. In other domains, like toxicity reduction, research has found that pre-training on more toxic data can lead to better controllability in post-training and an overall decrease in toxicity ([Li et al., 2025](#)).
- 3. Difficulties defining dangerous knowledge.** Building classifiers to distinguish harmful from legitimate dual-use knowledge remains challenging. Aggressive filtering risks degrading performance on beneficial scientific applications. Targeted removal of highest-risk content (e.g., protocols for obtaining live high-risk pathogens) may offer acceptable trade-offs.

4. **Failure due to in-context learning.** Models can leverage harmful information provided in-context or learned during malicious fine-tuning, necessitating layered defenses combining filtering with other safeguards (O'Brien et al., 2025). However, this may not apply to the most sensitive content (e.g., WMD specifications).

Bottlenecks to Implementation

1. **Lack of research and evaluation infrastructure.** Validating filter effectiveness requires red-teaming capabilities and domain-specific evaluation datasets that smaller organizations cannot develop in-house. Building and testing filters also requires running pre-training experiments—a substantial compute cost that academic researchers cannot easily bear. Further research is needed to build strong empirical evidence that filtering does not degrade general model performance and to verify that these methods are practical to implement at scale.
2. **Lack of domain expertise in companies and internationally.** AI companies often lack the domain expertise to independently identify dangerous CBRN knowledge and build specialized filters. While some specialized capacity resides in governments and research centers, few global entities (e.g., SecureBio) publish technical guidance bridging AI and biosecurity. Moreover, transferring existing production filters or the information necessary to build them beyond Western contexts may be restricted by regulations on hazardous information transfer.
3. **Language barriers.** Most current tools are designed and evaluated for English corpora, leaving non-English developers without proven language-specific methods or standard benchmarks for measuring filtering performance. Internet-scale datasets therefore require robust multilingual filtering capabilities.

Future Work Addressing Bottlenecks

1. **Establish research and governance infrastructure.** Governments or industry standard bodies should develop categorical exclusion lists for CBRN, child sexual abuse material (CSAM), and other hazardous data filtering. Governments should fund dedicated research hubs (e.g., AI×Bio Risk Research Groups). These hubs, equipped with both AI and threat-domain expertise, could develop and validate CBRN-specific filters, producing off-the-shelf tools and evaluation benchmarks that smaller developers can adopt. Collaboration with industry would ensure usability and alignment with production pipelines.
2. **Standardize filtering categories.** Expert groups could produce taxonomies of filtering categories that target highest-risk knowledge while preserving beneficial applications. BioTIER, for example, recommends excluding narrowly defined high-risk biomedical and related operational topics while retaining dual-use and benign biology (Marshall et al., 2026). Refining and building consensus around such frameworks could accelerate adoption of pre-training filtering.
3. **Measure effectiveness and trade-offs.** Future research should develop standardized benchmarks measuring the extent to which filters effectively mitigate harmful behaviors while preserving both general model performance and domain-specific performance in relation to high-risk knowledge subsets.

4. **Improve existing filtering methods and explore alternative approaches.** For example, rather than trying to perfectly classify and exclude dangerous data upfront, Selective Gradient Masking localizes dangerous knowledge into dedicated model parameters during training, which can then be removed while preserving general capabilities ([Shilov et al., 2025](#)). It may be possible to make this technique more effective by localizing specific knowledge using gradient routing ([Cloud et al., 2024](#)).
5. **Validate cross-lingual transfer.** Future research should validate the effectiveness of filters on models trained on non-English data.
6. **Explore alternative methods to remove harmful capabilities.** One alternative to filtering pre-training data would be to perform unlearning on the model, then distill the unlearned model. Early tests on a 100M-parameter model suggest this approach can achieve similar resistance to harmful behavior as pre-training filtering, potentially using less compute and labeled data ([Lee et al., 2025](#)).

2. Post-Training Techniques

Overview

There are two main post-training techniques for teaching models to handle harmful requests appropriately: reinforcement from human feedback (RLHF) and reinforcement from AI feedback (RLAIF) ([Christiano et al., 2017](#); [Ouyang et al., 2022](#); [DeepMind, 2022](#); [Bai et al., 2022](#); [Guan et al., 2024](#)). Targeted safety effects of such post-training include shifting model preferences toward safer responses, suppressing harmful capabilities or internal representations, and improving the robustness of safety behavior to adversarial elicitation or downstream modification.

Overall, both methods have been proven to be generally effective at steering models to handle harmful requests appropriately, with some limitations. Further research into best practices for RLAIF may discover more effective mechanisms.

Existing practices

Several companies report using post-training to mitigate harmful responses from AI models. Anthropic reports using RLHF and RLAIF to train their Claude models to respond harmlessly ([Anthropic, 2025a](#); [Anthropic, 2025b](#)). OpenAI has used RLAIF to train their most advanced models to align with a model spec “to prevent harm” ([Guan et al., 2024](#); [OpenAI, 2025a](#)) since o1 ([OpenAI, 2024](#); [OpenAI, 2025b](#); [OpenAI, 2025c](#); [OpenAI, 2025d](#)). They also reported training GPT-4.5 via RLHF ([OpenAI, 2025e](#)).² Google DeepMind trained Gemini 2.5 to comply with “safety policies” via RLAIF ([Comanici et al., 2025](#)). Zhipu used RLHF for the “safety” of ChatGLM ([Zhipu, 2024](#)), GLM-4.5 ([Zhipu, 2025](#)), and GLM-4.5V ([Zhipu, 2026](#)). Alibaba reported using techniques like RLHF and RLAIF to improve the ‘harmlessness’ of Qwen 2.5 ([Qwen Team, 2025](#)).³

Implementation

Reinforcement learning from feedback typically proceeds in two phases, which differ slightly between RLHF and two popular forms of RLAIF: constitutional AI (CAI), and deliberative alignment (DA).⁴

1. Supervised learning phase

- **RLHF:** A pre-trained model is fine-tuned on high-quality demonstrations produced by human annotators.
- **DA:** Fine-tuning data is generated by inserting safety principles into the system prompt and prompting the model to reason about these principles while responding to harmful queries ([Guan et al., 2024](#)).
- **CAI:** A helpful-only model first responds to harmful prompts; its responses are then critiqued and revised according to constitutional principles, and the revised outputs are used for fine-tuning ([Bai et al., 2022](#)).

2. Reinforcement learning phase

- **RLHF:** Human raters compare pairs of model outputs, producing preference data used to train a reward model.
- **DA:** The reward model itself is an AI system that has the safety specifications in its prompt and direct access to the specification during evaluation ([Guan et al., 2024](#)).
- **CAI:** The fine-tuned model generates paired responses to harmful prompts, which are evaluated (via multiple-choice questions) on how well they follow constitutional principles, and these judgments train the reward model ([Bai et al., 2022](#)).

Limitations

- 1. Proprietary models remain vulnerable to jailbreaking.** Public research by UK AISI, as well as private reports, indicate that even models fine-tuned to refuse harmful requests can still be jailbroken by experts ([UK AISI, 2025](#)). UK AISI also recently published evidence of automated jailbreaks that can bypass even the strongest safeguards used in industry ([Davies et al., 2026](#)).
- 2. Open models remain vulnerable to fine-tuning attacks.** Fine-tuning models can easily remove their safeguards ([Qi et al., 2024](#), [Murphy et al., 2025](#)). Methods reporting to produce models that are more robust to adversarial fine-tuning (e.g., [Tamirisa et al., 2024](#)) have been shown to be ineffective against cheap, realistic attacks such as adversarial fine-tuning and boundary point jailbreaking ([Kuo et al., 2026](#); [Zloczower et al., 2026](#); [Davies et al., 2026](#)). Furthermore, even if models became more robust, malicious actors might still be able to find specific ways to subvert safeguards by re-purposing deep-reinforcement learning algorithms ([Willison, 2023](#), [Liu et al., 2023](#), [Rao et al., 2023](#), [Wei et al., 2023](#); [Liu et al., 2025](#)).
- 3. Improper training may lead to overrefusal.** Post-training can cause models to refuse benign requests that share lexical or conceptual features with harmful queries. This is particularly problematic in biology, chemistry, and cybersecurity, where legitimate and malicious uses often rely on overlapping knowledge.
- 4. Generalization from reinforcement learning is uneven across deployment contexts.** Refusal training based on reinforcement learning with human feedback has been shown to be less effective for models operating in low-resource languages ([Yong et al., 2023](#)), under very long contexts ([Anil et al., 2024](#)), and in agentic or tool-use settings where the harmful action is composed of benign steps ([Glukhov et al., 2024](#)).

Future work addressing bottlenecks and limitations

1. **Further research into RLAIIF.** RLAIIF is still a nascent technique without a consensus for best practices, and may not always imbue the desired harmless behavior ([Zhang et al., 2025](#); [Schoen et al., 2025](#)). Building better evaluations for spec adherence could allow developers to more accurately compare the efficacy of different post-training processes ([Ahmed et al., 2025](#)).
2. **Tamper-resistant open weight models.** While existing research into training methods that make open-weight models more robust against fine-tuning attacks has not been effective ([Huang et al., 2024](#); [Sheshadri et al., 2024](#); [Tamirisa et al., 2024](#)), it is possible that training models against a wider variety of tampering threats might confer more generalizable tamper resistance ([Casper et al., 2025](#)).
3. **Training models to be confused about dangerous topics.** Even if models lack dangerous capabilities, it is possible that malicious actors could fine-tune them in order to acquire them ([Kaunismaa et al., 2026](#)). In this case, one way to mitigate risks is to instill confused or mistaken beliefs about dangerous topics ([Wang et al., 2025](#); [Slocum et al., 2025](#)); however, this requires further study.
4. **Building better datasets for safety fine-tuning.** Developing open-source or securely shared datasets that improve safety without degrading general capabilities could reduce the cost of building safer models ([Wang et al., 2025](#)).

Deployment Level Safeguards

3. Access Control

Overview

Access controls mitigate misuse risks *ex ante* by determining who can access an AI system, which capabilities they can use, and whether they can modify or remove its safeguards.

For proprietary models, developers can implement KYC requirements for end users and revoke access for those with documented histories of model misuse.

For open-weight models, however, weights can be downloaded and shared freely online, making it challenging to deny access to malicious actors. However, distributors can still usefully intervene. First, hosting platforms for open-weight models (e.g., [Hugging Face](#), [ModelScope](#), or the model developers themselves) can refuse to distribute jailbroken open-weight models or require KYC verification prior to permitting downloads of systems with significant misuse potential. Second, when AI service and application providers build upon open-weight models, they can implement access controls analogous to those employed for proprietary models.⁵

Existing practices

Foundation model developers use a number of techniques to control access to their models. For example, both Anthropic and OpenAI track chat and API interactions in order to enforce usage policies of their proprietary models ([Anthropic, 2025c](#); [OpenAI, 2025f](#)).⁶ This allows them to selectively monitor and block or ban users that might have violated the terms of the user policies. OpenAI also offers “organization verification” and provides “trusted access” modes to select users ratified by a government-issued ID ([OpenAI, 2026a](#); [OpenAI, 2026b](#); [Clymer et al., 2025](#)).⁷

Some frontier model developers have indicated that they will only release model weights publicly when they are confident that the model will not increase risks from malicious actors. For example, OpenAI released GPT-oss models after tests showed that they were not capable of “high risk” even when fine-tuned ([Wallace et al., 2025b](#)). Meta committed to performing and validating risk mitigation techniques before open-sourcing models with sufficient risk ([Meta, 2026](#)).

Model hosting platforms also control access to models to some extent. Model authors can request that users share their contact information before accessing model files and reject requests they do not approve ([HuggingFace, 2023](#); [ModelScope, 2024](#)). Model hosting platforms can also remove models that violate its content policy or violate copyright ([HuggingFace, 2023](#)).

Implementation

Implementation of access controls differs depending on the nature of the provider (e.g., foundation developer or model distribution platform) and the purpose of the access control.

Some example strategies include:

- **For developers:** Providing a chat interface or API; providing access to new models or greater numbers of tokens to specific registered users; monitoring and banning users that violate usage policies.
- **For distributors:** Gating models from being accessed without certain identification; removing models that violate hosting policies.

For each stage, developers and distributors can opt for different levels of user identification depending on their assessment of risks. These include:

- **Basic information:** Set up an account, have a registered email, share phone number, click through acceptable use terms.
- **Moderate justification:** Submit an application, share any affiliated institutions, justify the usage.
- **Robust authentication:** Share personal identifying information (such as government ID), check legitimacy by verifying user affiliation with the institution, check against government watch lists.

Limitations

1. **Even well-monitored proprietary APIs may fail to detect misuse.** For example, one developer reported cybercriminals used their models to develop, market, and distribute several variants of ransomware (e.g., [Anthropic, 2025d](#)). Although these accounts were eventually banned, the incident illustrates that sufficiently motivated actors can conduct meaningful misuse before being detected.

2. **Access to open-weight models is hard to revoke.** Current practices for releasing model weights tend to rely on assessments of the risk posed by that model ([Wallace et al., 2025b](#); [Meta, 2025](#)). However, a model deemed safe given fine-tuning at one point in time might be fine-tuned to elicit more dangerous capabilities by malicious actors in future; for example, fine-tuning on synthetic data generated by more powerful future models ([Kaunismaa et al., 2026](#)).
3. **Model distributors cannot prevent models downloaded on their platforms from being shared elsewhere.** While model distributors have some mechanisms by which they can influence which models get accessed on their platforms, they cannot prevent models downloaded from their platforms from being shared elsewhere, or gain insight into how those models are actually used.
4. **Limiting access can bear an opportunity cost in potential benefits from AI adoption.** Providing people with access to models, especially open-weight models, may have benefits for people who use modified versions of those models or wish to experiment with these models for science or safety purposes ([Seger et al., 2023](#)).
5. **The model might still be leaked or stolen.** Restricting public access to model assets cannot prevent theft from data centers or internal leaks. Nonetheless, it would still substantially reduce the number of people capable of accessing the model, since few have the means to steal it directly, and it may be possible to use internal access controls to mitigate risks from leaks ([Nevo et al., 2024](#)).

Future work addressing bottlenecks and limitations

1. **Better usage monitoring.** Being able to selectively ban or block users on chat or API interfaces for proprietary models is downstream of being able to reliably flag violations of the usage policies as they occur. Further research could identify mechanisms for detecting particular threat vectors like multi-agent cyber espionage, CBRN uplift, or better methods for monitoring computer use in general ([Sumers et al., 2025](#)).
2. **Controlled model deployment.** Developers don't need to release model weights without restrictions. In split deployment, only part of the model is shared, and the developer remains involved in inference ([Xie et al., 2025](#)). Alternatively, weights can be restricted to pre-approved hardware ([Clifford et al., 2025](#)). Although still nascent, both approaches warrant further research as potential access control options.

4. Monitoring and Response

Overview

Once access is granted to a particular group, misuse monitoring systems can analyze model interactions and broader usage patterns to identify potentially harmful activity. They flag suspicious patterns (e.g., repeated queries about synthesizing dangerous chemical compounds) and trigger appropriate responses (e.g., model downgrading, response rate limiting, access revocation, or notifying relevant authorities) either in real time or through post hoc log analysis.

For proprietary models, developers control the inference infrastructure and can implement monitoring effectively, making this the strongest current misuse safeguard.

For open-weight models, monitoring depends on deployment:

- **Managed deployments** (AI service and application providers): Downstream deployers can implement monitoring similar to proprietary models.
- **Self-hosted deployments** (local or cloud): No monitoring is currently feasible without platform or user cooperation.

Cloud compute monitoring, where providers like AWS and Alibaba Cloud automatically run extreme misuse monitors for models hosted on their platforms, could be a promising approach (Heim et al., 2024). However, even inference-level monitoring at this layer raises practical questions around privacy and liability, since cloud providers would be inspecting customers' inputs and outputs on their platforms. More ambitious goals, such as detecting malicious fine-tuning and scaffolding, face additional technical challenges in distinguishing malicious from legitimate workloads (Sun et al., 2026).

Existing Practices

Model developers globally already deploy input/output classifiers that flag potentially harmful requests, including CBRN threats, with flagged interactions subject to human review and potential account suspension (Anthropic, 2025e; OpenAI, 2025g). For example, Anthropic uses real-time classifiers to monitor model inputs and outputs and block the model from producing threat-related information, and uses offline classifiers to identify jailbreak attempts (Anthropic, 2025e). AI companies can also red-team the effectiveness of classifiers with bug bounties and other evaluations, developing a system of rapid response to new jailbreaks. OpenAI recently dedicated 700,000 A100e GPU hours to automated red-teaming to discover universal jailbreaks for GPT-5.6 Sol safeguards (OpenAI, 2026). Another strategy is to use activation probes to catch malicious outputs. For example, Google DeepMind uses misuse mitigation probes in production (Kramár et al., 2026).

Chinese regulatory frameworks require AI service providers to implement content monitoring systems, with mandatory reporting of illegal content to authorities (China Law Translate, 2023). Major Chinese companies operate content moderation systems, though these primarily target illegal content rather than CBRN or advanced cyber risks specifically. For example, DeepSeek uses real-time keyword matching to flag problematic queries, which are then reviewed against a preset risk prompt to determine whether to retract the dialogue (Guo et al., 2025). Alibaba Cloud provides AI Guardrails as a service for AI application providers' content compliance detection (Alibaba Cloud, 2025).

Implementation

Monitoring systems employ varied architectures with distinct cost-security trade-offs:

- **Online versus offline monitoring.** Online monitoring screens queries and outputs in real-time, blocking harmful content before delivery. Originally, this presented a cost—Anthropic's original Constitutional Classifiers imposed a ~24% inference overhead limiting widespread adoption (Sharma et al., 2025)—but recent techniques have reduced this to closer to 2% (Cunningham et al., 2025).

- Offline monitoring analyzes logs post hoc for policy violations, enabling account suspension without latency costs but failing to prevent first-instance harms. Companies typically combine both real-time blocking for CBRN risks and retrospective analysis for broader enforcement.
- **Monitoring architectures.** Different technical approaches offer varying cost-accuracy trade-offs.
 - **Keyword matching** uses predefined lists to flag problematic queries in real-time ([Guo et al., 2025](#)).
 - **LLM-based classifiers** employ separate language models to evaluate whether inputs or outputs violate safety policies. These can be fine-tuned (e.g., Anthropic’s Constitutional Classifiers, trained on synthetically generated harmful and harmless examples, achieves 95% jailbreak blocking. See more in [Sharma et al., 2025](#)) or prompted (with input caching). However, running language models as classifiers is computationally expensive.
 - **Probes** are simple classifiers used to determine what information a model’s internal representations contain. They offer a more efficient alternative, achieving similar performance to Constitutional Classifiers with just 2% of the parameters ([Cunningham et al., 2025](#); [Kramár et al., 2026](#)).
 - **Multi-stage systems** combine these approaches by using lightweight probes for initial screening with expensive classifiers for escalation, cutting costs 40x while maintaining robustness ([Cunningham et al., 2026](#)).
- **Monitoring scope.** Conventional misuse monitors inspect model inputs and outputs and, where available, reasoning traces. Even monitoring encrypted reasoning traces is important because encryption does not guarantee their confidentiality. [Panfilov et al. \(2026\)](#) demonstrate a decryption jailbreak that can recover hazardous information in proprietary models’ reasoning traces even when the visible response safely refuses the request.⁸ For AI agents, monitoring may also need to cover complete execution trajectories, including plans, tool calls, code execution, memory updates, and environmental effects, because harmful behavior can emerge over extended, multi-step workflows when no individual action is obviously malicious ([Williams et al., 2026](#); [Shanghai Artificial Intelligence Laboratory, 2026](#)).
- **Escalation mechanisms.** Flagged content can trigger automatic blocking (for high-confidence violations), human review (for ambiguous cases), and logging. This data can then provide a metric to improve on over time; for instance, Anthropic’s system reports 0.05% false positives through careful threshold calibration ([Cunningham et al., 2026](#)).

Limitations

1. **Inapplicability to self-hosted open-weight deployments.** Monitoring generally requires control over inference infrastructure. Once model weights are released, developers lose visibility into downstream usage, rendering monitoring ineffective for self-hosted open-weight models.
2. **Adversarial brittleness.** Despite extensive red-teaming, defenses remain imperfect ([McKenzie et al., 2025](#)). For example, UK AISI recently found an attack that bypassed Anthropic’s Constitutional Classifiers, as well as the safeguards on OpenAI’s GPT-5.2 ([Davies et al., 2026](#)). Attacks can also be broken up into seemingly benign steps and distributed through regular traffic for greater effectiveness ([Sun et al., 2026](#)).

3. **Dual-use ambiguity.** Training small models to distinguish legitimate research from malicious intent remains challenging, and can lead to high false positive rates. For example, Anthropic’s original paper noted that “while constitutional classifiers achieve a low false-positive rate of 1.50% on MMLU–Chemistry questions, we observe notably-higher FPRs of 26.05% on GPQA–Chemistry questions” ([Sharma et al., 2025](#)).
4. **Different systems may need separate classifiers.** Malicious code generation, fraud, and other misuse categories may require separate classifiers from those designed to mitigate CBRN risks, or broader training data, multiplying monitoring complexity and cost.
5. **Adversaries might use multiple models to perpetrate harm.** Adversaries may distribute a harmful workflow across multiple models, making each request appear benign in isolation. Classifiers may therefore miss the overall intent, allowing a capable model to complete legitimate-looking subtasks while less safeguarded models handle explicitly harmful steps ([Jones et al., 2024](#)).
6. **Inference cost constraints.** While recent advances in cascading architectures and representation reuse have greatly reduced the cost of classifiers ([Cunningham et al., 2025](#)), more thorough versions of classifier-based monitoring still impose some computational costs at scale.

Bottleneck to Implementation

1. **Technical complexity of probes.** Probes require significant technical expertise to develop and must be rebuilt each time model weights change, which may be prohibitive for low-resource developers. For open-weight models, the academic community could help by developing and sharing probes.
2. **Setup required to reduce inference costs and latency.** While multiple companies have adopted effective monitoring systems without prohibitive inference or latency costs, setting these up for deployment can require considerable time, organizational coordination, and upfront investment.
3. **Standardization challenges.** It may be difficult to develop a clear consensus as to what the precise classifier thresholds should be. Without widely-adopted standards for monitoring misuse, adversaries may “model shop” by routing harmful queries to providers with weaker safeguards.
4. **Classifier training data scarcity.** Effective monitoring requires datasets of adversarial attacks and harmful outputs, which are sensitive to share and difficult to curate comprehensively.

Future Work Addressing Bottlenecks

1. **Research in probe monitoring and other low-cost approaches.** Analyzing model internal states rather than running separate classifiers could dramatically reduce costs. The literature remains nascent and warrants continued validation in production.
2. **Coordination between companies on key threat categories.** Rather than pursuing comprehensive standardization, initial agreement on highest-priority threat categories could reduce model shopping while allowing companies flexibility on threshold calibration.

3. **Tiered monitoring architectures.** Most major companies already use expensive real-time classification for high-risk threats (e.g., CBRN) and cheaper offline analysis for broader policy violations. This allows them to balance cost with security. Further analysis into tiered systems could help companies to develop cheaper, more effective monitoring.
4. **Shared secured datasets for building effective classifiers.** Establishing secure data-sharing mechanisms, similar to biosecurity frameworks for pathogen research, could accelerate classifier development across the industry. One approach here might be to build consortium-managed datasets with controlled access for vetted researchers.
5. **Share best practices on evaluating and improving classifier robustness.** This includes: common benchmarks for measuring false positive/negative rates across different threat categories, red-teaming protocols to systematically identify classifier weaknesses, and disclosure norms for discovered jailbreaks while balancing security concerns.

Governance Level Safeguards

5. Pre-Deployment Evaluations

Overview

Pre-deployment evaluations encompass both dangerous-capability evaluations, which assess a model's ability to perform or meaningfully assist with dangerous tasks, and safeguard evaluations, which assess how effectively safeguards prevent malicious actors from accessing those capabilities. Their findings inform risk assessments and, in turn, release and deploy decisions to help iteratively improve safeguards ([Shanghai AI Lab and Concordia, 2025](#)). In this way, they are foundational to all other misuse safeguards.

Existing Practices

Anthropic, Google DeepMind, and OpenAI assess whether their models can “uplift” low- and medium-resourced threat actors to create, obtain, and deploy CBRN weapons beyond what is available through conventional means ([Anthropic, 2025b](#); [Google DeepMind, 2025](#); [OpenAI, 2025d](#)). Other developers (e.g., DeepSeek, Zhipu, Alibaba, ByteDance, Tencent) document general safety benchmarks but have not published CBRN threat assessments.

Government and third-party evaluations provide a cornerstone of accountability and visibility into misuse risks. For example, UK AISI has conducted biological, chemical, and cyber evaluations on over 30 AI systems ([UK AISI, 2025](#)), while CAICT evaluated the security of 15 open-weight systems for offensive cyber capabilities ([CAICT, 2025](#)). [Shanghai AI Lab \(2025\)](#) and [Concordia AI \(2026\)](#) have similarly evaluated diverse open-weight and proprietary models against frontier risk management frameworks covering biological, chemical, and cyber threats.

Implementation

Methods for evaluating model's dangerous capabilities include:

- **Automated assessments:** Multiple-choice questions testing domain knowledge, open-ended questions probing specific bottleneck steps, and agentic evaluations measuring multi-step task completion. Key benchmarks for bio risks include LAB-Bench (biological research capabilities; see [Laurent et al., 2024](#)), SecureBio VCT (virology protocol troubleshooting; see [SecureBio, 2026](#)), and other custom biosecurity evaluations. Developers and researchers can also build their own custom assessment environment cheaply and quickly using automated tools like Petri ([Fronsdal et al., 2025](#)).
- **Expert red-teaming:** Third-party red-teaming and biodefense experts conduct scenario-based exercises covering bacterial and viral threat pathways, assessing ideation, acquisition, magnification, formulation, and release stages.
- **Uplift trials:** Human studies measuring whether model access improves participant performance on sensitive tasks compared to internet-only baselines (e.g., [Anthropic's \(2025\)](#) virology uplift trial for challenging virus reconstruction protocols and [Frontier Model Forum's \(2025\)](#) wet-lab trials assessing real-world biological uplift).

Beyond assessing and eliciting dangerous capabilities, misuse evaluations also include **safeguard robustness evaluations**, which test whether safeguards remain effective under adversarial pressure. This includes red-teaming against jailbreaking, fine-tuning, and scaffolding attacks ([Zou et al., 2023](#), [Mazeika et al., 2024](#); [Murphy et al., 2025](#)). As models are already highly capable in CBRN and offensive cybersecurity domains, running safeguard robustness evaluations is a priority because they are critical to informing the overall risk posture of a deployment. They also provide a useful signal for how safeguards should be improved going forward. Recent examples include FAR.AI's AI Security Leaderboard, which benchmarks frontier models' resistance to jailbreaks in CBRN, cyber, and explosive threats contexts ([Timm et al., 2026](#)) and Concordia AI's red-teaming initiative, which tests safeguard robustness across multiple risk domains ([Concordia AI, 2026](#)). Beyond robustness, a growing body of research examines whether safeguards appropriately balance the need to block harmful dual-use requests against the risk of hindering legitimate scientific and defensive research through over-refusal ([OpenAI, 2026](#); [Zhao et al., 2025](#)). For example, SecureBio's BioTIER evaluates whether models refuse high-consequence biological requests without unnecessarily rejecting benign scientific inquiries ([Marshall et al., 2026](#)).

These assessments are especially important for open-weight models ([Wallace et al., 2025a](#)), where they can be modified and attacked with little deployment-time protections. One challenge of evaluating open-weight models is that they can be fine-tuned to increase their capabilities later on. Accordingly, frontier AI developers can use "malicious fine-tuning" (MFT) to assess these capabilities and estimate the worst-case outcome before release ([Wallace et al., 2025b](#)). Similar techniques can also be used for eliciting the capabilities of proprietary models ([Che et al., 2025](#)), which can be made substantially more capable when provided with greater inference compute ([Wei et al., 2025](#)).

Limitations

1. **Cost.** Some evaluations, like running models on existing refusal benchmarks, are relatively cheap. However, more ambitious evaluations should explore the ceiling of model capabilities: that is, the longest and hardest real-world harmful projects that the frontier model can complete for a malicious actor with realistic resources.

Evaluating this may require large token budgets, multiple repetitions lasting multiple days, expensive human baselines or expert grading, and more advanced and secure sandbox environments (see pricing in [METR, 2026](#)). These costs are likely to increase in the future as increasingly powerful models require more test-time compute to be fully elicited. Real-world uplift studies requiring more human participants, such as wet-lab trials, may also be more expensive and time-consuming to run.

2. **Lack of evaluation integrity.** As models become more capable and situationally aware, they may behave differently depending on who they think is instructing them ([Zhong et al., 2026](#)), including to strategically underperform on dangerous capability evaluations ([van der Weij et al., 2024](#)). Without sufficient countermeasures, such as evaluation awareness steering ([Anthropic, 2025b](#)) or fine-tuning on dangerous capabilities ([Kaunismaa et al., 2026](#)), evaluation results in the future may not reliably reflect true model capabilities in all circumstances.
3. **Lack of international alignment on threat vectors.** Without shared understanding within and across nation-states on which misuse vectors warrant priority evaluation and safeguards, inconsistent standards enable “model shopping,” where malicious actors gravitate toward models or jurisdictions with weaker oversight.
4. **Limited transparency.** Evaluation methodologies and results are often opaque to reduce potential infohazard risks, but this makes it difficult to independently assess the quality of the evaluations and verify claims about model safety or compare evaluations across companies ([Peppin et al., 2025](#); [Ho and Berg, 2025](#)).
5. **Absence of whole-chain risk analysis.** Most bio risk evaluations focus narrowly on specific stages of the bioweapon development pathway rather than assessing how LLMs interact with the complex, multi-stage process of developing and deploying a harmful biological agent. This fragmented approach may miss critical bottlenecks or fail to identify where AI assistance could prove most consequential.

Bottlenecks to Implementation

1. **Lack of high-quality evaluations and evaluators accessible to international developers.** Most third-party AI evaluators with CBRN expertise are concentrated in US and UK organizations or government research institutes (e.g., UK AISI, CAISI, SecureBio), making them inaccessible to developers in other jurisdictions. Ideally, all major AI jurisdictions would have domestic capacity to evaluate these risks for their developers and understand how to best elicit capabilities for their models.
2. **Cost and expertise barriers for developers.** Comprehensive biorisk evaluations (i.e., uplift trials, red-teaming, wet-lab studies) are resource-intensive and require specialized expertise beyond the reach of most developers. Absent external support or accessible tooling, smaller companies face significant barriers to conducting rigorous pre-deployment assessments.
3. **Limited threat intelligence outside governments.** Governments possess differentiated knowledge regarding threat actor capabilities, plans, and intentions. This information asymmetry complicates the design of evaluations appropriately calibrated to actual threats, limits the ability of non-governmental evaluators to prioritize effectively, and makes it challenging to share information between nations.

4. **Uneven regulatory coverage.** Despite emerging consensus on the necessity of evaluating CBRN risks prior to deployment ([CAC, 2025](#); [California SB 53, 2025](#)), absent regulatory mandates or company leadership buy-in, companies lack compelling reasons to prioritize these evaluations over competing development priorities.

Future Work Addressing Bottlenecks

1. **Build distributed evaluation capacity.** Governments in major AI jurisdictions (e.g., EU, China, Singapore) should fund domestic evaluation hubs and support partnerships between established evaluators (e.g., UK AISI and SecureBio) and institutions in underserved regions. The European Commission's AI Office provides a notable example, working with SecureBio, FAR.AI, and other organizations to develop CBRN risk assessments to support its EU AI Act implementation ([SecureBio, 2026](#)).
2. **Advance international alignment on evaluation and safeguard standards.** Governments should harmonize taxonomies of threat vectors and evaluation methodologies through bilateral or multilateral technical dialogues and working groups. Companies and expert groups should seek to improve the comparability of safeguard performance across companies and models to incentivize improvement of safeguards over time ([Sudarshan and Righetti, 2026](#)).
3. **Develop shared tooling and subsidized resources.** Lower barriers for resource-constrained developers by offering: open-source or securely shared evaluation frameworks and standardized benchmarks such as BioTIER ([Marshall et al., 2026](#)); automated red-teaming tools that reduce reliance on scarce expertise; publicly-funded evaluation services; and tiered protocols that reserve expensive wet-lab studies for frontier models.

Notes

2. Since OpenAI was one of the first companies to pioneer RLHF, it is possible that the number of models trained via RLHF is higher but unreported ([Christiano et al., 2017](#); [Ouyang et al., 2022](#)).
3. These techniques are also used to instill other behaviors beside harmlessness, such as helpfulness or capable reasoning. For example, Meta trained Llama 2 and 3.1-8B with RLHF ([Meta, 2023](#); [Meta, 2024](#)), and Deepseek reported training R1 in a manner similar to RLAIIF ([Deepseek, 2026](#)).
4. There are many variations of RLAIIF, and this summary aims to be illustrative rather than exhaustive.
5. This behavior could be incentivized by the original model provider. For example, a company might stipulate in its licence that all inference providers that served a model needed to ensure some sort of safeguard, such as biocapability classifiers or an API with robustness similar to the main company API.
6. According to OpenAI, this process includes using a combination of automated technologies (e.g., classifiers, reasoning models, hash-matching, blocklists) and other automated systems ([OpenAI, 2026](#)).
7. There is also a strong precedent for sharing models with trusted third parties for safety testing, either directly (e.g., [Homewood et al., 2025](#)) or through secure APIs (e.g., [OpenAI, 2023](#)).
8. However, training models to change their chain of thought without changing the underlying behavior can also harm monitorability. See [Baker et al., 2025](#).

The logo features a stylized white asterisk-like symbol to the left of the word "SAIF" in a bold, white, sans-serif typeface. The background is a dark teal color with several overlapping, semi-transparent geometric shapes in lighter shades of teal and grey, creating a layered, abstract effect.

*** SAIF**