SAIF

当前格局与未来方向

# 防范人工智能极端滥用的技术防护措施

Isabella Duan[1*], Edward Kembery[1*], Stephen Casper[2], Zhikai Chen[3], Jasper Götting[4], Geng Hong[5], Zaheed Kara[6], Lijun Li[7], Xiaojian Li[8,9], Kellin Pelrine[10], Ziyue Wang[11], Jia Xu[7], Ziwei Xu[10], Zhiyuan Zeng[8], James Zhang[12], Jie Zhang[7]

译者 : Zilan Qian [13]

[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
[13] Oxford China Policy Lab

* 为共同第一作者

# 执行摘要

随着人工智能系统能力持续增强，非国家行为体滥用该技术制造极端威胁——尤其是化学、生物、放射性与核（CBRN）武器及大规模网络攻击——的潜在风险，已引起各国政府、企业界、学术界和公众的高度警觉。[1]

本报告从三个层面梳理了目前五类主要防护实践：两个模型层面的干预措施——预训练数据过滤和后训练技术；两个部署层面的干预措施——监测和访问控制；以及一个治理层面的干预措施——部署前评估。部署前评估可以为部署决策提供依据，包括模型发布条件和防护措施设计，并支持模型部署之后对相关措施进行持续改进。

对于每一类干预措施，我们都会介绍当前实践、前沿模型开发者的实施情况、实证证据、局限性，以及未来研究与协作的优先方向。

我们的主要发现如下。

**防护托管式部署（包括专有模型和通过 OpenRouter 等第三方平台提供服务的开放权重模型）最有希望的措施是使用监测、访问控制和后训练干预。**因为服务提供者仍然掌握推理基础设施，所以使用监测是目前最强有力的防御手段：它既可以实时发现和阻止有害活动，也可以支持事后的调查和执法处置，包括暂停相关账户。同样，客户身份识别（KYC，Know Your Customer）等访问控制可以在滥用发生前降低风险，它能够决定谁可以访问一个人工智能系统、用户能够调用哪些能力，以及用户是否能够修改或移除系统的防护措施。后训练防护措施，例如基于人类反馈或人工智能反馈的强化学习，也可以提高模型拒绝有害请求的概率。不过，现有证据表明，在持续、专业的越狱攻击面前，这些措施仍然不够充分。如果将这些措施结合使用，就能够显著提高滥用的难度，并加快对新型越狱攻击的响应速度。

**对于本地或私有部署的开放权重模型，最有希望的防护措施是预训练数据过滤和后训练干预，但二者都有根本性的局限。**预训练数据过滤可以降低模型完成需要专门有害知识的任务的能力，而且与后训练防护相比，它可能更能抵御对抗性微调。然而，若模型的危险能力源于广泛存在的知识（例如编程知识中本就包含黑客能力）则很难在保留其通用性能的同时，仅凭数据过滤便有效抑制这些能力。此外，研究已经表明，部署后往往可以通过微调或模型再学习恢复危险能力。同样，只要行为者能够直接访问模型权重，也可以通过微调等技术轻易绕过后训练防护。总体而言，上述防护措施仍应成为标准实践，因为它们可以提高滥用开放权重模型的成本。然而，它们极易被绕过或移除，这意味着在对抗蓄意或经验丰富的行为体时，这些措施所能提供的保护相当有限。

[1] 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，其中包含核生化导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、使用核生化导武器能力，导致现有管控体系失效，加剧世界各地区和平安全威胁……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”。

**无论模型类型和部署方式，针对危险能力和防护措施鲁棒性的部署前评估，都是负责任发布决策不可或缺的依据。**无论专有还是开放权重，模型本身都并非天然危险。风险取决于模型能力，以及围绕这些能力设置的防护措施是否足够稳健。前沿模型获得越来越强大的双重用途能力，并可能造成现实世界中的伤害。对这些危险能力进行测量、采用基线防护措施并严格评估其有效性，应逐渐成为全球规范。这一做法可为部署决策提供信息支撑，并随时间推移不断增强防护能力。

图 1：下表是依据公开信息，对不同防护措施的采用情况、有效性、成熟度和估算成本进行的大致高层概览。需要注意的是，AI 企业内部采用的措施可能并未对外披露。

| 措施 | 主要开发者的采用情况 | 有效性 | 成熟度 | 估算成本 |
| --- | --- | --- | --- | --- |
| 预训练数据过滤 | 公开信息有限。OpenAI 报告使用了针对 CBRN 的过滤。Anthropic 已发表相关研究，但未确认在生产中使用。中国开发者根据监管要求进行一般安全过滤，但并非专门针对 CBRN。 | **中等。** 可减少模型接触危险知识，不过模型仍可能通过上下文学习和互联网访问获得这些知识。 | **低。** 最近的概念验证研究已证明其可行性，但生产环境中的部署仍有限。 | **低。** 通常只占预训练预算中可以忽略不计的一小部分，但需要开发领域专用分类器。 |
| 后训练技术 | 广泛采用。Anthropic、OpenAI、Google DeepMind、智谱、阿里巴巴等都报告使用人类或人工智能反馈强化学习。 | **中等。** 能有效引导模型行为，但仍容易受到专业越狱和恶意微调攻击。不过，它可以显著增加滥用成本。 | **高。** 已在业界广泛部署。 | **中等。** 通常只需要预训练所用数据和算力的一小部分，但提升鲁棒性仍需要持续研发。 |
| 访问控制 | 采用不均衡。**Anthropic** 和 **OpenAI** 会追踪使用情况并执行政策。**Meta** 和 **OpenAI** 会在发布权重前评估风险。**Hugging Face / ModelScope** 提供有限的访问门控。 | **中等。**客户身份验证、使用监测和账户暂停可以减轻托管部署中的滥用，但高水平行为者仍可能绕过这些控制。 | **高。** 对专有模型而言已属标准做法；对开放权重模型的应用仍处于早期阶段。 | **低。** 主要为运营和行政管理成本。 |

| 措施 | 主要开发者的采用情况 | 有效性 | 成熟度 | 估算成本 |
| --- | --- | --- | --- | --- |
| 监测与响应 | **专有模型中广泛采用。** Anthropic 使用宪法分类器（Constitutional Classifiers）；OpenAI 监测 API 和聊天。中国开发者根据监管要求进行监测，但主要针对违法内容，而非 CBRN。 | **高**。是目前托管部署中最强的滥用防护措施，不过复杂越狱仍然存在。 | **高**。已在主要实验室投入生产；近期进展降低了成本。 | **中低。** 高级分类器可能较昂贵，但探针等方式的推理成本较低。某些实时分类会增加延迟。 |
| 部署前评估 | **采用不均衡。** Anthropic、OpenAI 和 Google DeepMind 开展 CBRN 评估；英国人工智能安全中心（AISI） 已评估 30 多个系统。中国开发者报告通用基准，但没有公开 CBRN 专项评估。 | **高。**有助于做出充分知情的部署决策，并合理校准防护措施。 | **中等。** 已有成熟的评估实践，但尚未在行业内均匀普及。 | **中等。** 真实世界能力增益试验成本较高；廉价自动化测试可能缺乏效度。防护评估则可以较低成本地发现弱点并改善保护。 |

# 引言

随着人工智能系统能力持续提升，非国家行为者实施极端滥用的潜在风险——尤其是涉及化学、生物、放射性与核 (CBRN) 威胁以及网络威胁的风险——已成为全球研究人员和开发者共同关注的紧迫议题。虽然在开发缓解这些风险的技术防护措施方面已取得一些进展，但各组织和司法管辖区的实施实践并不一致；对于开发者应实施哪些防护措施，以及如何在纵深防御体系中有效分层部署这些措施，各方共识仍然有限。

本报告概述了目前用于应对极端滥用风险的常见模型层和部署层防护措施，并参考了近期研究成果以及主要人工智能公司公开披露的实践。通过描绘当前防护格局、提炼不同实践的共性，并指出实施瓶颈与未决问题，我们希望为各方搭建共享的技术基础，同时明确最需开展跨境协调的关键领域。我们希望本报告能够成为推动国际社会就采用基线技术防护措施、防范极端滥用风险形成共识的起点。

本报告主要面向基础模型开发者，包括模型开发公司、应用和服务提供商，同时也面向希望理解这些防护措施及其实施瓶颈的学术研究人员和政策制定者。虽然我们的重点是极端滥用风险，但这些防护措施同样可以用于其他类别的滥用。

# 模型层

## 1. 预训练数据过滤

### 概述

预训练数据过滤可以减少模型在训练过程中接触危险知识的机会，并降低模型对这些知识的掌握程度。虽然这项技术尚未被大规模、充分地测试，但目前来看，它仍是降低开放权重模型部署后极端CBRN滥用风险的最有潜力的缓解措施之一。

### 现有实践

从预训练语料库中移除有害数据，被普遍认为是训练更安全模型的一种关键手段（例如 Longpre et al., 2025; Casper et al., 2025）；但专门针对 CBRN 知识开展过滤，目前仍属于新兴实践。OpenAI 报告称，无论是开放权重模型还是专有模型 GPT-4o，都使用了预训练数据过滤来移除有害数据，尤其是危险的 CBRN 知识（OpenAI, 2024; Wallace et al., 2025）。Anthropic 已公开披露针对 CBRN 风险开展的预训练数据过滤实验（Chen et al., 2025），并发表了与此相关的模块化预训练研究（Roland et al., 2026），但尚未报告出于安全目的在生产环境中实际采用这些方法。Google DeepMind 和 IBM 也报告称在预训练数据中进行了安全过滤，但没有明确说明是否专门过滤 CBRN 风险相关数据（Google DeepMind, 2025, IBM, 2024）。

中国开发者受到监管要求约束，需要针对规定的一系列风险类别过滤预训练数据（TC260, 2023），但其中并没有明确针对 CBRN 风险。阿里巴巴、DeepSeek、腾讯和智谱均报告称，会出于安全考虑对预训练数据进行过滤，但同样没有具体提到 CBRN 风险（Qwen Team, 2025; DeepSeek Transparency Report, 2025; Tencent Hunyuan Team, 2024; Zhipu and Tsinghua, 2024）。

## 实施方式

Anthropic的研究 (Chen et al., 2025) 详细介绍了一套以CBRN为核心的过滤流程。预训练语料库中的文档通过训练好的分类器进行有害性评分，超过可调阈值的文档将被移除。研究人员评估了六种分类方法，包括宪法分类器 (constitutional classifiers)、保留损失评分 (holdout loss scoring)，二元词组词袋分类器，以及命名实体方法 (named-entity approaches)。其中，基于提示词的宪法分类器 (prompted constitutional classifiers) 达到了最高精度。分类器的训练数据通过大语言模型提示合成生成，避免了对敏感内容的人工标注。在中等规模的模型上，过滤使WMDP基准测试上的有害CBRN知识相对于随机基线降低了33%，且通用能力 (MMLU、代码、文本生成) 无显著下降。

与此同时，O'Brien等 (2025) 展示了另一种互补的多阶段流程，即首先进行关键词黑名单筛选，再将可疑内容升级交由分类器判断。实施这一流程所需成本约占总预训练预算的 0.8%。研究人员还发现，预训练数据过滤不仅能够提高模型开箱即用时的安全性，也可以显著增强模型抵御篡改的能力，而这一点对于降低开放权重模型的滥用风险尤为重要。他们经过过滤的 69 亿参数模型能够抵抗最多 10,000 个训练步骤和 3 亿 token 的对抗性微调，与后训练基线相比，鲁棒性提升了一个数量级。对于可能无意中恢复被压制能力的“良性微调”攻击，模型同样具有较强抵抗力。

## 局限性

1. **规模扩展的挑战。**更大的模型使用更多数据训练，即使采用安全缓解措施，也可能只需要更少的示例就能提取危险知识 (Souly et al., 2025; Bowen et al., 2025)。因此，要达到相同程度的能力压制效果，就必须使过滤系统具有更高的召回率。

2. **适用范围有限。**过滤对缓解CBRN滥用高度有效，因为CBRN的滥用依赖于特定有害领域的专业知识 (如生化武器合成) 。但对于网络攻击、影响力行动或生成有害言论，如果相关知识广泛存在于普通语料中，过滤所能提供的保护可能会更加有限。在毒性降低等其他领域，一些研究甚至发现，使用更多有毒数据进行预训练，可能会使后训练阶段更容易控制模型，并最终使整体毒性下降 (Li et al., 2025)。

3. **难以界定什么属于危险知识。**要构建能够区分有害知识和合法双重用途知识的分类器，仍然非常困难。过度激进的过滤可能损害模型在有益科学应用中的表现。相比之下，有针对性地删除最高风险内容（例如获得活体高风险病原体的具体操作流程）可能能够形成更可接受的权衡。

4. **因上下文学习而失效。**模型可以利用上下文中直接提供的有害信息，或者在恶意微调过程中重新学习有害知识。因此，有必要采取分层防御，将过滤与其他防护措施结合使用（O'Brien et al., 2025)。不过，模型对于最敏感的一类信息（例如大规模杀伤性武器的具体技术规格）的学习效果可能没有那么突出。

## 实施瓶颈

1. 缺乏研究和评估基础设施。验证过滤器是否有效，需要具备红队测试能力和特定领域的评估数据集，而较小的组织通常无法自行建立这些资源。构建和测试过滤器还必须运行预训练实验，这意味着大量算力成本，学术研究人员往往很难承担。各方还需要进一步研究，以形成充分的实证证据，证明过滤不会损害模型的一般性能，并验证这些方法能否在大规模训练中切实实施。

2. 企业和国际层面缺乏领域专业知识。AI 公司通常缺乏像SecureBio等组织所具备的独立识别危险 CBRN 知识、构建专门过滤器所需的领域专业能力。此外，如果想要传播现有生产过滤器或构建过滤器所需的信息，还可能受到危险信息转移相关法规的限制。

3. **语言障碍。**目前大多数工具都是针对英文语料设计和评估的，因此非英语开发者缺少经过验证的、适用于自身语言的方法，也缺乏衡量过滤性能的标准基准。因此，互联网规模的数据集需要具备稳健的多语言过滤能力。

### 应对瓶颈的未来工作

1. **建立关于研究与治理的基础设施。**政府或行业标准机构应针对CBRN、儿童性虐待材料 (CSAM) 及其他危险数据过滤制定分类排除清单；政府应资助专项研究中心 (如，人工智能+生物风险研究组)。这些中心应同时具备人工智能与威胁领域的专业知识，能够开发并验证针对CBRN的专用过滤器，提供可供小型开发者直接使用的现成工具和评估基准，并通过与行业合作，来确保过滤器的可用性以及与生产流程的兼容性。

2. 标**准化过滤类别。**专家组可以制定过滤类别的分类体系，重点针对风险最高的知识，同时保留有益用途。例如 BioTIER 建议，只排除定义狭窄的高风险生物医学知识及相关操作主题，同时保留双重用途和无害的生物学内容（Marshall et al., 2026）。进一步完善这类框架并形成共识，或会加速预训练过滤技术的采用。

3. 衡**量有效性与权衡。**未来研究应建立标准化基准，衡量过滤器能够在多大程度上减少有害行为，同时保留模型的一般能力，以及其在高风险知识子集相关专业领域中的表现。

4. 改**进现有过滤方法并探索替代方案。**例如，不同于试图进行完美分类或预先排除危险数据，选择性梯度掩码 (Selective Gradient Masking) 在训练期间将危险知识定位到专用模型参数中，然后可以将其移除，同时保留通用能力 (Shilov et al., 2025)。通过使用梯度路由 (gradient routing) 来定位特定知识，可能进一步提升该技术的有效性 (Cloud et al., 2024)。

5. 验**证跨语言迁移效果。**未来研究需要验证过滤器对于使用非英语数据训练的模型是否同样有效。

6. 探**索移除有害能力的方法。**预训练数据过滤的替代方案之一是对模型进行遗忘学习，再对遗忘后的模型进行蒸馏。在 1 亿参数模型上的测试表明，这种方法可实现与预训练过滤相当的抵抗有害行为的效果，且可能使用更少的计算量和标注数据 (Lee et al., 2025)。

## 2. 后训练技术

### 概述

目前有两种主要的后训练技术用于教导模型以适当方式处理有害请求：基于人类反馈的强化学习 (RLHF) 和基于人工智能反馈的强化学习 (RLAIF) (Christiano, 2017; Ouyang et al., 2022; DeepMind, 2022; Bai et al., 2022; Guan et al., 2024)。

这类后训练的针对性安全效果包括：使模型的偏好转向更安全的回答；抑制有害能力或相应的内部表征；提高模型安全行为在面对对抗性诱导或后续修改时的鲁棒性。总体而言，两种方法都已经证明能够有效引导模型适当地处理有害请求，但也存在一些局限。未来针对 RLAIF 最佳实践的研究，可能发现更加有效的机制。

### 现有实践

多家实验室报告使用后训练技术缓解人工智能模型的有害回答。Anthropic 报告使用 RLHF 和 RLAIF 训练其 Claude 模型以无害方式响应请求 (Anthropic, 2025a; Anthropic, 2025b)。

OpenAI 自 o1 模型起使用 RLAIF 训练其最先进的模型 (OpenAI, 2024; OpenAI, 2025b; OpenAI, 2025c; OpenAI, 2025d)，使其遵循以“防止伤害”为目标的模型规范（Model Spec）(Guan et al., 2024; OpenAI, 2025a)。OpenAI同时也报告通过 RLHF 训练了 ChatGPT 4.5 (OpenAI, 2025)。[2] Google DeepMind 通过 RLAIF 训练 Gemini 2.5 以符合"安全政策" (Google DeepMind, 2025b)。智谱将 RLHF 用于 ChatGLM (Zhipu, 2024), GLM-4.5 (Zhipu, 2025) 和 GLM-4.5V (Zhipu, 2026)。阿里巴巴报告使用 RLHF 和 RLAIF 等技术提升 Qwen 2.5 的 “无害性 (harmlessness)” (Yang et al., 2024)。[3]

## 实施方式

基于反馈的强化学习一般分为两个阶段。RLHF 与 RLAIF 的两种常见形式——**宪法式 AI（Constitutional AI，CAI）和审慎式对齐（Deliberative Alignment，DA）**——在具体操作上略有不同。[4]

1. **监督学习阶段**
    - **RLHF**: 对预训练模型进行微调，使用人类标注者生成的高质量示范数据。
    - **DA**：通过在系统提示中插入安全原则，促使模型在回应有害查询时对这些原则进行推理，从而生成微调数据 (Guan et al., 2024)。
    - **CAI**：首先让一个只针对“有帮助性”训练的模型回答有害提示；然后依据一套“宪法原则”批评和修改其回答，再用修改后的输出进行微调 (Bai et al., 2022)。
2. 强**化学习阶段**
    - **RLHF**：人类评分员对模型输出进行成对比较，产生的偏好数据用于训练奖励模型。
    - **DA**：奖励模型本身也是一个 AI 系统，其提示词中包含安全规范，并且在评估时可以直接访问这些规范 (Guan et al., 2024)。
    - **CAI**: 经过微调的模型针对有害提示生成成对回答，再通过多项选择题等形式，评估这些回答在多大程度上遵循“宪法原则”，最后利用这些判断来训练奖励模型 (Bai et al., 2022)。

## 局限性

1. **专有模型仍易遭受越狱攻击。**英国 AISI 的公开研究及内部报告均表明，即使经过微调以拒绝有害请求的模型仍可被专业攻击者越狱 (UK AISI, 2025)。英国 AISI 最近还发布了自动化越狱的证据，显示这种方法甚至能够绕过业内最强的一些防护措施 (Davies et al., 2026)。

2. **开放模型仍容易受到微调攻击**。对模型进行微调可轻易移除其防护措施 (Qi et al., 2024, Murphy et al., 2025)。一些声称能够让模型更能抵御对抗性微调的训练方法 (如 Tamirisa et al., 2024)，被证明无法抵御廉价、现实的攻击方式，例如对抗性微调和边界点越狱 (Kuo et al., 2026; Zloczower et al., 2026; Davies et al., 2026)。即使模型变得更具鲁棒性，恶意行为者仍可能通过重新利用深度强化学习算法找到特定的方式绕过防护措施 (Willison, 2023, Liu et al., 2023, Rao et al., 2023, Wei et al., 2023, Liu et al., 2025)。

3. **不恰当的训练可能导致过度拒绝**。后训练可能导致模型拒绝无害请求，尤其是那些与有害问题在词汇或概念上相似的请求。这一问题在生物学、化学和网络安全领域尤其突出，因为合法用途和恶意用途利用的知识往往有大量的重叠。

4. **强化学习的泛化效果在不同部署场景下并不均衡**。研究表明，研究表明，基于 RLHF 的拒绝训练在以下情境中效果欠佳：低资源语言 (Yong et al., 2023)，极长上下文 (Anil et al., 2024)，以及智能体或工具使用场景，特别是当一个有害行为由多个表面无害的步骤组合而成时 (Glukhov et al., 2024)。

### 应对瓶颈与局限性的未来工作

1. **深化RLAIF研究。**RLAIF仍是一项新兴技术，尚无有关最佳实践的共识，且该技术可能无法保证持续赋予模型理想的无害行为 (Zhang et al., 2025; Schoen et al., 2025)。构建更好的规范遵守评估体系可使开发者更准确地比较不同后训练流程的效果 (Ahmed et al., 2025)。

2. **构建防篡改的开放权重模型。**目前的研究未能找到使开放权重模型更能抵抗微调攻击的有效训练方法 (Huang et al., 2024; Sheshadri et al., 2025; Tamirisa et al., 2025)。不过，如果模型在训练期间针对更广泛的篡改威胁接受训练，仍有可能形成泛化性更强的抗篡改能力 (Casper et al., 2025)。

3. **训练模型对危险话题产生困惑。**即使模型不具备危险能力，恶意行为者也可能通过微调来获取这些能力 (Kaunismaa et al., 2026)。在这种情况下，一种缓解风险的方式是使模型对危险话题形成困惑或错误认知 (Wang et al, 2025; Slocum et al., 2025)。但此课题需要进一步研究。

4. **构建更好的开源安全微调数据集。**开发开放源代码或能够安全共享的数据集，使模型在不损害一般能力的情况下提高安全性，可以降低构建安全模型的成本 (Wang et al., 2025)。

# 部署层防护措施

## 3. 访问控制

### 概述

缓解滥用的一种方法是限制恶意行为者访问人工智能系统或以规避安全防护措施的方式修改它们的能力。

对于闭源模型，开发者可以对最终用户实施“客户身份验证 (Know Your Customer, KYC)”要求，并撤销对有滥用模型历史记录的用户的访问权限。

不过，对于开放权重模型，有效阻止恶意行为者获取模型仍是一项重大挑战。主要存在以下干预点。首先，开放权重模型托管平台 (例如 Hugging Face、ModelScope，或模型开发者自身) 可以拒绝分发已被越狱的开放权重模型，或在允许下载具有重大滥用潜力的系统前要求完成 KYC 验证。其次，人工智能服务和应用提供商基于开放权重模型构建产品时，可以实施与闭源模型相似的访问控制。[5]

### 现有实践

基础模型开发者使用了多种技术来控制用户访问他们的模型。例如，Anthropic和OpenAI都跟踪聊天和API交互，以强制执行其闭源模型的使用政策 (Anthropic, 2025; OpenAI, 2025)。[6] 这使他们能够选择性地监控、阻止或禁止可能违反用户政策条款的用户。OpenAI还提供“组织验证 (organization verification)”，并向通过政府签发的身份证明认证的选定用户提供“可信访问 (trusted access)”模式 (OpenAI, 2025b; OpenAI, 2025c; Clymer et al., 2025)。[7]

一些前沿模型开发者表示，只有当确信模型不会增加恶意行为者滥用的风险时，他们才会公开发布模型权重。例如，此前测试表明，OpenAI发布的GPT-oss模型即使经过微调也不具备“高风险”能力 (Wallace et al., 2025b)。Meta 承诺，在开源风险达到一定水平的模型前，会实施并验证风险缓解技术 (Meta, 2025)。

模型托管平台也在一定程度上控制模型访问。模型作者可以要求用户在访问模型文件前提供联系信息，并有权拒绝未获批准的访问请求 (HuggingFace, ModelScope)。托管提供商也可以拒绝托管违反其内容政策或侵犯版权的模型 (HuggingFace)。

## 实施方式

访问控制的具体实施方式取决于服务提供方的性质（例如基础模型开发者还是模型分发平台）以及访问控制的目的。一些示例策略包括：

- **对于开发者：** 提供聊天界面或 API；只向特定注册用户提供新模型访问权限或更高的词元数量；监测并封禁违反使用政策的用户。
- **对于分发方：** 在用户未提供特定身份证明时限制其访问模型；移除违反托管政策的模型。

在每个阶段，开发者和分发方都可以根据风险评估选择不同程度的用户身份验证，包括：

- **基本信息：** 创建账户、拥有注册邮箱、提供电话号码、点击接受可接受使用条款。
- **中等程度的用途说明：** 提交申请、说明所属机构、解释使用目的。
- **严格身份认证：** 提供个人身份信息（例如政府签发的身份证件）、通过验证用户与机构的隶属关系确认其真实性、与政府观察名单进行比对。

## 局限性

1. **即便是受到良好监控的闭源API，也可能无法发现部分滥用行为。**例如，来自模型开发者的一份报告称，网络犯罪分子曾使用其模型开发、营销和分发多个勒索软件变种 (如Anthropic, 2025b)。尽管这些账户最终被封禁，但这一事件表明，动机足够强烈的行为者仍能在被发现之前实施实质性的滥用。

2. **开放权重模型的访问权限难以撤销。**当前发布模型权重的实践往往依赖于对该模型风险的评估 (Wallace et al., 2025b; Meta, 2025)。然而，一个在某一时间点考虑微调后仍被认为安全的模型，未来可能被恶意行为者通过微调诱导出更危险的能力。例如，可以利用未来更强大的模型产生的合成数据进行微调 (Kaunismaa et al., 2026)。

3. **模型分发者无法阻止在其平台上下载的模型被分享到别处。**虽然模型分发者有一些机制可以影响哪些模型在其平台上被访问，但他们无法阻止在其平台上下载的模型被分享到别处，也无法获取这些模型实际使用方式的信息。

4. **限制访问会带来 AI 应用潜在收益方面的机会成本。** 允许人们访问模型，尤其是开放权重模型，可能为那些使用模型修改版本的人，或者希望为了科学或安全目的开展模型实验的人带来收益 (Seger et al., 2023)。

5. **模型仍然可能遭到泄露或窃取。** 限制公众对模型资产的访问，并不能防止有人从数据中心窃取模型或发生内部泄露。不过，这仍然能够大幅减少可以访问模型的人数，因为真正有能力直接窃取模型的人很少，而且还可能通过内部的访问控制来降低模型泄露带来的风险 (Nevo et al., 2024)。

## 应对瓶颈和局限性的未来工作

1. **更好的使用监控。**能否在专有模型的聊天或 API 接口上有选择地封禁或阻止用户，取决于平台能否在违反使用政策的行为发生时可靠地将其标记出来。未来研究可以寻找用于发现特定威胁路径的机制，例如多智能体网络间谍活动、CBRN 能力提升，或者开发更好的通用计算机使用监测方法 (例如，Sumers et al., 2025)。

2. **有限制地部署模型**。开发者无需无限制地公开模型权重。在拆分式部署（split deployment）中，仅共享模型的部分组件，开发者仍参与推理过程 (Xie et al., 2025)。另一种方式是将权重限制在预先批准的硬件上 (Clifford et al., 2025)。虽然这些选项目前尚处于起步阶段，但进一步的研究可以探索它们是否能够成为可行的政策选择。

# 4. 监测与响应

## 概述

一旦特定群体获得访问权限，滥用监测系统就可以分析模型交互以及更广泛的使用模式，以识别潜在有害活动。系统会标记可疑模式（例如反复询问如何合成危险化合物），并实时或通过事后日志分析触发适当响应，例如降低用户所使用的模型等级、限制响应速率、撤销访问权限，或者通知有关部门。

对于专有模型，开发者控制着推理基础设施，因此能够有效实施监测，这使它成为目前最强的滥用防护措施。

对于开放权重模型，能否进行监测取决于部署方式：

- **托管式部署（AI 服务和应用提供商）：** 下游部署方可以实施与专有模型类似的监测。
- **自行托管部署（本地或云端）：** 如果平台或用户不予合作，目前无法进行监测。

云计算监测或许是一种颇具前景的方案，即由 AWS、阿里云等服务提供商自动对托管在其平台上的模型运行极端滥用监测 (Heim et al., 2024)。不过，即使只是在这一层进行推理级监测，也会引发隐私和法律责任方面的实际问题，因为云服务商需要检查客户在其平台上的输入和输出。像检测恶意微调和脚手架式增强这样更具雄心的目标还会面临额外的技术挑战，即如何区分恶意工作负载和合法工作负载 (Sun et al., 2026)。

## 现有实践

世界各地的模型开发者已经部署输入/输出分类器，用于标记可能有害的请求，包括 CBRN 威胁。被标记的交互会接受人工复核，并可能导致账户暂停 (Anthropic, 2025; OpenAI, 2025)。例如，Anthropic 使用实时分类器监测模型输入和输出，拦截威胁性内容的生产，并借助离线分类器识别越狱攻击 (Anthropic, 2025)。他们通过漏洞赏金 (bug bounty) 和其他评估对分类器的有效性进行红队测试，开发了一个针对新越狱方法的快速响应系统。

中国监管框架要求 AI 服务提供商建立内容监测系统，并强制要求将违法内容报告给有关部门 (中央网信办, 2023)。中国大型企业都运行内容审核系统，不过这些系统主要针对违法内容，而不是专门针对 CBRN 或高级网络风险。例如，DeepSeek 使用实时关键词匹配来标记有问题的查询，随后根据预设的风险提示对这些查询进行复核，以判断是否撤回相关对话 (Guo et al., 2025)。阿里云提供人工智能防护墙作为服务，用于人工智能应用提供商的内容合规检测 (Alibaba Cloud, 2025)。

## 实施方式

监控系统采用不同的架构，具有不同的成本-安全性权衡：

在线与离线监控。在线监测会实时检查有关查询和输出的内容，在交付前阻止有害内容。这起初带来了成本问题：Anthropic最初的宪法分类器带来了约24%的推理开销，限制了其被广泛采用 (Sharma et al., 2025)。但最近的技术已将其降低到接近2% (Cunningham et al., 2026)。离线监测则在事后进行日志分析，以查找政策违规行为。它可以在不增加延迟的情况下支持账户暂停，但无法阻止首次发生的伤害。公司通常会将两种方式结合起来，对 CBRN 风险进行实时阻断，同时针对更广泛的政策执行进行事后分析。

监控架构。监测系统可以采用多种架构。不同架构在成本和安全性之间存在不同的权衡：

- **关键词匹配：**使用预定义列表实时标记有问题的查询(Guo et al., 2025)。
- **基于大语言模型的分类器**：基于 LLM 的分类器使用独立语言模型判断输入或输出是否违反安全政策。这些分类器可以经过微调 (例如 Anthropic 的 Constitutional Classifiers 使用合成生成的有害和无害示例训练，据报告可以阻止 95% 的越狱攻击 (Sharma et al., 2025))，也可以通过提示词运行并使用输入缓存。不过，将语言模型作为分类器运行需要较高的计算成本。
- **探针 (probes)**：用于判断模型内部表征包含哪些信息的简单分类器。它们提供了更加高效的替代方案，只需 Constitutional Classifiers 约 2% 的参数量就能够达到类似性能 (Cunningham et al., 2026; Kramár et al., 2026)。
- **多阶段系统**：结合了上述方法，例如使用轻量级探针进行初步筛查，使用昂贵的分类器进行升级处理，在保持鲁棒性的同时将成本降低40倍 (Cunningham et al., 2026)。

监控范围。传统滥用监测会检查模型输入、输出，以及在能够获得时检查推理轨迹。即使推理轨迹经过加密，对其进行监测仍然很重要，因为加密并不能保证其机密性。Panfilov et al. (2026) 展示了一种“解密越狱”，即使模型面向用户的可见回答已经安全拒绝请求，也可以从专有模型的推理轨迹中恢复危险信息。[8] 针对AI智能体，监测范围可能还需覆盖其完整的执行轨迹，包括计划制定、工具调用、代码执行、记忆更新及环境效应。原因在于，有害行为可能在长周期、多步骤的工作流中渐次浮现，而任何单一操作单独看来未必具有恶意 (Williams et al., 2026; 上海人工智能实验室, 2026)。

升级机制。被标记的内容可以触发自动阻断（针对高置信度违规）、人工复核（针对含糊案例）以及日志记录。随后，这些数据可以成为持续改进的指标；例如，通过精细校准阈值，Anthropic 的系统报告假阳性率为 0.05% (Cunningham et al., 2026)。

## 局限性

1. **不适用于自托管的开放权重部署。**监控需要API中介访问。一旦模型权重被发布，开发者就失去了对下游使用的可见性，使监控对自托管的开放权重模型失效。
2. **面对对抗攻击时较为脆弱。**尽管研究人员进行了大量的红队测试，模型的防御措施仍然不完善(McKenzie et al., 2025)。例如，英国AISI最近发现了一种攻击，可以绕过Anthropic的宪法分类器以及OpenAI GPT-5.2上的保障措施 (Davies et al., 2026)。攻击还可以被拆解为多个表面无害的步骤，并分散到正常流量中，从而提高有效性(Sun et al., 2026)。
3. **双重用途带来的模糊性。**训练小型模型区分合法研究与恶意意图仍然具有挑战性，并可能导致较高的假阳性率。例如，Anthropic 最初的论文指出：“虽然宪法式分类器在 MMLU-Chemistry 问题上的假阳性率较低，为 1.50%，但我们在 GPQA-Chemistry 问题上观察到明显更高的 26.05% 假阳性率” (Sharma et al., 2025)。
4. **不同系统可能需要不同分类器。**恶意代码生成、欺诈等滥用类型，可能需要有别于化生放核（CBRN）风险分类器的专门分类器，抑或更广泛的训练数据。这将大幅增加监测系统的复杂性与成本。

5. **对手可能使用多个模型来实施伤害。**攻击者可以将一项有害工作流分散到多个模型上，使每一次请求单独看来都无害。因此，分类器可能无法识别整体意图，使能力较强的模型完成表面合法的子任务，而防护较弱的模型负责处理明确有害的步骤 (Jones, Dragan and Steinhardt, 2024)。

6. **推理成本限制。**虽然级联架构 (cascading architectures) 和表征复用 (representation reuse) 方面的近期进展已经大幅降低了分类器成本 (Cunningham et al., 2026)，但更加全面的分类器式监测在大规模运行时仍然会产生一定计算成本。

## 实施瓶颈

1. **探针的技术复杂性。**探针的开发需要大量的技术专业知识，并且每次模型权重更改时都必须重新构建，这对于资源有限的开发者来说可能是难以承受的。对于开放权重模型，学术界可以通过开发和共享探针来提供帮助。

2. **降低推理成本和延迟需要前期部署工作。**尽管多家企业已成功部署有效监测系统，且未造成难以承受的推理成本或延迟，但这些系统的建设与落地仍可能耗时甚久，并需要大量前期投资及组织协调。

3. **标准化方面的挑战。各方可能很难就分类器的具体阈值形成明确共识。**如果缺乏能被广泛采用的滥用监测标准，对手可能会进行“模型挑选”，把有害查询转向防护措施较弱的服务提供商。

4. **分类器训练数据稀缺。**有效的监控需要对抗性攻击和有害输出的数据集，这些数据共享起来很敏感，并且很难全面整理。

## 应对瓶颈的未来工作

1. **研究探针监测和其他低成本方法。**分析模型内部状态，而不是运行单独的分类器，可以大幅降低成本。相关研究目前仍处于早期阶段，需要继续在生产环境中进行验证。

2. **企业之间在关键威胁类别上的协调。**与其追求全面的标准化，不如就最高优先级的威胁类别达成初步协议。这可以减少模型选购，同时允许实验室在阈值校准方面具有灵活性。

3. **分层监控架构。**大多数主要公司已经针对高风险威胁（例如 CBRN）采用成本较高的实时分类，针对更广泛的政策违规使用成本较低的离线分析。这样公司能在成本和安全之间取得平衡。进一步研究分层系统，可以帮助企业开发更便宜、更有效的监测方案。

4. **建立用于构建有效分类器的安全共享数据集。**建立像病原体研究领域采用的生物安全框架一类的安全数据共享机制，可以加快整个行业的分类器开发。其中一种方法可以是建立由联盟管理的数据集，只向经过审查的研究人员提供受控访问。

5. **分享评估和提高分类器鲁棒性的最佳实践。**包括建立用于衡量不同威胁类别中假阳性率和假阴性率的共同基准；建立红队测试协议，以系统识别分类器弱点；以及在兼顾安全问题的同时，为新发现的越狱攻击建立披露规范。

# 治理层防护措施

## 5. 部署前评估

### 概述

部署前评估既包括危险能力评估 (用于判断模型执行危险任务或为危险任务提供实质性帮助的能力) 也包括防护措施评估 (用于判断防护措施在多大程度上能够阻止恶意行为者获取这些能力)。评估结果将支撑风险研判，进而影响模型的发布与部署决策，推动防护措施持续迭代优化 (上海人工智能实验室 and 安远AI, 2025)。从这个意义上说，部署前评估是其他所有滥用防护措施的基础。

### 现有实践

Anthropic、Google DeepMind 和 OpenAI 会评估其模型是否能够在传统手段之外进一步“提升”资源较少或中等资源水平的威胁行为者制造、获取和部署 CBRN 武器的能力 (Anthropic, 2025; Google DeepMind, 2025; OpenAI, 2025)。其他开发者，例如 DeepSeek、智谱、阿里巴巴、字节跳动和腾讯，虽已记录通用安全基准，但尚未发布化生放核（CBRN）威胁评估。

政府和第三方评估是建立问责机制以及提高滥用风险透明度的重要基石。例如，英国 AISI 已经对 30 多个 AI 系统开展生物、化学和网络安全评估 (UK AISI, 2025)。中国信通院（CAICT）则针对 15 个开放权重系统的进攻性网络能力进行了安全评估 (CAICT, 2025)。上海人工智能实验室 (2025) 和安远AI (2026) 也针对多种开放权重和专有模型进行了评估，所采用的前沿风险管理框架涵盖生物、化学和网络威胁。

### 实施方式

评估模型危险能力的方法包括：

- **自动化评估**：使用多项选择题测试领域知识，使用开放式问题探测具体瓶颈步骤，并通过智能体评估衡量多步骤任务完成能力。生物风险方面的重要基准包括 LAB-Bench（生物研究能力；见 Laurent et al., 2024）、SecureBio VCT（病毒学实验流程排错；见SecureBio, 2026），以及其他定制化生物安全评估。开发者和研究人员还可以利用 Petri 等自动化工具，低成本、快速地构建自己的定制评估环境（Fronsdal et al., 2025）。

- **专家红队测试**：第三方红队与生物防御专家会针对特定场景开展测试，涵盖细菌与病毒的威胁路径，并评估从构思、获取、扩增、配制到释放等各阶段。

- **能力提升试验 (uplift trials)**：衡量与仅使用互联网的基线相比，获得模型访问权限是否能够提高参与者完成敏感任务的表现的人类研究。例如 Anthropic's (2025) 针对高难度病毒重建方案开展了病毒学能力提升试验。Frontier Model Forum's (2025) 评估了真实世界生物能力提升的湿实验。

除了评估和诱导危险能力之外，滥用评估还包括防护措施鲁棒性评估，用于检验防护措施在对抗压力下是否仍然有效。这其中也包括针对越狱、微调和脚手架式攻击进行的红队测试 (Zou et al., 2023, Mazeika et al., 2024; Murphy et al., 2025)。由于模型已经在 CBRN 和进攻性网络安全领域具有很强能力，开展防护措施鲁棒性评估是一个优先事项，因为它们对于判断一个部署的整体风险状况至关重要。它们也可以提供有用的信号，帮助确定未来应如何改进防护措施。

近期的例子包括 FAR.AI 的 AI Security Leaderboard，该榜单测试了前沿模型在 CBRN、网络和爆炸物威胁情境下抵御越狱攻击的能力 (Timm et al., 2026)；以及安远 AI 的红队测试计划，用于测试多个风险领域中的防护鲁棒性 (安远, 2026)。除鲁棒性外，越来越多的研究也在考察防护措施能否适当地平衡两种需要：一方面阻止有害的双重用途请求，另一方面避免因过度拒绝而妨碍合法的科学和防御性研究 (OpenAI, 2026; Zhao et al., 2025)。例如，SecureBio 的 BioTIER 会评估模型能否拒绝可能造成严重后果的生物请求，同时又不会不必要地拒绝无害的科学问题 (Marshall et al., 2026)。

**这些评估对于开放权重模型尤其重要** (Wallace et al., 2025)，因为这类模型可以被修改和攻击，而部署阶段几乎没有额外保护。评估开放权重模型面临的一项挑战是，它们以后可以通过微调进一步提高能力。因此，前沿 AI 开发者可以使用“恶意微调”(malicious fine-tuning，MFT）来评估这些能力，并在发布前估算最坏情况下的结果 (Wallace et al., 2025)。类似的技术也可用于激发专有模型的能力 (Che et al., 2025)；当提供更多的推理计算时，这些模型的能力可以大幅提升 (Wei et al., 2025)。

## 局限性

1. **成本。** 一些评估相对便宜，例如让模型运行现有的拒绝基准。然而，更具雄心的评估应当探索模型能力的上限，即前沿模型能够帮助拥有现实资源的恶意行为者完成多长时间、难度多高的现实世界有害项目。评估这些能力可能需要很大的词元预算、持续多天的多次重复试验、昂贵的人类基线或专家评分，以及更加先进和安全的沙盒环境（价格参见 METR, 2026）。随着能力越来越强的模型需要更多测试时计算才能充分诱导其能力，未来这些成本可能还会增加。需要更多真人参与者的现实世界能力增益研究，例如湿实验室试验，也可能更加昂贵且耗时。

2. **缺乏评估完整性。**随着模型能力增强并具有更强的情境意识，它们可能根据自己认为是谁在向它们下达指令而采取不同表现 (Zhong et al., 2026)，其中包括在危险能力评估中策略性地表现不佳 (van der Weij et al., 2025)。如果没有足够的对抗措施，例如评估感知导向 (Anthropic, 2025) 或针对危险能力进行微调 (Kaunismaa et al., 2026)，未来的评估结果可能无法可靠反映模型的真实能力。

3. **国际社会对威胁路径缺乏一致认识。** 若各国国内及国家间无法就哪些滥用途径应优先评估与防范达成共识，标准分歧将催生“模型挑选”现象，令恶意行为者转向监管薄弱的模型或司法辖区。

4. **透明度有限。** 评估方法与结果往往出于降低潜在信息危害（infohazard）风险的考量而不予公开，但这给独立评判评估质量、核实模型安全声明以及开展跨公司评估比较带来困难 (Peppin et al., 2025; Ho and Berg, 2025)。

5. **缺少全链条风险分析。**多数生物风险评估仅狭隘地聚焦于生物武器开发路径中的特定阶段，却未评估大语言模型（LLM）如何与有害生物因子的复杂多阶段开发及部署流程产生交互。这种碎片化方法可能遗漏关键瓶颈，亦无法判明 AI 的辅助在哪些环节可能产生最深远影响。

## 实施瓶颈

1. **国际开发者难以获得高质量的评估和评估人员。**目前大多数具备 CBRN 专业知识的第三方 AI 评估机构都集中在美国和英国的组织或政府研究机构中，例如 UK AISI、CAISI 和 SecureBio，这使其他司法辖区的开发者难以获得这些资源。理想情况下，各主要AI司法辖区均应建立本土化的风险评估能力，并掌握最有效的方式，以引导本地区开发者充分释放其模型的能力。

2. **开发者面临成本与专业知识的障碍。**全面的生物风险评估 (例如能力提升试验、红队测试和湿实验室研究) 不仅需要大量资源投入，还依赖于超出多数开发者自身能力范围的专门知识。如果没有外部支持或易得的工具，小型公司在开展严格部署前评估时会面临重大障碍。

3. **非政府机构掌握的威胁情报有限。**政府机构往往掌握关于威胁行为者能力、意图及行动计划的差异化信息。这种信息不对称既增加了设计出与真实威胁相契合的评估方法的难度，也制约了非政府评估者有效确定优先事项的能力，同时还使各国之间的情报共享更加困难。

4. **监管覆盖不均。**尽管各方在部署前评估CBRN风险的必要性上正在形成共识 (CAC, 2025; California SB 53, 2025)，但如果缺乏监管强制要求或公司领导层的支持，企业就缺乏足够有力的理由，在其他相互竞争的研发优先事项之上优先开展这些评估。

## 应对瓶颈的未来工作

1. **建立分布式的评估能力。**主要 AI 司法辖区的政府（例如欧盟、中国和新加坡）应资助本国评估中心，并支持成熟评估机构（例如 UK AISI 和 SecureBio）与服务不足地区的机构建立合作关系。欧盟委员会的 AI Office 是一个值得关注的例子：它正与 SecureBio、FAR.AI 等组织合作开发 CBRN 风险评估，以支持《欧盟人工智能法案》的实施(SecureBio, 2026)。

2. **推进评估与防护标准的国际协调。**政府应通过双边或多边技术对话和工作组，对威胁路径分类体系和评估方法进行协调。企业和专家组织应当努力提高不同公司和模型之间防护性能的可比性，从而形成激励，推动防护措施不断改善 (Sudarshan and Righetti, 2026)。

3. **开发共享工具并提供补贴资源。**通过以下方式降低资源有限开发者的门槛：提供开源或安全共享的评估框架和 BioTIER 等标准化基准 (Marshall et al., 2026)；提供减少对稀缺专业知识依赖的自动化红队工具；提供公共资金支持的评估服务；以及采用分级评估流程，只对前沿模型使用成本较高的湿实验室研究。

# 尾注

2. 由于 OpenAI 是最早开创 RLHF 的公司之一，通过 RLHF 训练的模型数量有可能更高，只是尚未被报告 (Christiano, 2017; Ouyang et al., 2022)。

3. 这些技术也被用于灌输无害性以外的其他行为，例如帮助性或强大的推理能力。例如，Meta 使用 RLHF 训练了 Llama 2 和 3.1-8B (Meta, 2023; Meta, 2024) ，而深度求索 (Deepseek) 报告以类似于 RLAIF 的方式训练了 R1 (Deepseek, 2026) 。

4. RLAIF 存在多种变体，本报告旨在举例说明，而非列举所有种类。

5. 原始模型提供方也可以通过激励方式推动这种行为。例如，一家公司可以在许可协议中规定，所有提供该模型推理服务的服务商都必须确保部署某种防护措施，例如生物能力分类器，或者提供稳健性与原公司主要 API 相当的 API。

6. OpenAI表示：“这一过程包括综合运用多种自动化技术 (分类器、推理模型、哈希匹配 (hash-matching)、阻止名单及其他自动化系统) ” (OpenAI, 2025)。

7. 与可信第三方共享模型进行Beta测试也有成熟的先例，无论是直接共享 (例如，Homewood et al., 2025) 还是通过安全的 API (例如，OpenAI, 2023)。

8. 不过，如果训练模型在不改变底层行为的情况下改变其思维链，也可能损害系统的可监测性。参见Baker et al., 2025。

SAIF

WWW.SAIF.ORG