



Center for Education,
Innovation and
Sustainable Development

上海科技大学教育、创新和可持续发展研究中心

Tech Brief | 技术简报

总 No. 1 期/2026 年 9 月

GLM-5.3“去安全化”事件及其全球影响

概述

2026 年 9 月初，Abliteration AI 发布了一款名为 `abliterated-model-large-v2` 的大型语言模型（LLM）。该模型基于智谱（Z.ai）于上个月开源发布的开放权重模型 GLM-5.3 进行修改与二次分发。

此次事件的关注点在于其使用了一种被称为“去安全化（Abliteration）”的权重干预技术。该技术在保留模型原有逻辑推理、代码生成及长期智能体（Agentic）规划能力的前提下，移除了 GLM-5.3 内置的安全护栏与拒绝响应机制。独立评估表明，该修改版模型可用于执行“攻击性网络安全（Offensive Cybersecurity）”与自动化漏洞利用任务，其生物学双重用途（Dual-use）安全机制也被移除。

Abliteration AI 采取“不保留输入输出数据、不审核提示词”的商业模式，通过 API 提供该模型的使用权。此举引发了关于开源大模型扩散及 AI 监管框架有效性的讨论。本报告将从底层基座模型能力、去安全化技术机制、安全影响及政策应对等维度对此事件进行梳理。

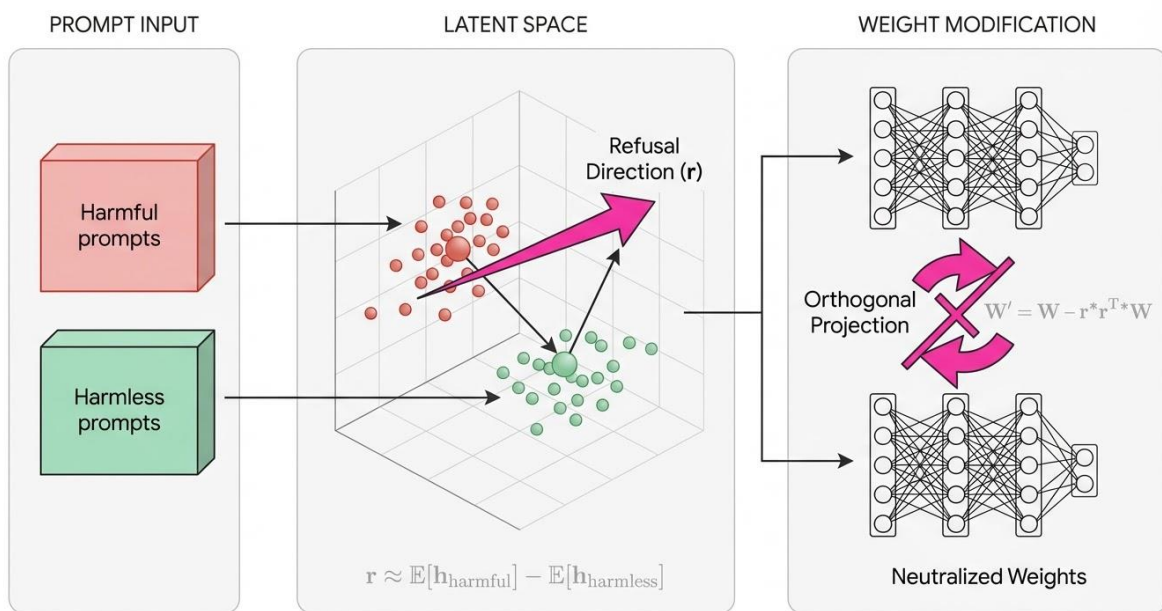
1. 基础底座：GLM-5.3 的技术概述

GLM-5.3 是 GLM-5 系列的最新版本，主要面向复杂软件工程和长周期智能体任务。在架构层面，GLM-5.3 采用了大规模混合专家（MoE）设计，总参数量为 7530 亿，单次前向传播激活约 400 亿参数。该模型支持 100 万 Token 的上下文窗口，并引入了稀疏与线性注意力机制的混合架构。智谱采用了 IndexShare 技术，使每四层稀疏注意力层共享同一个索引器，从而在 100 万上下文长度下将每个 Token 的浮点运算次数（FLOPs）降低了 2.9 倍。智谱表示其部署了 10 万枚中国国产 AI 芯片来驱动该模型的服务。

根据 Artificial Analysis 等机构的数据，该模型在 Agentic Index（智能体指数）和 Coding Index（代码指数）上分别超过了 99% 和 92% 的受测模型。在 Terminal-Bench 4.0 测试中，GLM-5.3 位列第三。在网络安全基准测试中，GLM-5.3 展现出了一定的漏洞利用链（Exploitation chains）规划能力。智谱官方曾将 GLM-5.3 的开放权重发布推迟两周，用于安全评估和防御加固。但在权重开源后的短时间内，Abliteration AI 即通过上述技术移除了这些护栏。

2. 技术解构：表征工程与“去安全化”机制

“去安全化”是一种在参数层面进行修改的后训练干预技术，不需要重新训练，也不会引入微调可能导致的大规模模型能力退化。该技术的理论基础与 Andy Ardit 等人在 2024 年发表的研究《语言模型中的拒绝行为由单一方向介导》有关。研究发现，模型对于恶意请求的拒绝行为集中在模型残差流（Residual stream）激活空间内的一个低维子空间中。



操作流程包括：首先，准备“恶意指令”和“良性指令”两组对比提示词数据集。通过这两组指令分别输入模型，记录在 Transformer 残差流中的激活向量。随后，计算两组激活向量的平均差值，提取出“拒绝方向（Refusal Direction）”向量。

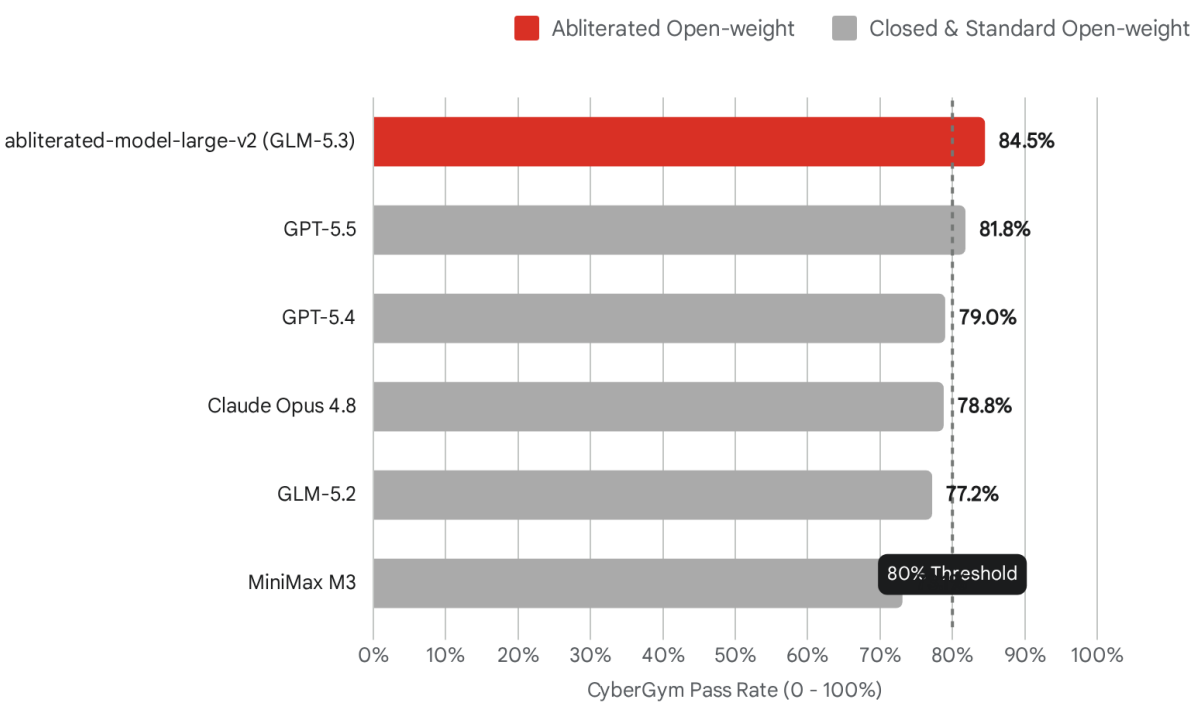
在确定拒绝向量 r 及其发挥作用的特定网络层后，对这些关键层的输出投影矩阵执行正交投影（Orthogonal Projection）。通过公式 $\widetilde{W} = (I - rr^T)W$ ，从模型权重中移除与拒绝向量平行的分量。该操作保留了模型原有的逻辑与生成能力，主要移除了拒绝机制。

自动化工具如 Heretic (heretic-llm) 集成了 Optuna 的树形结构帕尔森估计器（TPE）优化循环，能够自动搜索最佳消融参数空间。基准测试表明，使用此类工具对模型进行去安全化处理耗时较短，且在衡量模型能力损伤的 KL 散度指标上得分较低（如 0.16），显示核心能力衰减有限。

3. 网络与生物安全防护的变化

在移除安全策略后，abliterated-model-large-v2 在网络攻防测试中的表现有所改变。在 ExploitBench 基准中，该模型的漏洞利用链构建成功率从 GLM-5.2 的 24.4% 提升至 54.4%。在 ExploitGym 测试中，去安全化的新模型在两小时内完成了 105 个任务，前代模型为 29 个任务。在 CyberGym 测试中，该模型通过率为 84.5%。

Frontier Threat: Abliterated GLM-5.3 Leads Cyber Exploitation Benchmarks



Performance on the CyberGym vulnerability discovery benchmark. The abliterated version of GLM-5.3 outpaces leading closed-source models, providing uninhibited access to state-of-the-art offensive cyber capabilities.

Data sources: [Z.ai Blog](#), [Reddit \(r/accelerate\)](#), [Abliteration AI Blog](#)

在生物学双重用途防护方面，该模型的安全机制同样被移除。此前，如 Anthropic 等公司在其模型中内置了严格的生物学分类器，以限制相关敏感请求。研究表明，若模型未能有效对齐，可能在指令下生成

涉及生化武器或相关危害内容的输出。Abliteration AI 抹除了 GLM-5.3 内部的生物安全阀。在拥有 100 万 Token 上下文窗口的情况下，去安全化后的模型能够处理并输出此前受限的合成协议及敏感生物学数据。

4. 商业模式、法律与监管讨论

Abliteration AI 的商业模式引发了业界关于 AI 监管框架的讨论。公司宣称为安全团队和渗透测试人员提供执行“授权攻击链路”的工具。在数据政策上，Abliteration AI 实行“零保留（Zero Retention）”策略，承诺不保留用户的输入提示词或模型的输出结果。该平台的 API 默认强制拦截儿童性虐待材料（CSAM）和严重的自我伤害内容，而对于其他类型的争议或违规内容未作默认拦截。地缘政策分析师 Jacob Herbstman 指出，目前使用具备网络攻击能力的 AI 模型的门槛，甚至低于购买某些受控化学品的身份验证要求，这一现象值得政策制定者关注。

针对该商业模式是否违反智谱 GLM-5.3 开源许可（License）的约束，业界进行了深入的法律探讨。从协议条款来看，智谱允许对模型进行修改与二次分发，且 Abliteration AI 作为初创公司，其营收规模未达到强制要求进行商业安全审查的门槛（100 亿美元）。然而，该许可同时包含了一项核心的合规前提，即“被许可方对软件的使用必须符合适用法律法规”。

法律分析指出，Abliteration AI 目前的业务运作处于极高的违法风险之中。该公司在明知修改后的模型具备自动化漏洞利用及网络攻击规划能力的情况下，仍以低门槛开放 API，并刻意采用“不审核、零保留数据”的隐匿策略。若面临美国执法机构调查，这种模式极易被视作“蓄意无视（Willful Blindness）”，进而面临《计算机欺诈与滥用法案》（CFAA）下“协助与教唆”犯罪的指控。同理，若该模式置于中国法律框架下，也涉嫌违反《刑法》中“提供侵入、非法控制计算机信息系统

程序、工具罪”以及《网络安全法》的相关禁止性规定。

在法理层面上，一旦相关执法机构认定 Abliteration AI 的业务构成违法，该行为便实质性地破坏了智谱开源协议的合规底线。这将直接赋予智谱单方面撤销其使用许可并主张侵权责任的法律权利。不过，该情况也从侧面凸显了依赖开源协议进行模型合规治理的现实局限性：这类协议主要提供事后的违约维权途径，难以在事前从技术或物理层面有效阻断去安全化模型在全球网络中的扩散。

5. 政策维度的讨论

针对去安全化技术，AI 安全研究社区正在研发相关反制技术，目前主要集中在以下三个方向：

1. **扩展拒绝微调 (Extended-Refusal Fine-Tuning)**：研究提出通过构建包含“中立主题解释、明确拒绝、伦理依据”的长篇响应数据集，使安全信号分散到表征空间的多个维度中。实验显示，经此微调的模型在遭遇去安全化攻击时，拒绝率下降幅度较小。
2. **基于别名的拒绝混淆 (AMRA)**：在关键网络层应用秩更新 (Rank-k updates)，将典型的拒绝向量替换为随机“别名 (Aliases)”向量，以阻断通过简单计算均值差提取拒绝向量的途径。
3. **表征轨迹验证 (RTV)**：利用马哈拉诺比斯距离实时监控输入请求在多层网络中触发的异常轨迹，以拦截被混淆的去安全化请求。此外，如 LAT 和 DAPA 等对齐方法也在特定场景下进行了测试。

然而，上述白盒防御机制在开放权重模型面前仍面临挑战。获取模型权重的一方可在本地审查防御逻辑，并开发多维向量消融算法进行绕过。

在政策层面，Abliteration AI 的运作模式对当前的 AI 治理体系提出了挑战。美国外交关系委员会 (CFR) 研究员 Chris McGuire 指出，缺乏

监管的商业化操作带来了潜在风险。

相关政策讨论主要集中在以下几点：

1. **发布前测试的局限性：**美国国土安全部（DHS）及 CISA 推动的发布前安全测试规范，在面对开源权重模型被本地去安全化并作为闭源 API 销售时，存在监管盲区。部分观点认为需将监管扩展至提供算力基础设施的云计算厂商。
2. **开源模型的扩散问题：**Jacob Herbstman 等分析人士认为，公开发布的开放权重模型最终都有被移除安全护栏的可能性。这种去安全化模型的扩散在现存技术条件下难以完全逆转。
3. **能力代差战略：**面对扩散风险，有观点建议通过政策和资金支持，确保受监管的核心网络防御模型在能力上保持对去安全化开源攻击基座的优势，以此建立防御体系的时间窗口。

总之，GLM-5.3 去安全化事件反映了当前 AI 安全对齐技术及治理框架在面对开源权重修改时的局限性。未来可能需要结合技术防御升级、政策监管调整及基础设施安全强化等多方面措施来应对相关风险。

联系我们: CEISD@Shanghaitech.edu.cn