安远 AI
CONCORDIA AI

智谱

前沿开放权重AI风险管理框架

Frontier Open-Weight AI Risk Management Framework



2026 年 9 月

执行摘要

开源奠定了现代软件的基础，开放权重则正在成为全球人工智能的基石。它重塑了前沿人工智能走向全球的方式：任何国家和机构都能结合自身条件，审视、改造并部署相关模型、工具与运行框架（harness）。这种可及性带来了显著价值，但也衍生出独特的治理挑战。专有系统的优势在于部署可控、监督集中；开放权重系统的优势，则体现在助推学术研究，激发经济创新，提升透明度，并支持本地化的数据治理。开放权重系统还能定制化构建防御能力，从而与专有系统形成互补，减少对少数闭源提供方的依赖，降低单点故障带来的风险。

然而，开放也会带来独特且可能难以逆转的风险。经微调的开放权重模型可能被移除安全防护措施，进而被用于恶意目的；篡改版本一旦广泛传播，就难以对任何中心节点进行有效问责。其中最值得关注的，是前沿模型的两用能力及其在缺乏监督的情况下可能被滥用的风险。联合国《人工智能问题独立国际科学小组初步报告》、《2026年国际人工智能安全报告》、《新加坡共识：全球人工智能安全研究优先事项》等多项凝聚科学共识的重要报告，均指出应对此类风险的紧迫性，同时也强调了不能以损害开放创新带来的社会价值为代价。

为应对这一抉择背后的核心治理挑战，安远AI与智谱共同编制了本框架。此前，安远AI与上海人工智能实验室联合开发了《前沿人工智能风险管理框架 2.0》；本框架是在其基础上编制的配套文件，但结合开放权重模型的特点，对完整的风险管理生命周期——从风险评估、风险缓解到风险治理——进行了针对性调整。

尽管开放与安全之间存在多种张力，本框架仍致力于在二者之间探寻兼顾平衡的帕累托前沿，并力求搭建首个完备、基于实证的基础，供各方进一步探讨。随着前沿开放权重模型的快速发展，各界对风险管理最佳实践与相关标准的需求已十分迫切。我们诚挚邀请研究者、开发者、政策制定者及社会组织在此基础上继续探索，共同推动开放权重人工智能治理的科学研究与实践落地。

贡献与致谢

贡献者

安远AI：谢旻希、Oskar Galeev、王伟冰、程远、方亮

智谱：乔文斌、王笑尘、张晨

机构简介

安远AI：安远AI是人工智能安全与治理领域的第三方研究和咨询机构，也是目前中国该领域唯一的社会企业。机构立足中国、连接全球，以技术研究为基础，贯通产业实践、政策治理与国际合作，在中国与全球人工智能生态之间发挥可信、专业的桥梁作用。

智谱：智谱由清华大学技术成果转化而来，融合AGI使命与国产大模型情怀，是国内首个、国际较早研发AGI大模型的企业。自成立以来，陆续推出中国首个百亿模型、首个开源千亿模型、首个对话模型、首个多模态模型、全球首个设备操控Agent等。公司持续探索AGI的技术边界与实践应用，历次模型发布旨在推动突破性的技术革新。

致谢

我们感谢以下两场专题研讨会的参与者对本研究部分内容提出的宝贵意见：一是国际人工智能安全与伦理协会（International Association for Safe and Ethical AI）2026年在巴黎举办的年会期间，由安远AI等联合举办的开放权重人工智能风险管理研讨会；二是2026年上海世界人工智能大会期间，由安远AI与FAR.AI联合举办的开放权重人工智能治理研讨会。

感谢Stephen Casper、Mick Yang、易婧玮、李朝卓、Gabriel Wagner以及吴君仪提供的反馈与建议。同时感谢陈欣怡、李晨舟、安远AI其他同事对中文版翻译的贡献。

如何引用本报告

Concordia AI and Z.AI. 2026. *Frontier Open-Weight AI Risk Management Framework*.

安远AI、智谱，2026，《前沿开放权重AI风险管理框架》。

目录

执行摘要 i

贡献与致谢 ii

框架总览 1

前沿开放权重通用型人工智能的定义与范围 2

开放权重通用型人工智能的价值 3

透明度与监督 3

促进人工智能研究 3

应用与创新 4

权力扩散与收益共享 4

风险识别 5

滥用风险 5

意外风险 6

风险阈值 7

部署环境（E） 7

威胁源（T） 8

使能能力（C） 9

社会韧性（R） 9

风险分析 10

开发前 10

部署前 10

部署后 11

风险评价 13

风险缓解 15

开发前 15

开发阶段 15

部署阶段 16

部署后 16

风险治理 18

内部治理 18

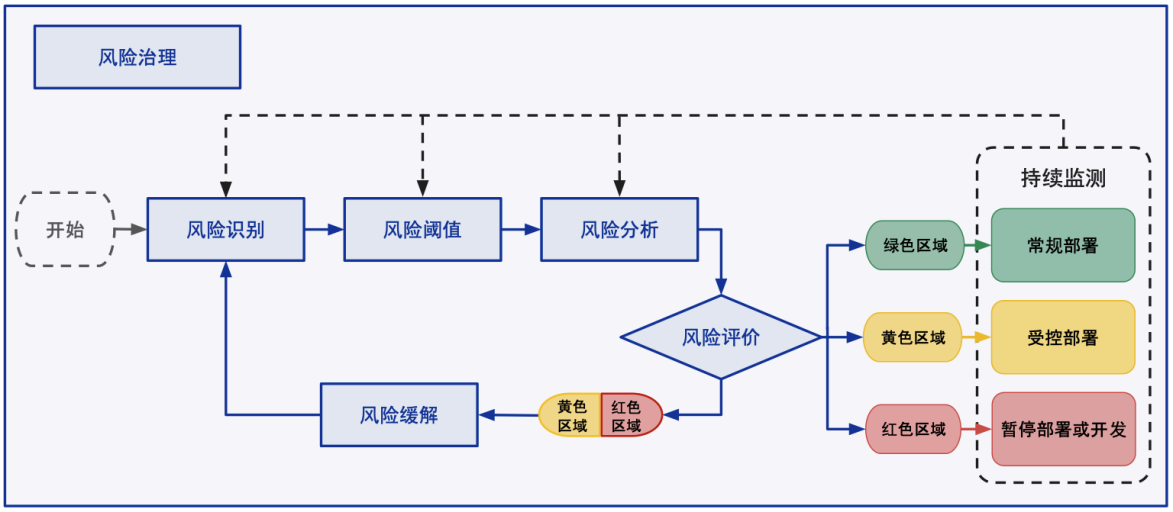
透明度与社会监督 19

参考文献 20

框架总览

《前沿开放权重AI风险管理框架》聚焦开放权重通用型人工智能模型在广泛可得后所带来的特殊、复杂风险。本框架可作为由上海人工智能实验室、安远AI联合发布的《前沿人工智能风险管理框架 2.0》（以下简称《框架 2.0》） [1] 的配套文件，专为适配开放权重发布这一情境，旨在为开发者和更广泛的开放权重生态进行主动风险管理提供可操作的指导方针。同时，通过为开源社区制定技术安全指南，本框架也与联合国人工智能问题独立国际科学小组所建议的方向相契合：以证据为基础，评估人工智能带来的机遇、风险与影响 [2]。该小组及其他重要科学共识文件 [3, 4] 均强调，管理两用能力的风险已十分紧迫。此外，全国网络安全标准化技术委员会（TC260）等国家标准化机构也提出，要推动开源模型提供方与开源社区加强协作，在风险隐患告知、明确禁止性行为、厘清防范滥用之安全义务等方面强化相关规则和责任 [5]。

本框架以六阶段生命周期为核心架构：风险识别、风险阈值、风险分析、风险评价、风险缓解、风险治理。该“六阶段”架构与《框架 2.0》保持一致；不同之处在于，在每个阶段，本框架都针对开放权重模型发布的特点，对相应的风险管理环节作出调整。《框架 2.0》中的多项风险管理指导方针预设开发者可以持续管控模型的部署环节，例如可以设置访问准入机制、监测模型使用行为，或是撤回已发布版本。开放权重发布则从根本上改变了这一点：管控抓手去中心化，监督和风险缓解能力也从原始开发者转移到下游生态系统。本框架将逐一甄别哪些风险管理指导方针依赖开发者层面的管控措施，标示出相关风险管理指导方针在开放权重发布场景下不再适用的情形，并给出无法依赖开发者层面的管控时的替代方案。



需要这种逐阶段适配设计的原因在于：虽然开放权重模型具备诸多价值，例如加快创新节奏、扩大模型可及范围、支撑强有力的外部安全评估，但恰恰是这些特性带来了独特的治理难题，使原开发者无法对下游使用进行管控和监测。例如，开放权重让安全防护措施更容易被移除，

也会让恶意行为者更易通过微调将模型用于有害目的。本框架探讨了上述各类挑战及其背后的取舍权衡。框架并未给出终极解决方案，而是致力于探索如何在保障公共安全的前提下，充分发挥开放带来的价值。

前沿开放权重通用型人工智能的定义与范围

本框架中前沿人工智能模型风险管理方法的适用范围，限于开放权重通用型人工智能模型中的一个子集：在风险相关领域，能力已接近或达到当前前沿水平的模型。现阶段开放权重的定义和治理机制仍在形成之中，所需的风险管理路径尚无共识 [6]。为此，精准划定适用范围十分重要：治理资源应优先投向权重公开发布后会产生重大风险与收益的场景 [7]，同时尽可能不对更大范围的开源生态造成额外阻力。

开放权重模型，即模型参数可供公开下载的模型。前沿模型对外发布后虽常被称为“开源”，但大多模型仅属于“开放权重”发布，并未附带训练流程与数据集 [3]，不属于完整开源。只要权重公开，即符合“开放发布”（open release）的定义 [8]。

对于人工智能的开放程度（openness），目前业界尚未形成统一定义。目前较受认可的界定方式包括开放源代码促进会（Open Source Initiative）发布的《开源人工智能定义》（Open Source AI Definition） [9] 以及Linux基金会发布的《模型开放性框架》（Model Openness Framework）[10]。二者均未把“开放”视为非黑即白的二元属性，而是依据发布了哪些模型组件、适用何种许可协议，将其理解为一个连续的谱系。本框架也延续这一视角，讨论开放带来的风险与收益。

开放权重发布具有单向性 [4，11]。任何专有模型都可以后续开放发布；而开放权重模型一旦公开，就无法再逆向收回、重新闭源。这种非对称性，是开放权重风险管理的核心特征之一。

开放权重通用型人工智能的价值

对开放权重模型进行风险管理，首先需认识到：开放能够带来显著价值；而价值的高低取决于所开放分享的内容有多少——是仅开放模型权重，还是同步开放训练数据、训练代码及配套文件。《2026年国际人工智能安全报告》也指出，相关收益可能超过公开权重所产生的新增风险 [3]。本框架聚焦前沿开放权重模型，在这类模型上，开放的收益与新增风险之间的权衡，对众多下游部署方的影响最为重大。

开放权重发布的价值大致可分为四类。第一是**透明度与监督**：开放便于外部主体审查、验证前沿人工智能模型的属性。第二是**促进人工智能研究**，尤其有助于人工智能安全组织和学术机构开展相关工作。第三是通过拓宽模型获取渠道、降低使用门槛，由此推动**更广泛的应用与创新**。第四是**权力扩散与收益共享**：开放权重把开发、评估、治理人工智能并从中受益的能力，分散到少数具备雄厚实力的国家和前沿模型开发者之外。

透明度与监督

开放权重模型可被外部监督与问责，封闭权重模型却无法以同样方式受到外部制约。透明度带来的价值包括：

- **独立审查：**开发者之外的研究人员可以审计模型及其配套组件，检查漏洞、偏见，核验开发者声明与模型实际表现之间的出入 [12, 13]。
- **更广泛的评测主体：**开放发布把对前沿模型的监督，直接扩展到社会组织、监管机构、下游开发者和学界。规模更广的评测群体不仅能够识别关键安全缺陷，也能为有益应用确立最佳实践 [3, 13]。
- **贴近真实场景的红队测试：**红队可利用攻击者同样能够公开获取的工具开展模拟对抗性攻击。相比黑盒探测，该方式能够更真实地复刻攻击行为 [12]。

促进人工智能研究

开放权重降低了前沿人工智能研究和人工智能安全研究的壁垒，使以下工作成为可能：

- **对齐与可解释性研究：**部分可解释性研究和前沿对齐研究需要直接访问模型参数与内部模式，仅依靠应用程序接口（API）无法提供这种访问条件 [12]。
- **科学可复现性：**开放发布为独立同行评审及前沿人工智能成果复现提供可能，从而核验开发者所宣称的模型能力和安全属性 [8, 14]。但需要说明的是，若无底层数据集和训练代码，仅依靠开放权重尚无法实现完整端到端训练过程的复现。

- **累积式科研进展：**开放发布让资源有限的研究人员能够在既有系统基础上继续开展工作，推动科研成果持续迭代积累 [3]。

应用与创新

开放权重拓展了人工智能可被有效部署的场景，进而在以下方面带来提升：

- **防御能力：**防御方可在自有基础设施上检查、适配并运行开放权重模型，用于漏洞研究和事件响应，从而提高响应速度，减少对少数闭源提供方的依赖 [15, 16, 17]。
- **下游开发效率：**开放权重让开发者能够依托前沿模型能力进行后续开发，降低开发新应用所需的算力、资金和时间成本 [12, 18]。
- **隐私保护与离线部署：**在敏感数据必须本地留存，或是无法获得稳定网络连接的场景下，开放权重可支持端侧与本地部署 [3, 12]。
- **面向服务不足用户的本地化适配：**开放权重便于针对低资源语言、区域化场景和专业细分任务，对模型进行微调，而这类需求往往不是专有模型开发者的优先事项 [3]。

权力扩散与收益共享

开放权重可以把开发、治理人工智能并从中受益的能力，从少数头部主体向外扩散。

- **市场去集中化：**开放权重能够降低行业准入门槛、促进市场竞争、扩大开发者群体，从而缓解人工智能能力集中在少数资源充足企业的局面 [12, 15]。
- **弥合数字鸿沟：**开放发布使发展中地区各类主体得以获取、研究并改造适配前沿人工智能，缩小数字鸿沟，避免其参与人工智能发展的机会被预先阻断 [3, 19, 20, 21]。
- **分布式治理：**开放权重发布扩大了能够实质性影响人工智能开发、评估与治理的主体范围，助力构建分布式风险缓解策略与监督机制，纳入除前沿模型开发者外多方主体的视角 [13]。

上述价值说明了开放权重发布所具备的重要意义与巨大潜力，其中，提升人工智能安全与安保能力也是重要价值之一。然而，这些价值本身并不能决定某一具体前沿模型的权重是否应当开放，以及应当以何种方式开放。这些问题正是风险管理生命周期的起点。本框架将在以下部分按风险管理流程的六个阶段进行分析，并在每个阶段权衡开放带来的收益和边际风险。

风险识别

风险识别是发现、确认和描述风险的过程。风险管理领域的国际标准IEC/ISO 31010 [22] 同时涵盖了基于证据的评审与系统性专家评审，并明确指出，无论采用何种方法，“识别风险时，必须充分考量人员与组织因素”。《2026年国际人工智能安全报告》则指出，对于通用型人工智能系统的风险识别，威胁建模与风险分类体系是目前最成熟的方法 [3]。

《框架 2.0》针对前沿通用型人工智能模型识别出了若干“滥用风险”与“意外风险” [1]，本节说明这两类风险如何因开放权重发布而被放大或改变。《框架 2.0》中还涉及对“失控风险”和“系统性风险”的分析，但本节暂不涉及；因为尚无充分证据表明，相较于专有模型的发布，开放权重发布会加剧上述两类风险。

前沿开放权重通用型人工智能模型的风险识别十分关键，因为此类模型存在多项可能促成风险的特征：权重一旦发布便难以甚至无法撤销，模型通过去中心化分发可以触达大量使用者，其能力也日益处于前沿水平。去中心化的发布模式，也加大了追踪开放权重模型在现实世界中的传播、使用与影响的难度，从而使全面识别其各类风险更具挑战 [6]。本版本所列风险清单反映当前已有证据的实际情况。该风险分类体系将随新型风险的出现迭代更新；若后续有证据表明开放权重发布会放大失控风险与系统性风险，这两类风险也将被进一步纳入考量。

滥用风险

- **开发者失去对模型的可见性：**权重一旦公开发布，开发者便无法观察或追踪模型的最终用途。他们无法监测部署情况、识别滥用模式，也无法把某个具体的滥用事件追溯至对应来源，导致有害应用在事前更难预防，在事后更难补救 [2, 3]。这之中的风险在于，各类主体均可能绕过内置的安全防护措施（如内容过滤器），对模型进行恶意微调 [23]。由于无法掌握部署场景的相关信息，开发者无从知晓此类行为何时发生，进而加剧了滥用风险。
- **降低资源充足的恶意行为者的门槛：**开放权重模型可能让更多主体在不受监督的情况下，强化自身现有能力以用于恶意目的 [24]。当部署成本较低时，为恶意用途而改造开放权重模型的投入也会相应减少 [25]。但前沿模型的门槛相对更高：部署与微调最大规模的开放权重模型需要充足的算力与专业技术能力支撑，因此前沿开放权重模型的该项风险主要集中体现在资源充足的恶意行为者身上。
- **可篡改性：**由于开放权重模型支持自定义配置，恶意行为者可直接移除其内置的安全防护措施 [26]。近期研究表明，对开放权重模型进行此类恶意微调，可能促成危险应用的落地 [27, 28, 29]。经恶意微调的模型版本一旦生成，还可进一步传播扩散；故而单次修改行为就可能导致安全防护措施失效的模型流入大量使用者手中。

- **缺乏可强制执行的人工智能内容标识：**开放权重发布从根本上削弱了确保人工智能生成内容附带可靠来源标识的能力。专有模型API提供方可以在推理层嵌入水印、元数据标签或内容凭证；开放权重模型却可以在本地运行，不具备任何此类监测标记机制；而内置于模型权重中的标识机制，也可能通过微调、权重编辑或推理阶段修改等方式被移除或规避 [30]。这会直接放大滥用风险：恶意行为者可生成大量不携带可检测特征的合成文本、图像、音频或视频，使平台、事实核查机构和终端用户更难将人工智能生成的虚假信息、深度伪造内容、网络钓鱼同真实内容区分开来。这会降低大规模说服与有害操纵的成本，扩大其触达范围，并加大事后溯源和监管执法的复杂度。

意外风险

- **安全机制的非预期退化：**已对齐模型的安全护栏在微调过程中可能遭到削弱，即便该微调并无恶意意图：研究表明，常用的良性数据集用于微调时，可能无意间削弱模型的安全对齐水平。因此，模型的初始对齐效果在经过自定义微调后不一定能够保持，进而产生现有基础设施尚不足以应对的新风险 [31]。
- **缺陷传播：**当大量下游应用依赖少数开放基座模型时，模型内嵌偏见、对抗性漏洞，或规避部署后监测的能力等未解决的缺陷，会随共享组件扩散，引发相关性失效与级联失效 [4, 24, 32]。专有模型可由托管方统一推送修复补丁，开放权重模型的更新则仅在下游开发者自愿采纳的前提下才能生效，且下游开发者是否落实相关更新亦是无从保证的。

风险阈值

“风险阈值”指不可接受风险的明确界限，是判定模型可否发布、能否授权访问、如何缓解其风险以及如何作出治理决策的依据 [1，33]。长期以来，制定风险阈值一直是国际人工智能治理反复强调的优先事项；然而，专门针对开放权重模型的风险阈值，目前仍未形成清晰界定[34]。联合国《人工智能问题独立国际科学小组初步报告》指出，建立各方共同商定、标准化的风险阈值已成为当务之急 [2]。但值得注意的是，开放对治理方法提出了新的要求——模型权重一经发布，许多原先适用于完全封闭专有模型的管控手段便不再适用 [35]。

本节沿用《框架 2.0》中“红线”与“黄线”的定义。“红线”指不应逾越的不可容忍阈值；“黄线”则是潜在风险的早期预警指标，提示相关方需要升级安全与安保措施。《框架 2.0》已就若干关键领域的特定场景作出红线规定，这些领域可能对公共安全和国家安全构成威胁，包括网络攻击、生物威胁、大规模说服与有害操纵以及失控风险，详细内容参见《框架 2.0》表2.1至表2.4 [1]。但就开放权重模型而言，权重一经发布，部署侧的风险缓解措施大多不再具备实施条件，其安全余量须另行设定 [36]。

为校准开放权重模型所需的安全余量，本框架采用四个相互关联的分析维度，分别从部署环境（Deployment Environment）、威胁源（Threat Source）、使能能力（Enabling Capability）和社会韧性（Societal Resilience）四个方面，对潜在危害的可能性与严重程度进行估计。

部署环境（E）

该维度指人工智能模型部署运行的具体场景和约束条件，如部署领域、运行参数、监管环境、用户群体特征、基础设施依赖关系，以及可用的监督机制。即使模型能力保持不变，部署环境的变化也可能显著改变其风险特征。

开放权重的基础模型以去中心化的方式传播和被采用，评估其与部署环境相互作用所产生的风险变得尤为复杂。开放权重模型的运行环境千差万别：任何人都可下载权重、开展微调，再将其应用于医疗、金融等各个领域，乃至不受监管的个人项目；其监管环境和用户群体也各不相同。这将导致环环相扣的脆弱性——由于基础设施依赖关系或监督机制的变化更难追踪，而一旦某个修改版本嵌入关键系统，还可能引发级联失效。相比之下，专有模型通常通过API等受控渠道部署。尽管其对下游实际使用场景的掌握仍然有限，但提供方可以统一设定运行参数并监测使用情况。

由于开放权重模型缺乏集中式的部署后管控措施，因此对于同样一种两用能力，开放权重模型发生未被遏制的滥用的概率会更高。相应地，模型发布前的能力评测的触发条件需要设置得更为保守。专有模型可以适用更高的风险阈值，因为在内部治理机制有效运转的前提下，专有模

型内置的安全防护措施与访问限制能够在风险进一步扩大之前消化一部分风险，而开放权重模型则把风险管理的重任转移给了下游用户和社区。

威胁源（T）

威胁源指通过与人工智能模型交互，可能引发有害后果的源头或行为主体，常见类型包括外部主体（如恶意用户、对抗方）、内部因素（如模型未对齐、训练数据偏差）、操作因素（如人为失误、系统集成故障），以及复杂人机环境互动中产生的涌现行为。

在上述每个类型中，开放权重模型都会使威胁源的范围进一步扩大：外部主体可以更轻易地获取模型权限，对其进行修改和利用；社区层面的微调可能催生涌现行为，进一步增加不可预测性；而由于缺乏集中管控，各类部署场景下的运行故障也更难预判。从理论上讲，专有模型的威胁源范围更有限，因为提供方可以监测使用情况，设置访问速率限制，并封禁存在滥用行为的账号。但在实际场景中，二者并非泾渭分明：即便是托管模型，其微调API也可能使原有的安全护栏失效；闭源系统面向用户的交互界面本身，同样可能成为滥用行为的载体 [33]。

在结合“威胁源”维度校准风险阈值时，必须先明确哪些主体能够实际运用前沿开放权重模型的能力。相关主体至少可分为四类：

- **普通用户：**不修改模型权重，仅依托现有推理框架、已公开的越狱方法及其他现成的绕过工具使用模型。
- **熟练开发者：**借助单节点算力，配合公开可用的低秩适配（LoRA）、微调及模型编辑工具对模型进行调整。
- **专业团队：**具备开展全量微调、模型编辑、智能体工程及系统级优化的工程能力。
- **资源充足的行为体：**拥有专业技术能力、专有数据，以及与开发团队规模相当的资金与算力资源。

相比某一层级主体的独立影响，上述各层级主体之间的相互作用更为关键。资源有限的使用者无需自行对模型进行复杂修改：一旦专业团队或资源充足的行为体移除了安全防护措施，并将不安全的衍生版本二次发布，普通用户便可直接下载运行。

开放权重模型一经发布，便同时向各层级主体开放，不具备筛选准入、排除特定主体的机制。因此，开放权重模型的能力阈值可能需要设定得低于同级别的专有模型。专有模型可以依靠准入管控过滤威胁，因而能够适用更高的能力阈值，不过这也可能掩盖只有在黑盒部署中才会暴露的漏洞。与之相反，开放权重模型可以受益于社区主导的威胁识别。随着安全领域集体努力的积累，其能力阈值或可随时间推移逐步上调。

使能能力（C）

使能能力指人工智能模型具备的核心功能性能力。当模型在未配备额外安全防护措施的状态下部署时，这些能力可能催生特定的风险场景。其涵盖两类能力：一是预期能力，如科学推理、代码编写、任务规划等；二是因模型规模扩大或训练而出现的涌现能力。其中需重点关注瓶颈能力，即对特定危害结果能否发生具有决定性影响的能力。

对开放权重模型而言，使能能力的评估不能以模型发布时的能力为准，而应对照下游资源充足的行为体通过微调所能挖掘出的能力水平。相比之下，专有模型的评估可依据系统在提供方预设配置下运行时所展现的能力，此时系统提示词、拒答训练、监测机制均已部署到位。**因此，对开放权重模型进行风险评估时，必须追踪模型权重所能实现的能力上限（即最坏情况）。**

设想一种情形：如果专有模型部署时的能力超出既定水平，触及了关于恶意使用的特定阈值；那么对于开放权重模型而言，其发布时的能力即使处于相对较低的水平，也应认定已触及同一阈值。这是因为模型权重公开后，部署环境（E）的可变性更强、约束更宽松，威胁源（T）的范围也随之扩大，可对模型进行修改的主体更多。从实践层面看，这意味着开放权重发布中危险能力的判定门槛应被相应降低：同样的能力，在专有模型部署模式下可能处于阈值以下、属于安全范畴；但当部署环境与潜在威胁源的风险等级提高时，便可能突破该阈值。

社会韧性（R）

社会韧性是社会承受人工智能引发的危害并从中恢复的能力，其评估应从多个维度展开，可考虑的方面包括：提升识别虚假信息能力的公共教育项目（例如广泛推进数字素养计划，向用户传授批判性思维与事实核查技能，降低其受操纵的可能性）；能够改变攻防平衡的网络安全基础设施进步（例如采用健全的协议、人工智能辅助的威胁检测手段）；以及强化的生物安全措施（例如提供方执行强制性的DNA序列筛查，防范人工智能赋能的生物工程滥用）[4, 37]。

在增强社会韧性（R）方面，开放权重发布具备封闭部署难以企及的独特优势。通过将模型权重直接交给防御方，开放发布可支持本地化防御和更快速的响应，促进众包安全工具的发展，并降低对集中式基础设施的依赖。但这能否切实降低净风险，取决于防御适配措施的推进节奏与覆盖范围，能否赶超开放权重模型自身的能力迭代速度以及威胁主体获取相关能力的速度。

综合来看，本框架建议：设定风险阈值时，应基于上述维度的组合作出整体判断，而非针对每个维度分别设置测试。具体而言，部署环境（E）、威胁源（T）与使能能力（C）三者共同决定一次开放权重发布所产生的风险暴露水平；社会韧性（R）则决定在这一暴露水平变得不可接受之前，社会还能在安全的前提下承受多大的风险。因此，当E、T、C三者叠加，超出R所能承受的范围时，即视为突破阈值。换言之，同样的能力在发布时，究竟位于风险阈值之上还是之下，将取决于社会在发布时点的韧性能力处于何种水平。

风险分析

风险分析应贯穿人工智能开发生命周期的全流程，用以判定人工智能是否越过“黄线”——即需要升级安全措施的预警指标。我们建议开发者进行开发前与部署前分析，以支撑关键部署决策；同时开展持续的部署后监测，为下一代系统的安全开发提供指引。以下小节将着重说明，开放权重发布为何会使每个分析阶段的风险特征都更为复杂。其中最重要的是，部署前分析必须充分考量开放权重发布特有的风险因素，不能直接沿用为封闭权重模型而设计的评测方法[11]。此外，开放权重的版本一旦扩散，可追溯性十分有限，部署后监测也会因此受限 [38]。

开发前

对于开放权重发布，开发前的风险分析应聚焦设计与训练阶段的决策。这些决策在权重公开后决定着模型的下游风险特征，且一经发布便无法再做修正。在模型开发流程的初始阶段，可运用的核心技术手段有两项：

- **训练数据筛选：**预训练数据或下游微调数据中若混入对抗性内容，会破坏模型的安全对齐效果。研究表明，即便开发者无法预知攻击者将采用何种技术和流程，对训练文本开展迭代式、结构化的筛选，也能够降低模型被越狱的风险 [39]。
- **预期安全缺口：**开发者应当测算同一模型在两种状态下的危险能力差距：安全防护措施完好时，以及安全防护措施被移除后。证据显示，该差距会随模型规模扩大而增大。这意味着，开发前就目标规模与安全训练投入作出决策时，必须以预期安全缺口为依据，而不能仅参考模型训练完成后、安全防护措施完好状态下的能力 [40]。

部署前

由于下游主体可以在模型发布后对其进行微调和修改，在开展开放权重发布的部署前风险分析时，开发者必须推演发布后可能出现的一系列下游模型变体，而不能仅针对单一部署系统进行分析。相关的风险分析技术主要分为两类：

最坏情况风险分析技术

- **模型篡改攻击模拟：**篡改攻击指直接对已发布的模型进行修改的攻击方式，具体手段包括恶意微调、权重编辑、潜在激活变更等，旨在高影响领域中激发出模型的最大能力。常规的“输入—输出”评测只能反映最坏情况行为的最小值；而篡改攻击模拟则通过测试下游主体实际可从模型权重中释放出的最大能力，给出更贴近真实值的风险上限 [41]。例如近期一项研究中，研究人员针对生物与网络滥用的场景，对开放权重

模型进行定制化改造，并分别与开放权重、封闭权重模型的基线水平做对比，据此估计此类篡改带来的增量风险 [23]。

- **完全访问审计：**由于篡改涉及对模型的直接修改，因此，审计开放权重系统必须能够完整访问模型权重及其内部状态 [42]。围绕对抗性微调开展的最坏情况能力评测，应被列为开放权重发布的必要环节 [41]。
- **将抗篡改作为“最坏情况评测”的目标：**对开放权重模型而言，开发者无法依赖系统级安全防护。这意味着，最坏情况分析必须考虑下游主体对权重级安全防护可能施加的全部干预手段，包括微调、引导、模型编辑、剪枝、量化、蒸馏、模型合并与后门植入等。目前既有的安全防护措施普遍难以抵御上述各类干预，抗篡改性因此也是开放权重模型风险管理领域当前尚待攻克的核心技术挑战之一 [6, 43]。

边际风险分析技术

- **基线能力比对：**评估开放权重发布的风险时，需对照既有的工具和专有模型，从而识别出开放权重发布带来的真实“边际风险”。若某个开放权重模型只是复制了在其他渠道已能低成本、轻易获取的同类能力，其净增量风险就很小 [7]。尽管这一“边际风险”视角有助于厘清关于开放权重模型安全性的争论，但多数现有评测方法仍未能对照这些基线对模型进行系统性检验 [24]。
- **评测缺口与风险演化：**仅在发布时点进行基线能力比对，并不能完全回答开放权重发布会在下游造成多大危害的问题。DeepSeek-R1的发布正说明了这一点：如果模型仍然容易被越狱，推理能力的增强反而会让滥用变得更具现实可行性，而下游微调还可能进一步削弱安全防护措施 [44]。与此同时，部署前安全测试本身也越来越难以开展——前沿模型日益能够区分测试环境与真实部署场景，进而利用评测中的漏洞 [3]。要应对这两类问题，可以设置明确的增量评测触发机制：每当训练数据、后训练方法、模型架构、工具使用与智能体能力或发布模式发生重大变化时，开发者都应重新测试受影响的组件，而不是默认沿用先前的安全评估结果。

借助边际风险分析方法，可根据开放权重发布版本相对于闭源前沿的能力水平，分级校准风险缓解措施。能力超过前沿封闭权重模型的发布版本，应当接受最严格的发布前评测、第三方审计以及持续的发布后监测；能力与前沿持平的发布版本，需执行标准化的评测方案，并遵循共享的安全指引；能力明显低于前沿水平的发布版本，则只采取更轻量的措施，例如简化版评测以及社区和行业自治。

部署后

- **开放权重模型的发布后监测存在特有约束：**模型权重一经下载，开发者便不再拥有集中式的观测节点，无法全面掌握模型的使用、适配与再分发情况，下游使用呈现实质去中心化特征，已发布的模型也无法再被收回 [45, 46]。部分新兴技术试图缓解这一困

境，其思路是在所发布的模型中嵌入识别性信号：或是在模型生成内容时添加水印；或是植入指纹，从而对衍生模型进行溯源定位。但与封闭API的场景相比，这两类信号在开放权重环境下的有效性均更难维持 [47, 48]。而且，即便依托此类新兴技术，开放权重模型的部署后监测在方法体系上仍发展不足，这一短板也为开发者群体自身所承认 [49]。

- **间接与代理监测技术：**在开发者无法直接观测下游模型使用情况的情形下，可依托间接信号开展监测。例如，可以参照同类专有系统上的API滥用特征，推断已下载开放权重衍生模型的滥用模式；再如，对公开托管的衍生模型实施技术监测；此外，还可在生态范围内建立共享的风险预警与事件响应机制。这些技术并不能完全弥合可见性缺口，但能够识别各类滥用行为，为风险管理流程提供支撑。

风险评价

本节按风险等级将模型划入三个区域：绿区（采取常规措施即可安全部署）、黄区（需强化安全防护并履行授权程序）、红区（需采取限制开发与部署等非常规管控措施）。我们建议采取迭代式风险降低措施，直至风险降至可接受水平。

与一般的前沿人工智能风险管理流程一致，模型所处的区域直接决定了后续所需的风险缓解措施和治理方案。对于开放权重发布场景而言，核心差异集中体现在黄区。在风险评价环节，开放权重模型面临的关键难点在于安全防护措施的有效性难以持续：一旦模型权重公开发布，通常用于管控黄区风险的访问限制、部署管控等常规手段便会失去效用。因此，在制定部署决策时必须将这一约束条件纳入考量：

- **剩余风险评估取决于安全防护措施的持久性。**权重一经公开发布，安全训练便成为唯一持续生效的风险缓解手段；因此，剩余风险的高低，在很大程度上取决于模型在发布后接受进一步训练时，原有安全训练的防护效果能否保持稳定。然而，少量有害训练样本即可抵消安全训练的作用 [31]。攻击者甚至无需自行收集这类样本，因为经对齐训练的模型可通过提示指令自行生成 [29]。目前通常仅需数步微调，即可移除用于实现模型拒绝有害请求、遗忘危险知识的各类技术手段 [43]。与此同时，用于评估此类防护能力的评测方法，还可能误导评审人员，使其高估安全防护措施的实际持久度 [50]。目前，构建持久有效的安全防护措施并对其开展严格评测，仍是开放权重模型安全领域亟待解决的核心技术问题 [6]。因此，相较于经由受控API部署的模型，开放权重模型的剩余风险估计存在更高的不确定性；这种不确定性应充分体现在风险估计结果的披露方式及其在后续环节的应用之中。
- **开发者应考虑多种部署方案。**分阶段发布，面向经审核的研究人员与审计人员的结构化访问（structured access），面向获批用户的受控下载等模式，既能保留开放模式的核心优势——如外部监督审查、拓展高能力模型的覆盖面——又无需全盘承受无限制开放权重所伴随的风险 [51, 52]。对此类决策而言，核心问题往往不在于“是否发布”，而在于“通过何种渠道发布”。GLM-5.3的发布是该模式的典型实践：发布方以存在两用风险为由，采用结构化访问机制，先由经审核的安全合作伙伴在受控环境下完成模型评测，随后逐步扩大访问范围并开放API接口，待安全评测与发布筹备工作全部完成后，才正式公开完整权重 [53]。归根结底，开发者应结合具体场景，综合权衡开放发布带来的增量风险与新增收益；对于同一模型，不同主体可能基于自身的风险容忍度、所处的生态系统以及正在考虑的具体发布机制，得出不同结论 [24]。

上述考量适用于单次发布场景。但在连续迭代发布的模式下，即便单次发布本身的风险均处于可接受水平，每次带来的小幅风险增量也可能随时间不断累积、逐步放大，最终导致总体风险

显著上升 [3]。因此，有必要针对开放权重发布场景专门设计风险评价方法 [11]；开发者应定期研判模型能力的累积演进态势，判断适宜的发布阈值是否已因此发生偏移。

各区的风险应对策略，都应针对开放权重情形作出调整：

风险级别	决策及处置策略
绿区 （广泛可接受）	**完全开放权重发布** • 基座模型及发布后可能出现的各类衍生修改版本（如未经审核或被恶意微调的衍生版本），其剩余风险整体处于可接受范围。 • 可推进完全开放权重发布。需落实基线安全实践，完善模型文档，并通过现有反馈渠道开展发布后影响监测。
黄区 （可容忍）	**受限和/或分阶段发布** • 只有当开放权重开发者（在模型开发全生命周期中）以及托管平台两者都在上游实施一系列缓解措施时，相关风险才处于可容忍范围。 • 模型发布可在一系列可选方案中择一推进，包括分阶段或渐进发布，面向经审核的研究人员和外部审计人员的结构化访问，针对特定应用场景的受控访问，或在风险评价期间通过API接口提供模型托管服务。
红区 （不可接受）	**暂停发布** • 若模型触及特定领域风险红线[1]，即应暂停或推迟发布，待重新完成风险评价后，再行研判发布策略。 • 若开放权重模型发布后被重新划入红区，鉴于开放权重发布具有不可逆性，风险缓解需转向生态层面的响应措施。

模型权重发布后，风险评价的结论就难以逆转。由于开放权重模型可被任意修改、脱离监管使用，且传播过程不可逆，因此风险评价环节对后续的风险缓解与风险治理阶段具有决定性作用 [6]。鉴于在部署层面和模型使用过程中的管控手段已不再适用，开放权重模型的风险缓解，必须立足于在发布前即已生效、或在权重分发后仍可持续发挥作用的措施。至于风险治理，其范围则不能止步于原始开发者，而应延伸至由开放发布所衍生的各类下游版本构成的生态。

[1] 有关网络攻击风险、生物安全风险、化学安全风险、大规模说服和有害操纵风险以及失控风险的红线阈值，请参见《框架 2.0》第2.2节。

风险缓解

本节梳理开放权重人工智能的风险缓解措施，涵盖从开发前到部署后的模型全生命周期。这一思路与近年多项分类体系的框架逻辑一致 [34, 54, 55]，即风险缓解需在多个节点开展。本节围绕开放权重模型与封闭权重模型之间的一个核心差异展开：开放权重发布后，多数常规部署层的管控手段将不再生效，风险缓解的重心一方面向预训练阶段前移，另一方面向更广泛的生态扩散。

这也是本框架与《框架 2.0》最核心的差异所在 [1]。《框架 2.0》中，黄区的风险缓解工具箱依托三类手段：“了解你的客户”（KYC）机制、API输入输出内容限制，以及对模型部署场景与部署方式的监督。这三类技术手段的生效前提，是开发者保留对用户与模型交互接口的控制权，而开放权重这一发布模式恰恰会导致该控制权的丧失。

开发前

开发前的风险缓解措施在模型权重定型之前便已生效，因而具有独特的持久性：下游主体无法激活模型从未习得的能力。训练数据筛选是这一阶段的首选方案。近期证据显示，它是开放权重发布可用的最有力防线之一。

- **从训练数据中滤除有害内容目前看来是可用的各类防护措施里最具稳固性的手段之一。**近期研究表明，相较于训练后追加防护补丁的模型，基于筛选后数据进行预训练的模型抵御对抗性微调的能力明显更强，且不会折损无关任务的性能 [56, 57, 58]。
- **训练数据筛选无法单独构成完整防御。**若在推理阶段通过检索增强或工具增强注入有害信息，训练数据筛选只能约束模型在训练中所学到的固有知识，却无法阻止模型调用外部输入的危险信息 [57]。此外，当前研究尚未厘清各类训练内容与模型能力之间的对应关系，也还未形成可规模化应用的可靠方法以界定筛选边界 [54]。基于上述局限，风险缓解应采取纵深防御策略，而非仅依赖训练数据筛选这一种手段。

开发阶段

预训练完成后，风险缓解的重心便转移到训练后模型的防护技术上。对于开放权重模型而言，任何安全防护措施都必须能够抵御对抗性微调，才能在权重公开后保持防护效力。受这一约束限制，该领域的研究主要聚焦于抗篡改微调技术；与此同时，一批旨在将安全防护深度嵌入模型或其运行环境的新兴技术也逐步产生。

- **抗篡改微调。**当下，越来越多风险缓解技术致力于把安全防护措施深度内置到模型之中，让攻击者难以简单通过少量微调就移除安全防护措施，想要绕过防护的技术门槛几乎等同于完整重训练模型 [43, 59, 60]。《2026年国际人工智能安全报告》认为该技

术路线具备发展潜力，但现阶段还无法达到稳定可靠的防护效果 [3]。具体而言，现有研究显示，“机器遗忘”技术只能将有害知识弱化、模糊化，无法做到彻底清除；即便是当前效果最好的方案，依旧存在被攻破的可能：攻击者只需用10个无关样本对模型进行微调，或是在激活空间修改特定向量方向，就能绕过防护 [61]。就算机器遗忘能够发挥部分作用，面对强上下文学习能力的模型依旧会暴露出短板——借助提示词，已经被清除的知识可以重新恢复 [62]。另有最新研究探索在预训练阶段把特定能力拆解成独立模块，模型发布时可按需保留对应模块、不予发布；不过这类方案目前只在小规模模型上完成了验证 [63]。

- **脚手架式防护方法。**该类方法属于附加型防护机制，护栏模型是其典型代表：在模型输出送达用户之前，先对提示词开展恶意意图筛查，并校验回复内容是否存在安全违规情形。任何掌握模型权重的主体均可轻易禁用或绕过此类防护，但它们仍具备实际作用：防范意外误用、震慑技术能力较弱的行为主体，同时为下游开发者构建自身安全体系提供起点 [54]。
- **密码学与硬件层面方法。**参数级加密、GPU端密钥管理等新兴技术手段，可以让模型仅在完成授权解密步骤后方能被修改或运行 [64]。近期安全研究正在探索人工智能模型可验证的数学安全边界 [65]。在理论层面上，这类技术路径有望缓解开放权重模型的风险，但尚未在前沿模型规模下得到验证。

部署阶段

部署环节本身就是风险缓解的关键节点。开放权重模型一经发布便不可撤回，因此该阶段有两项决策至关重要：选择何种发布渠道，以及哪些文档内容应与模型权重同步发布。

- **不同的发布策略分布于一条连续光谱上。**部署方式涵盖完全封闭的API调用访问，到彻底开放模型权重，这两者中间还有受控访问、仅面向研究者开放等过渡方案 [51]。分阶段发布则是逐步放开访问权限，让开发者和防御方能够在每个阶段监测模型的使用情况 [16, 54, 66]。
- **与权重同步发布的配套文档可用于开展独立审查。**详实的模型卡、技术报告及评测数据，可为外部研究者提供研究系统本身及其下游影响的必要背景信息 [54]。最佳实践示例包括：随技术报告、基准测试结果与权重一并发布安全分类器 [67]，同时引入同行评审机制 [44, 68]。

部署后

模型权重一经发布，风险管理的重心便转向衍生模型所构成的下游生态的风险监测与缓解。

- **模型溯源技术。**相关方法包括：通过推断衍生模型对应的基础模型，还原完整的模型谱系 [69]；在模型权重中嵌入水印，使模型副本即便经过修改仍可被识别 [70]。这两

类技术并非无法突破，但突破均需投入相应技术成本；其核心价值在于显著抬高匿名滥用的门槛，并为事后责任溯源提供技术支撑 [54]。另有配套研究提出对开放模型生态本身进行图谱化构建，将公开模型仓库映射为可检索的模型图谱，支撑属性预测、衍生模型识别等下游应用场景 [71]。

- **数据溯源技术可用于追溯人工智能生成内容的来源。**依托不可感知的水印与元数据标准，研究人员可追踪人工智能生成内容（含开放权重模型的输出内容）的传播路径。这类技术同样存在被移除的可能，但目前已具备实用价值，正逐步在全生态范围内落地推广 [54]。
- **分发平台是生态层面实施访问管控的天然节点。**由于开发者不再承担访问中介职能，KYC机制与使用行为监测工作向外转移至分发平台与下游部署方，这也是当前仅存的主要管控抓手 [54]。

风险治理

风险治理阶段的核心目标是建立组织架构、监督机制与问责框架，确保整个风险管理流程得到严格落地和持续监测，并随威胁态势与技术能力的演进而动态迭代优化。国际风险管理标准强调，无论采用何种风险评估方法，在识别与描述风险时都必须充分考量人员因素与组织因素[22]。模型权重一经发布，治理责任主体便不再局限于单一开发主体，而是遍布价值链各环节的多元参与方。TC260发布的《人工智能安全治理框架 3.0》明确回应了这一问题，该框架指出，开源生态需强化安全保障措施，呼吁模型提供方与开源社区共同制定通用规则，向用户告知下载模型对应的风险与安全责任，并明确开放权重模型的禁止性行为 [5]。人工智能合作组织（Partnership on AI）也持相近观点，将模型提供方、模型适配方、托管服务方、下游应用开发者均纳入治理责任主体范畴，各方角色定位各异，可运用的治理抓手也不尽相同 [34]。

内部治理

《框架 2.0》的内部治理章节聚焦开发者自身管理流程，包括风险委员会、审查委员会及内部问责体系 [1]。对于开放权重发布场景，这些机制虽仍不可或缺，但仅靠其已不足以覆盖全部风险：开发者的内部控制无法延伸至模型发布后进行微调、托管或二次部署的各类主体。

- **开发者层面的内部治理仍是整个治理体系的基础。**与欧盟《人工智能法案》配套的《通用型人工智能行为准则》（General-Purpose AI Code of Practice，以下简称《行为准则》）指出，开放许可的通用型人工智能模型提供方，若模型具有系统性风险，则需要开展模型评测，实施对抗性测试，跟踪并上报严重事件，以及落实网络安全防护措施 [72]。行业指引则强调，应记录开发流程，共享评测框架，发布可复用的开放许可成果 [73]，同时通过内部审计与文档记录监测模型失效、性能漂移与不安全行为 [74]。另有建议提出，开发者应设立开源治理委员会，开展发布前的安全风险评估，并在组织内部制定应急响应机制 [75]。
- **价值链下游各环节同样需要建立配套的内部治理流程。**由于开发者层面的管控措施无法随模型权重同步传导，模型适配方、托管服务方、下游部署方均需根据自身角色定位建立相应的治理架构 [34]。《行为准则》明确了开发者与下游提供方的边界划分规则：当下游修改方对模型的修改显著改变其通用性、能力水平或系统性风险时，该修改方即成为该通用型人工智能模型的 “提供方” [76]。
- **问责归属不明确仍是当前的一大治理空白。**即便建立了分布式内部治理体系，开放权重模型发布后若被恶意修改并被用于有害目的，其责任主体仍不明确 [3]。本框架将其作为有待进一步探讨的开放性问题，下一小节所述的透明度与外部监督机制，正是为部分回应这一问题而设计。

透明度与社会监督

模型权重一旦公开，外部各方就能开展封闭权重模式下无法进行的审查、评测与漏洞识别。正因如此，透明度本身构成一项独立的治理机制。

- **开放权重使外部主体得以开展封闭权重模式下无法进行的工作。**已有研究指出，开放并不与人工智能安全目标相矛盾，反而有助于提升安全水平：偏见问题与安全漏洞能够更早暴露并被及时修复，评测结论经开发者社区交叉核验，可信度更高；开放代码也为外部审计提供了可验证的依据 [77, 78]。但这一优势不会自动实现，它取决于开源生态是否具备相应的技术能力与协同机制，能否对发现的问题及时采取行动。
- **监测与披露义务需延伸至前沿模型的内部推理过程。**如今前沿模型越来越多地引入扩展式内部推理与思考机制，而这些过程无法从最终输出中完整观测到，并会在长时间、多轮的智能体交互中持续发挥作用。这就为欺骗性对齐留下了更大空间——即模型内部已经形成有害意图，但对外输出仍保持合规的情形 [79, 80]。因此，模型卡与系统卡应披露以下内容：是否启用扩展式内部推理功能，采用了何种安全监测机制（是覆盖完整推理链，还是仅监测最终输出）以及监测尚未覆盖的剩余风险。
- **模型权重与推理托管平台是开展外部监督的核心载体。**HuggingFace、GitHub、ModelScope等平台汇聚了可供检索和下载的模型卡、技术报告与评测数据，研究人员、审计机构与社会组织可在模型发布后对其开展审查 [81]。不过平台监督也存在局限：原始模型的使用限制无法在重新上传至平台的衍生模型中得到有效落实 [38]。对此，开发者可采取配套措施加以补充，包括发布前向审计人员开放完整权限、采用逐步扩大受众范围的分阶段发布，以便在模型广泛开放前排查缺陷 [3]。

参考文献

[1] Shanghai Artificial Intelligence Laboratory and Concordia AI, Frontier AI Risk Management Framework 2.0, Concordia AI, Jul. 2026. [Online]. Available: https://concordia-ai.com/wp-content/uploads/2026/08/Frontier-AI-Risk-Management-Framework-v2.0-EN-2026-07.pdf

[2] Independent International Scientific Panel on AI, *Preliminary Report of the Independent International Scientific Panel on AI: Evidence-based assessment of opportunities, risks and impacts of AI*, United Nations, Jul. 2026. [Online]. Available: https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report

[3] Y. Bengio et al., International AI Safety Report 2026, U.K. Department for Science, Innovation and Technology, Feb. 2026. [Online]. Available: https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026_1.pdf

[4] International Scientific Exchange on AI Safety, The 2026 Singapore Consensus on Global AI Safety Research Priorities, Infocomm Media Development Authority (IMDA), Singapore, Jul. 2026. [Online]. Available: https://www.imda.gov.sg/activities/activities-catalogue/international-scientific-exchange

[5] 全国网络安全标准化技术委员会 (National Technical Committee 260 on Cybersecurity of SAC), 人工智能安全治理框架3.0 (AI Safety Governance Framework 3.0), Sep. 14, 2026. [Online]. Available: https://www.cac.gov.cn/2026-09/14/c_1791137092283345.htm

[6] S. Casper et al., Open Technical Problems in Open-Weight AI Model Risk Management, *Transactions on Machine Learning Research*, Mar. 2026. [Online]. Available: https://openreview.net/pdf?id=8QyGLnFkzc

[7] S. Kapoor et al., On the Societal Impact of Open Foundation Models, Feb. 27, 2024. arxiv: 2403.07918 (cs). [Online]. Available: https://arxiv.org/abs/2403.07918

[8] IEEE-USA, Open-Weight AI Models, IEEE-USA Position Statement, Nov. 2025. [Online]. Available: https://ieeeusa.org/assets/public-policy/positions/ai/Open_Weight_AI_Models_1125.pdf

[9] Open Source Initiative, The Open Source AI Definition 1.0, Open Source Initiative, Oct. 28, 2024. [Online]. Available: https://opensource.org/ai

[10] Linux Foundation AI & Data, Model Openness Framework, Linux Foundation AI & Data, n.d. [Online]. Available: https://isitopen.ai/

[11] P. Paskov, C. Rodriguez, S. Dev, and S. Casper, Open Weight AI Models Require Proportional Evaluation Approaches, Jun. 18, 2026. arxiv: 2606.19890 (cs). [Online]. Available: https://arxiv.org/abs/2606.19890

[12] OECD, AI Openness: A Primer for Policymakers, OECD Artificial Intelligence Papers, No. 44, Aug. 14, 2025. [Online]. Available: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/08/ai-openness_958d292b/02f73362-en.pdf

[13] C. François et al., A Different Approach to AI Safety: Proceedings from the Columbia Convening on Openness in Artificial Intelligence and AI Safety, Jun. 27, 2025. arxiv: 2506.22183 (cs). [Online]. Available: https://arxiv.org/pdf/2506.22183

[14] DeepSeek AI, DeepSeek-R1, Hugging Face model card, Jan. 20, 2025. [Online]. Available: https://huggingface.co/deepseek-ai/DeepSeek-R1

[15] NVIDIA, Industry Leaders Unite in Open Secure AI Alliance for AI Safety and Security, NVIDIA Blog, Jul. 27, 2026. [Online]. Available: https://blogs.nvidia.com/blog/open-secure-ai-alliance/

[16] Thinking Machines Lab, A Safe Path to Open Weights, Thinking Machines Lab, Jul. 31, 2026. [Online]. Available: https://thinkingmachines.ai/blog/a-safe-path-to-open-weights/

[17] UK AI Security Institute, How Far Behind the Frontier are Leading Open Weight Models on Cyber?, AI Security Institute Blog, Jul. 17, 2026. [Online]. Available: https://www.aisi.gov.uk/blog/how-far-behind-the-frontier-are-leading-open-weight-models-on-cyber

[18] 安远AI (Concordia AI), 基础模型的负责任开源 (Responsible Open Sourcing of Foundation Models), Concordia AI, Apr. 2024. [Online]. Available: https://concordia-ai.com/wp-content/uploads/2024/04/%E5%9F%BA%E7%A1%80%E6%A8%A1%E5%9E%8B%E7%9A%84%E8%B4%9F%E8%B4%A3%E4%BB%BB%E5%BC%80%E6%BA%90.pdf

[19] Ministry of Foreign Affairs of People's Republic of China, Global AI Governance Initiative, Oct. 20, 2023. [Online]. Available: https://www.mfa.gov.cn/eng/zy/gb/202405/t20240531_11367503.html

[20] United Nations General Assembly, Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development, United Nations, A/RES/78/265, Mar. 21, 2024. [Online]. Available: https://docs.un.org/en/A/res/78/265

[21] Orinoco Tribune, Final Declaration of Havana's G77+China Summit, Orinoco Tribune, Sep. 18, 2023. [Online]. Available: https://orinocotribune.com/final-declaration-of-havanas-g77china-summit/

[22] IEC/ISO, Risk management — Risk assessment techniques, IEC/ISO 31010:2019, 2019. [Online]. Available: https://www.iso.org/standard/72140.html

[23] E. Wallace, O. Watkins, M. Wang, K. Chen, and C. Koch, Estimating Worst-Case Frontier Risks of Open-Weight LLMs, Aug. 5, 2025. arxiv: 2508.03153 (cs). [Online]. Available: https://arxiv.org/html/2508.03153v1

[24] Y. Bengio et al., International AI Safety Report 2025, U.K. Department for Science, Innovation and Technology, Jan. 29, 2025. [Online]. Available: https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf

[25] Shanghai Artificial Intelligence Laboratory et al., Frontier AI Risk Management Framework in Practice: A Risk Analysis Technical Report, Jul. 22, 2025. arxiv: 2507.16534 (cs). [Online]. Available: https://arxiv.org/pdf/2507.16534

[26] Y. Gal and S. Casper, Customizable AI systems that anyone can adapt bring big opportunities — and even bigger risks, *Nature*, vol. 646, pp. 286–287, Oct. 7, 2025. [Online]. Available: https://www.nature.com/articles/d41586-025-03228-9

[27] A. Chang, N. Conley, H. S. Ganesan, and A. Swanda, Death by a Thousand Prompts: Open Model Vulnerability Analysis, Nov. 5, 2025. arxiv: 2511.03247 (cs). [Online]. Available: https://arxiv.org/html/2511.03247v1

[28] E. Wallace, O. Watkins, M. Wang, K. Chen, and C. Koch, Estimating worst-case frontier risks of open-weight LLMs, OpenAI, Aug. 5, 2025. [Online]. Available: https://openai.com/index/estimating-worst-case-frontier-risks-of-open-weight-llms/

[29] J. Yi, R. Ye, Q. Chen, B. Zhu, S. Chen, D. Lian, G. Sun, X. Xie, and F. Wu, On the Vulnerability of Safety Alignment in Open-Access LLMs, *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9236–9260, Aug. 2024. [Online]. Available: https://aclanthology.org/2024.findings-acl.549.pdf

[30] Y. Zhao, Z. S. Wu, and A. Block, MarkTune: Improving the Quality-Detectability Trade-off in Open-Weight LLM Watermarking, Dec. 3, 2025. arxiv: 2512.04044 (cs). [Online]. Available: https://arxiv.org/pdf/2512.04044

[31] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!, Oct. 5, 2023. arxiv: 2310.03693 (cs). [Online]. Available: https://arxiv.org/abs/2310.03693

[32] R. Bommasani, K. A. Creel, A. Kumar, D. Jurafsky, and P. Liang, Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?, *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/17a234c91f746d9625a75cf8a8731ee2-Abstract-Conference.html

[33] I. Solaiman, R. Bommasani, D. Hendrycks, A. Herbert-Voss, Y. Jernite, A. Skowron, and A. Trask, Beyond Release: Access Considerations for Generative AI Systems, Feb. 23, 2025. arxiv: 2502.16701 (cs). [Online]. Available: https://arxiv.org/pdf/2502.16701

[34] M. Srikumar, J. Chang, and K. Chmielinski, Risk Mitigation Strategies for the Open Foundation Model Value Chain, Partnership on AI, Jul. 11, 2024. [Online]. Available: https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/

[35] United Nations High-level Advisory Body on Artificial Intelligence, Governing AI for Humanity: Final Report, United Nations, 2024. [Online]. Available: https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

[36] D. Raman, N. Madkour, E. R. Murphy, K. Jackson, and J. Newman, Intolerable Risk Threshold Recommendations for Artificial Intelligence, Mar. 4, 2025. arxiv: 2503.05812 (cs). [Online]. Available: https://arxiv.org/pdf/2503.05812

[37] J. Bernardi, G. Mukobi, H. Greaves, L. Heim, and M. Anderljung, Societal Adaptation to Advanced AI, Centre for the Governance of AI (GovAI), May 23, 2024. [Online]. Available: https://cdn.governance.ai/Societal_Adaptation_to_Advanced_AI.pdf

[38] W. Xu, H. Ye, H. Ye, K. Gao, V. Filkov, and M. Zhou, A governance horizon for ethical-use constraints in open-weight AI models, May 23, 2026. arxiv: 2605.24383 (cs). [Online]. Available: https://arxiv.org/abs/2605.24383

[39] X. Liu, J. Liang, M. Ye, and Z. Xi, Robustifying Safety-Aligned Large Language Models through Clean Data Curation, May 24, 2024. arxiv: 2405.19358 (cs). [Online]. Available: https://arxiv.org/abs/2405.19358

[40] A.-K. Dombrowski, D. Bowen, A. Gleave, and C. Cundy, The Safety Gap Toolkit: Evaluating Hidden Dangers of Open-Source Models, Jul. 8, 2025. arxiv: 2507.11544 (cs). [Online]. Available: https://arxiv.org/pdf/2507.11544

[41] Z. Che, S. Casper, R. Kirk, A. Satheesh, S. Slocum, L. E. McKinney, R. Gandikota, A. Ewart, D. Rosati, Z. Wu, Z. Cai, B. Chughtai, Y. Gal, F. Huang, and D. Hadfield-Menell, Model Tampering Attacks Enable More Rigorous Evaluations of LLM Capabilities, Feb. 3, 2025. arxiv: 2502.05209 (cs). [Online]. Available: https://arxiv.org/pdf/2502.05209

[42] S. Casper et al., Black-Box Access is Insufficient for Rigorous AI Audits, Jan. 25, 2024. arxiv: 2401.14446 (cs). [Online]. Available: https://arxiv.org/abs/2401.14446

[43] R. Tamirisa, B. Bharathi, L. Phan, A. Zhou, A. Gatti, T. Suresh, M. Lin, J. Wang, R. Wang, R. Arel, A. Zou, D. Song, B. Li, D. Hendrycks, and M. Mazeika, Tamper-Resistant Safeguards for Open-Weight LLMs, Aug. 1, 2024. arxiv: 2408.00761 (cs). [Online]. Available: https://arxiv.org/pdf/2408.00761

[44] D. Guo et al., DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning, *Nature*, vol. 645, no. 8081, pp. 633–638, Sep. 17, 2025. pubmed: 40962978. [Online]. Available: https://www.nature.com/articles/s41586-025-09422-z

[45] A. Rao, D. Keller, N. Kalra, R. Steed, K. Kwegyir-Aggrey, K. Klyman, D. Staheli, and S. Bergman, Challenges to the Monitoring of Deployed AI Systems, NIST AI 800-4, National Institute of Standards and Technology, Mar. 2026. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-4.pdf

[46] J. Pratt and A. Tanjaya, Documenting the Impacts of Foundation Models: A Progress Report on Post-Deployment Governance Practices, Partnership on AI, Feb. 26, 2025. [Online]. Available: https://partnershiponai.org/wp-content/uploads/2025/02/PAI_documentation-progress-report.pdf

[47] T. Gloaguen, N. Jovanović, R. Staab, and M. Vechev, Towards Watermarking of Open-Source LLMs, Feb. 14, 2025. arxiv: 2502.10525 (cs). [Online]. Available: https://arxiv.org/pdf/2502.10525

[48] J. Xu, F. Wang, M. D. Ma, P. W. Koh, C. Xiao, and M. Chen, Instructional Fingerprinting of Large Language Models, Jan. 21, 2024. arxiv: 2401.12255 (cs). [Online]. Available: https://arxiv.org/abs/2401.12255

[49] Meta, Meta Advanced AI Scaling Framework v2, Meta, n.d. [Online]. Available: https://ai.meta.com/static-resource/Meta_Advanced-AI-Scaling-Framework-v2

[50] X. Qi, B. Wei, N. Carlini, Y. Huang, T. Xie, L. He, M. Jagielski, M. Nasr, P. Mittal, and P. Henderson, On Evaluating the Durability of Safeguards for Open-Weight LLMs, Dec. 10, 2024. arxiv: 2412.07097 (cs). [Online]. Available: https://arxiv.org/pdf/2412.07097

[51] I. Solaiman, The Gradient of Generative AI Release: Methods and Considerations, Feb. 5, 2023. arxiv: 2302.04844 (cs). [Online]. Available: https://arxiv.org/abs/2302.04844

[52] E. Seger et al., Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives, Centre for the Governance of AI (GovAI), Sep. 29, 2023. [Online]. Available: https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf

[53] Z.AI (@Zai_org), GLM-5.3 Release Statement, X (formerly Twitter), Aug. 14, 2026. [Online]. Available: https://x.com/Zai_org/status/2088280509474320693

[54] UK AI Security Institute, Managing risks from increasingly capable open-weight AI systems, AI Security Institute Blog, Aug. 29, 2025. [Online]. Available: https://www.aisi.gov.uk/blog/managing-risks-from-increasingly-capable-open-weight-ai-systems

[55] Frontier Model Forum, Preliminary Taxonomy of AI-Bio Misuse Mitigations, Frontier Model Forum Issue Brief, Jul. 30, 2025. [Online]. Available: https://www.frontiermodelforum.org/issue-briefs/preliminary-taxonomy-of-ai-bio-misuse-mitigations/

[56] P. Maini, S. Goyal, D. Sam, A. Robey, Y. Savani, Y. Jiang, A. Zou, M. Fredrikson, Z. C. Lipton, and J. Z. Kolter, Safety Pretraining: Toward the Next Generation of Safe AI, Apr. 23, 2025. arxiv: 2504.16980 (cs). [Online]. Available: https://arxiv.org/pdf/2504.16980

[57] K. O'Brien, S. Casper, Q. Anthony, T. Korbak, R. Kirk, X. Davies, I. Mishra, G. Irving, Y. Gal, and S. Biderman, Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs, Aug. 8, 2025. arxiv: 2508.06601 (cs). [Online]. Available: https://arxiv.org/pdf/2508.06601

[58] Y. Chen, M. Tucker, N. Panickssery, T. Wang, F. Mosconi, A. Gopal, C. Denison, L. Petrini, J. Leike, E. Perez, and M. Sharma, Enhancing Model Safety through Pretraining Data Filtering, Anthropic Alignment Blog, Aug. 19, 2025. [Online]. Available: https://alignment.anthropic.com/2025/pretraining-data-filtering/

[59] D. Rosati, J. Wehner, K. Williams, Ł. Bartoszcze, D. Atanasov, R. Gonzales, S. Majumdar, C. Maple, H. Sajjad, and F. Rudzicz, Representation Noising: A Defence Mechanism Against Harmful Finetuning, May 23, 2024. arxiv: 2405.14577 (cs). [Online]. Available: https://arxiv.org/pdf/2405.14577

[60] P. Henderson, E. Mitchell, C. Manning, D. Jurafsky, and C. Finn, Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models, *Proc. 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*, Aug. 8, 2023. [Online]. Available: https://doi.org/10.1145/3600211.3604690

[61] J. Łucki, B. Wei, Y. Huang, P. Henderson, F. Tramèr, and J. Rando, An Adversarial Perspective on Machine Unlearning for AI Safety, Sep. 26, 2024. arxiv: 2409.18025 (cs). [Online]. Available: https://arxiv.org/pdf/2409.18025

[62] I. Shumailov, J. Hayes, E. Triantafillou, G. Ortiz-Jimenez, N. Papernot, M. Jagielski, I. Yona, H. Howard, and E. Bagdasaryan, UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI, Jul. 2024. arxiv: 2407.00106 (cs). [Online]. Available: https://arxiv.org/pdf/2407.00106

[63] E. Roland, M. Cubuktepe, E. Martinez, S. Servaes, K. Pepper, M. Vaiana, D. S. de Lucena, J. Rosenblatt, A. Foote, C. Anil, and A. Cloud, Modular Pretraining Enables Access

Control, Anthropic Alignment Blog, Jul. 8, 2026. [Online]. Available: https://alignment.anthropic.com/2026/modular-pretraining/

[64] OpenAI, Reimagining secure infrastructure for advanced AI, OpenAI Blog, May 4, 2024. [Online]. Available: https://openai.com/index/reimagining-secure-infrastructure-for-advanced-ai/

[65] 王小云 (Xiaoyun Wang), 密码技术与人工智能安全 (Cryptography and AI Security), 中国科学院数学与系统科学研究院系统科学研究所 (Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences), Dec. 8, 2025. [Online]. Available: http://iss.amss.cas.cn/gb2019/lt/xtkxfzlt/202512/t20251208_804832.html

[66] Z.AI, GLM-5.3: Frontier Coding with Emergent Cyber Capabilities, Z.AI Blog, Aug. 14, 2026. [Online]. Available: https://z.ai/blog/glm-5.3

[67] Qwen Team, Qwen3Guard, Qwen (Alibaba) Blog, Sep. 23, 2025. [Online]. Available: https://qwen.ai/blog?id=f0bbad0677edf58ba93d80a1e12ce458f7a80548&from=research.research-list

[68] Nature, Bring us your LLMs: why peer review is good for AI models, *Nature*, editorial, Sep. 17, 2025. [Online]. Available: https://www.nature.com/articles/d41586-025-02979-9

[69] E. Horwitz, A. Shul, and Y. Hoshen, Unsupervised Model Tree Heritage Recovery, May 28, 2024. arxiv: 2405.18432 (cs). [Online]. Available: https://arxiv.org/abs/2405.18432

[70] F. Boenisch, A Systematic Review on Model Watermarking for Neural Networks, Dec. 2021. arxiv: 2009.12153 (cs). [Online]. Available: https://arxiv.org/abs/2009.12153

[71] E. Horwitz, N. Kurer, J. Kahana, L. Amar, and Y. Hoshen, We Should Chart an Atlas of All the World's Models, Mar. 13, 2025. arxiv: 2503.10633 (cs). [Online]. Available: https://arxiv.org/abs/2503.10633

[72] Future of Life Institute, High-level summary of the AI Act, EU AI Act Website, Feb. 27, 2024. [Online]. Available: https://artificialintelligenceact.eu/high-level-summary/

[73] C. Osborne, Charting Strategic Directions for Global Collaboration in Open Source AI: Key Takeaways from the GOSIM Open Source AI Strategy Forum, The Linux Foundation, Jul. 2025. [Online]. Available: https://www.linuxfoundation.org/hubfs/Research%20Reports/lfr_gosim25_072325.pdf

[74] Y. Jernite and I. Solaiman, Hugging Face Comments on AI Accountability for the National Telecommunications and Information Administration, Hugging Face (NTIA RFC Response), Jun. 12, 2023. [Online]. Available: https://huggingface.co/blog/assets/151_policy_ntia_rfc/HF_NTIA_RFC.pdf

[75] G. Wagner, J. Zhou, K. Y. Ng, and B. Tse, State of AI Safety in China 2025, Concordia AI, Jul. 2025. [Online]. Available: https://concordia-ai.com/wp-content/uploads/2025/07/State-of-AI-Safety-in-China-2025.pdf

[76] J. Farrell and T. Emborg, An Introduction to the Code of Practice for General-Purpose AI, Future of Life Institute, EU AI Act Website, Jul. 3, 2024. [Online]. Available: https://artificialintelligenceact.eu/introduction-to-code-of-practice/

[77] 王哲 (Zhe Wang) and 薛澜 (Lan Xue), 大模型开源创新公地：历史演进、价值逻辑与中国叙事 (Open-Source Innovation Commons for Large Models: Historical Evolution, Value Logic

and Chinese Narrative), 探索与争鸣 (Exploration and Free Views), no. 7, 2025. [Online]. Available: https://ciss.tsinghua.edu.cn/info/zlyaq/8618

[78] 傅宏宇 (Hongyu Fu) and 彭靖芷 (Jingzhi Peng), 开源人工智能治理的全球实践及路径选择 (Global Practices and Path Choices in Open-Source AI Governance),中国信息安全 (China Information Security), no. 2, 2025. [Online]. Available: https://www.secrss.com/articles/80066

[79] T. Korbak et al., Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety, Jul. 15, 2025. arxiv: 2507.11473 (cs). [Online]. Available: https://arxiv.org/abs/2507.11473

[80] B. Baker, J. Huizinga, L. Gao, Z. Dou, M. Y. Guan, A. Madry, W. Zaremba, J. Pachocki, and D. Farhi, Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation, Mar. 14, 2025. arxiv: 2503.11926 (cs). [Online]. Available: https://arxiv.org/abs/2503.11926

[81] Concordia AI, State of AI Safety in China 2026, Concordia AI (aisafetychina.com), Jul. 2026. [Online]. Available: https://aisafetychina.com/