启明星辰
www.venustech.com.cn

大模型安全威胁框架

Ver1.0

启明星辰信息技术集团股份有限公司

2025年6月

一、 引言

近年来，以 ChatGPT、GPT-4、Claude、DeepSeek 等为代表的大语言模型(Large Language Models, LLMs)技术取得突破性进展，在全球范围内掀起 AI 发展热潮。大模型凭借其强大的学习、泛化和生成能力，推动了各行各业的深刻变革。在中国，大模型技术亦被视为国家战略性新兴产业的重要组成部分，众多科技巨头、创业公司、科研机构纷纷推出多元化应用场景的大模型产品与服务。大模型正逐步成为数字经济时代不可或缺的“新一代人工智能基础设施”，其影响力已渗透到社会经济的方方面面，预示着一个智能驱动的新时代的到来。

新兴技术的发展和广泛应用，往往伴随着新的安全挑战。尽管大模型展现出巨大的应用潜力，但由于其特有的技术特性与交互模式，也带来了传统信息安全之外的全新挑战。大模型具有复杂决策黑箱、数据深度依赖、内容边界模糊、能力动态演化等独有特性，使得大模型应用场景下的安全威胁呈现出新的维度与复杂性，对国家安全、社会稳定和企业发展构成严峻挑战。

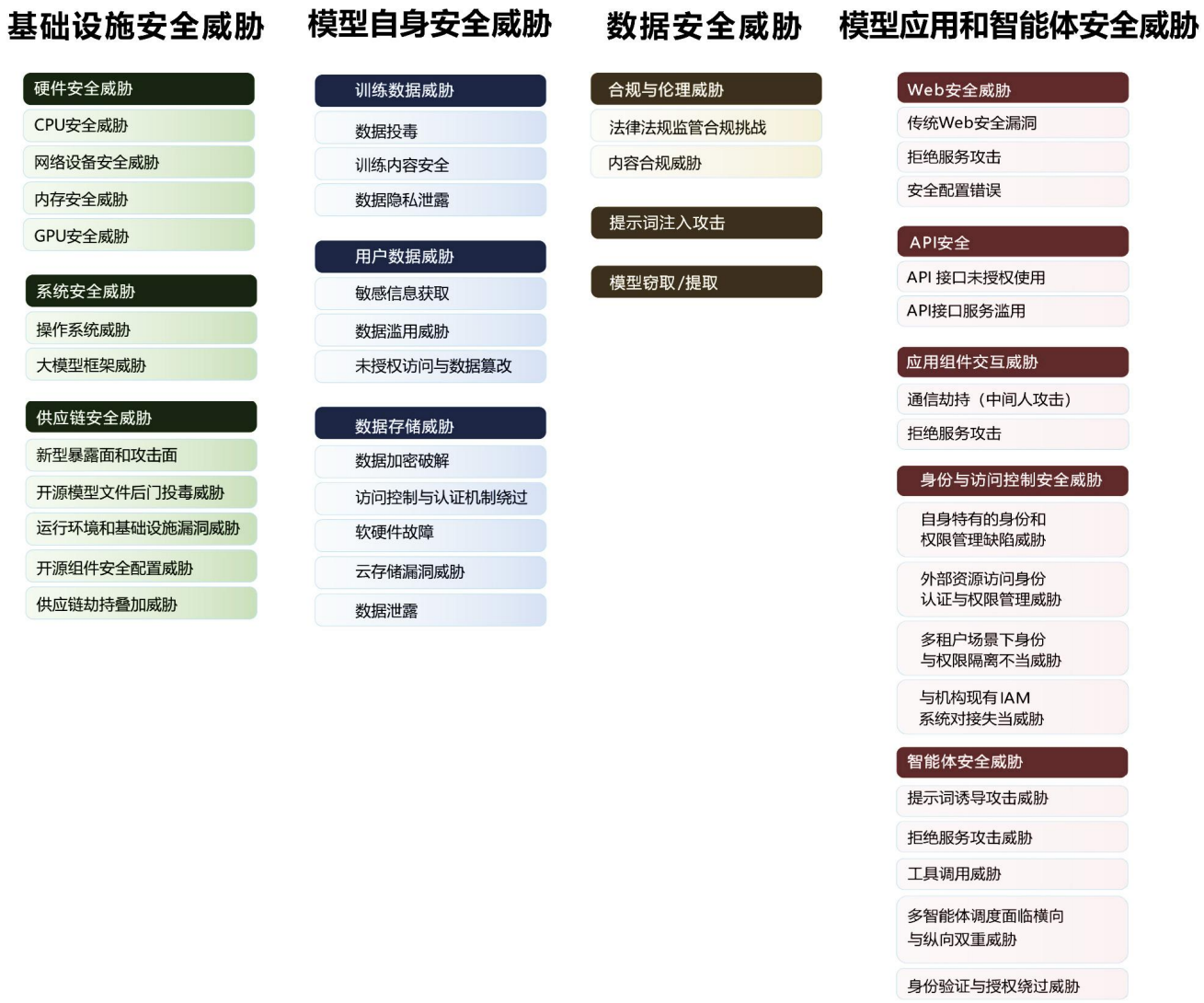
因此，构建一个全面、系统且具有前瞻性的大模型安全威胁框架，对于我们识别潜在风险、评估现有漏洞、制定有效的防御策略具有极其重要的意义。它将帮助企业和机构在享受大模型红利的同时，能够有针对性地部署安全措施，从而推动负责任的 AI 发展，确保大模型技术的可信、安全、可控。

鉴于上述背景，本文旨在从安全核心技术研究、实战化攻防技术研究和安全智库的多重视角出发，构建一个针对大模型应用场景的多维度威胁框架。文章将系统梳理并分析当下主流的大模型安全风险与威胁框架，并围绕大模型应用场景下的安全风险进行剖析，分析大模型从数据、模型自身、模型应用到合规伦理等各环节可能存在的威胁。

二、 大模型安全威胁框架

与传统的网络安全威胁相比，大模型应用场景下的威胁不仅继承了传统的安全风险，更衍生出了一系列与大模型自身特性和应用场景紧密相关的新型威胁。为系统性地理解和应对这些威胁和挑战，我们必须全方位审视大模型系统的各个关键组成部分所面临的潜在威胁。

如下图所示，本章节将从大模型基础设施、数据安全与隐私、模型自身安全及大模型应用安全等多维度出发，展开相关安全威胁初步探讨与分析。



大模型安全威胁框架图

2.1 大模型基础设施安全威胁分析

大模型的训练、推理和运行依赖于庞大的计算、存储和网络基础设施，这些底层组件的安全是整个大模型系统稳健运行的基石。

- **基础设施硬件安全威胁**

大模型运行所依赖的硬件资源（如 CPU、GPU、内存、网络设备等）一旦出现安全问题，将直接冲击大模型应用的稳定运行与数据安全。

CPU 安全威胁：从硬件核心来看，CPU 主要面临微架构攻击威胁、固件与后门威胁、硬件退化及威胁等。不仅可能通过利用微架构漏洞（如分支预测缺陷）可实现权限绕过、窃取缓存数据；还可因固件漏洞或预置的硬件后门实现对硬件的未授权控制，破坏大模型的稳定性和完整性；甚至通过恶意负载加剧硬件退化，导致硬件性能严重下降或损坏。

GPU 安全威胁：在 GPU 层面，尤其是在多租户共享的云环境中，攻击者可利用 GPU 缓存、功耗等物理特性实现共享资源的侧信道攻击，或利用 GPU 内存泄露漏洞、虚拟化漏洞，窃取模型参数与数据或其他租户的数据，破坏大模型应用运行的公平性和安全性；同时，驱动程序漏洞也为权限提升和远程代码执行提供了可乘之机，直接威胁模型的运算核心。

内存安全威胁：Rowhammer 攻击或恶意代码注入能修改内存中敏感数据，直接破坏数据一致性，进而影响大模型应用的可靠性；而内存泄漏、缓冲区溢出攻击及缓存侧信道攻击则为窃取模型权重或训练数据打开了通道。

网络设备安全威胁：作为支撑大模型基础运行环境的重要组成部分，网络设备面临固件漏洞威胁、中间人攻击威胁、带宽拒绝服务攻击威胁。其固件漏洞可导致设备被控，中间人攻击能拦截并篡改传输中的关键数据，而带宽拒绝服务攻击则能轻易瘫痪分布式训练和推理所依赖的数据通信

- **系统安全威胁**

大模型应用的运行依赖于操作系统、应用程序及底层框架。而攻击者一旦针对这些关键组件发起攻击，将对大模型应用的安全构成严重威胁。

操作系统威胁：大模型基础设施操作系统层面的威胁与传统操作系统面临的安全威胁基本一致，主要包括内核权限提升、系统组件漏洞威胁、恶意软件威胁等。

大模型基础框架威胁：大模型的运行依赖 TensorFlow、PyTorch、Numpy 等基础深度学习框架，其学习框架中存在的设计缺陷和漏洞，可能实现破坏大模型训练过程、窃取或篡改模型参数与训练数据甚至控制大模型应用的运行。

- **供应链安全威胁**

攻击者可对大模型供应链发起隐蔽性攻击，通过污染其依赖的第三方组件（如库、API、云服务或知识库），利用第三方组件漏洞或植入恶意代码（如伪装成正常更新的依赖包），导致模型运行异常、数据泄露或系统崩溃，威胁模型稳定性。主要面临如下威胁。

新型的暴露面和攻击面。大模型系统及其支撑组件在网络上开放了不必要的服务端口、管理接口或 API 端点，或者这些暴露面缺乏足够的身份认证、授权和安全加固措施，从而形成了薄弱的攻击面，容易被攻击 者扫描发现并利用来进行未授权访问、数据窃取、拒绝服务攻击或进一步渗透。比如，开源跨平台大模型工具 Ollama 存在未授权访问漏洞，攻击者可绕过身份认 证直接访问服务接口，访问获取业务数据、模型参数和配置文件等敏感数据。

开源模型文件面临后门投毒威胁。开源模型通常使用模型文件形式保存和分发，将带来大模型本地化部署或安装文件被仿冒并用于实施网络攻击的风险。攻击者通过“高频关键词搜索引擎投毒”、“仿冒网站”等方式， 诱导用户下载伪造的大模型本地部署工具包，传播含有木马病毒等恶意程序。用户环境一旦被植入木马，攻击者即可进一步控制用户服务器，窃取敏感信息、破坏系统数据，甚至入侵内部网络。同时，**用户被诱导安装仿冒的大模型应用 APP 造成敏感信息泄露**。攻击者从非官方渠道发布仿冒 DeepSeek 等 APP 安装包的手机木马病毒，

诱导用户安装 APP、获得高权限并阻止用户卸载，窃取用户在手机上存储或传输的敏感数据及个人信息，从而侵害用户个人隐私和经济利益。

大模型运行环境和基础设施漏洞利用的威胁。大模型多运行在云环境上，那么云环境自身的漏洞等安全问题将会影响到大模型的安全运行。例如 虚拟机/容器逃逸、虚拟机漂移、不安全挂载、镜像安全、云服务配 置管理安全等。除此之外，大模型底层基础设施的漏洞利用也带来巨大威胁。比如，在大模型开发过程中利用 CI(持续集成)/CD(持续交付部署)基础设施的漏洞在模型推送到生产环境的过程中实施攻击;在大模型部署过程中利用容器和集群系统漏洞执行恶意代码、窃取数据、干扰服务运行等，造成隐私信 息泄露;在 RAG 应用开发过程中利用向量数据库的漏洞获取未授权的数据、篡改数据、执行恶意代码或发起其他攻击，以获取敏感信息、远程操控恶意代码等。

开源组件安全配置问题带来的安全威胁。开源组件(如训练框架、模型服务接口)的默认配置常以易用性优先，忽略安全加固(如未启用身份验证、开放无认证的 API 端口等)。攻击者可利用默认配置的漏洞直接入侵模型服务，窃取数据、注入恶意负载或操控模型行为，导致数据 泄露、模型窃取、服务中断等问题。

如 Ollama 使用了 0.0.0.0 的默认 IP 设置、默认开放 11434 端口且无任何鉴权机制，攻击者可通过公网并利用该脆弱点访问模型的训练数据、参数和配置文件，从而窃取敏感信息或篡改模型行为。

供应链劫持带来的叠加威胁。攻击者通过篡改软件、硬件或服务的供应链 环节(开发工具、第三方库、更新包、外包组件、硬件、源码等)， 将恶意代码或后门植入合法产品中，从而在用户使用这些产品时实 施攻击。该方式利用了供应链中的上下游信任关系，使得恶意代码能够在未被察觉的情况下进入大模型的开发、训练或部署阶段，以“叠加威胁”的方式对大模型的安全性造成大范围的严重威胁。

2.2 大模型数据隐私与保护威胁

大模型的训练、微调及推理过程对数据高度依赖，其安全性贯穿于全生命周期，覆盖整个大模型应用场景的各个层级。从静态存储到动态交互，从模型构建的基石到用户输入的隐私，任何一环的数据安全失守，都将影响模型的可靠性和用户信任，甚至引发严重的法律与合规危机。

- **训练数据安全威胁**

在训练数据层面，威胁源自其采集、清洗和标注的全过程。“数据投毒”攻击，即在训练数据集中注入恶意的、有偏见的或错误的信息，能够从源头上诱使模型学习错误模式，导致其产生错误的推理、固化的偏见或危险的输出。未经充分脱敏的训练数据可能包含大量个人身份信息或受版权保护的内容，可能被数据提取或逆向还原出来，导致数据隐私泄露。而如果大模型训练数据清洗过程不够完善，并未完全过滤出包含色情、暴力、仇恨、敏感政治话题、隐私信息等违规内容，训练后的大模型则可能在运行过程中出现相关违规内容。

- **用户数据安全威胁**

大模型应用落地的过程中，会产生大量用户数据，用户在交互过程中输入的提示词、个人文件和敏感信息等数据。敏感信息可能因隐私保护不足或加密存储被侧信道攻击、反向推理等手段破解；攻击者可能突破权限限制非法访问或篡改数据；此外，用户数据存在未经授权收集、滥用（如商业推断或广告定位）的风险，甚至被恶意挖掘用于进一步攻击，这些行为不仅破坏用户信任，还可能违反相关法律法规（如 GDPR 等）。

- **数据存储安全威胁**

在数据存储环节，无论是训练数据集、模型权重文件还是用户日志，若存储系统（如云存储、数据库）存在访问控制不当、加密措施缺失或配置错误，都可能导致数据被未授权访问、批量下载或恶意篡改，内部人员的恶意行为同样是不可忽视的重大风险。

- **数据传输安全威胁**

在数据交互的动态过程中，模型与用户之间、以及分布式集群节点之间的数据传输，若缺乏端到端的强加密链路保护，极易遭受“中间人攻击”，导致通信内容被窃听、篡改或劫持。

2.3 大模型自身安全威胁分析

大语言模型作为 AI 系统的核心架构基础，无论是通用型亦或是垂直领域专用模型，均已成为企业核心战略资产。

- **合规与伦理威胁**

内容合规威胁：大模型在内容生成或信息处理过程中，可能无意或被恶意诱导地生成违反国家法律法规、社会公序良俗、行业规范或特定文化价值观的内容。这包括但不限于：涉及色情、暴力、恐怖主义、虚假信息、诽谤、侵犯知识产权、煽动分裂、种族歧视、宣扬邪教、侮辱英烈等各类非法或有害内容。

法律法规监管合规威胁：在数据处理与使用、知识产权等方面，大模型可能面临国家和地区层面法律法规和监管层的合规挑战。

- **提示词注入攻击**

攻击者通过构造巧妙的语言指令，能够欺骗模型、绕过预设的安全防护，诱使其执行非预期的恶意操作。一旦攻击成功，模型可能会泄露隐私数据、传播虚假信息或者执行非法操作，严重威胁数据安全、信息真实性以及系统的合法合规运行。尽管现有防御机制已经采用了输入过滤与对齐微调，但由于对抗性攻击的形式多样，系统在鲁棒性方面仍然存在明显短板，难以有效抵御此类威胁。

- **模型窃取/提取攻击**

攻击者通过未授权的非法手段获取模型权重参数，或通过系统性查询目标模型并解析其输

出来构建高相似度的替代模型。当大模型应用系统暴露了模型的输入输出接口或内部结构细节等信息时，攻击者即可利用这些信息构建定向化查询策略，通过大规模输入向量注入与输出记录，最终训练出具备功能等效性的克隆模型，威胁企业核心模型资产安全。

2.4 大模型应用和智能体安全威胁分析

AI 应用系统作为承载和展现 AI 能力的软件实体，除了面临 AI 模型特有的风险外，其自身作为一个应用程序，同样继承了传统软件开发中存在的各类安全漏洞。此类风险涵盖了常见的 Web 应用漏洞(如注入、XSS)、因不安全地处理 AI 模型输出而引入的新型漏洞(如 AI 生成内容直接执行导致的安全问题)，以及攻击者利用 AI 应用的业务逻辑缺陷进行欺诈或资源滥用。忽视 AI 应用自身的代码和逻辑安全，将使得 AI 系统的整体安全性大打折扣，即使其核心 AI 模型本身是安全的。

大模型应用和智能体也是一类应用系统，所面临的安全威胁是多方面的，既有常规应用的通用安全问题，也有人工智能应用和智能体的特有安全问题。大模型应用平台的前端和后端安全性直接关系到大模型应用安全，平台组件安全问题对大模型的实际应用场景构成多重威胁。智能体自身可能面临代码漏洞、配置错误等威胁，易被攻击者利用，导致行为异常或数据泄露，智能体组件间通信可能被篡改、伪造，破坏协同工作，如数据传输被截获篡改，影响决策准确性，多智能体协同时，个别智能体被恶意控制或行为异常，会干扰整体任务执行，甚至引发连锁反应，使整个系统陷入混乱。

- **大模型 Web 应用安全**

大模型应用作为一种人工智能应用系统，同样面临着传统 Web 应用所面临的各种已知安全漏洞，如 API 接口滥用、跨站脚本（XSS）、SQL 注入、跨站请求伪造（CSRF）、 失效的访

问控制、安全配置错误、使用含有已知漏洞的组件、不安全的 API 接口和拒绝服务攻击，这些漏洞可能被攻击者利用来窃取数据、破坏系统或获取未授权访问。

- **API 安全威胁**

大模型应用在调用外部 API(包括模型推理 API、数据服务 API、第三方工具 API 等)时，由于 API 接口本身的设计缺陷、认证授权机制薄弱、输入参数校验不足、返回数据处理不当或 API 网关配置错误等原因，导致应用受到攻击，如数据泄露、权限提升、拒绝服务或被用于进一步攻击其他系统。

主要威胁来自于 API 接口的未授权使用和 API 接口服务滥用。

API 接口未授权使用风险。主要表现为大模型接口组件因缺少 API Key 验证等访问控制、安全配置不当或使用默认配置导致 API 风险点位暴露而被滥用，造成大模型服务过载引发拒绝服务或数据泄露。比如 DeepSeek 因 Clickhouse 数据库的配置错误暴露百万条聊天记录和 API 密钥的日志泄露事件等。

API 滥用带来大模型算力消耗风险。攻击者编写脚本或者通过僵尸网络(Botnet)利用未限制的 API 接口执行高负载任务(如长文本生成、恶意爬虫高频调用等)，以每秒数百次的频率向模型 API 接口发送简单但高消耗的请求，模型被迫反复生成无意义长文本，占用大量 GPU 显存和计算资源，导致算力资源耗尽而拒绝服务，比如，某大模型曾遭遇每秒 2.3 亿次恶意请求，导致服务中断。

- **大模型应用组件交互威胁**

大模型应用的正常运行需要实现各部件之间的数据传输和通信，包括访问数据库、外部工具、计算资源等。若数据传输或通信过程中缺乏加密，攻击者可通过中间人攻击（MITM）截获、篡改 API 请求与响应，导致敏感信息被窃取或篡改。攻击者还可通过频繁调用 API 或系统资源导致关键 API 不可用或系统资源耗尽，实现拒绝服务攻击，使大模型应用系统陷入瘫痪状态。

- **大模型身份与访问控制安全威胁**

健全的身份验证和权限管理是保障大模型应用及其所访问资源安全的核心机制。根据启明星辰发布的《AI 就绪的大模型身份与访问管理白皮书（AI-R-IAM）》发现，大模型广泛应用引入了 AI 身份（AI Identity）和非人类身份（Non-Human Identity, NHI），同时，大模型的应用场景复杂多样，存在用户访问公域、私域大模型，大模型调用 RAG 知识库等，除了存在大模型应用自身身份和权限管理缺陷之外，相对于与传统的身份与访问管理 IAM 系统而言，由于在资产属性、内容、节点、细粒度以及集成对接方面的复杂性，同时存在着身份冒用、授权边界不清或授权过高等风险。

大模型应用自身特有的身份和权限管理缺陷威胁。在大模型应用中，涉及到模型训练、数据标注、内容审核、应用系统管理员等自身特有的多种用户角色，其权限模型设计不合理、权限划分不清、存在权限冲突或未能遵循最小授权原则将会导致某些角色拥有超出其工作职责所需的过多权限，或不同角色之间的权限边界模糊，从而可能被内部人员误用、滥用或被外部攻击者利用进行权限提升。

大模型应用访问外部资源的身份认证与权限管理威胁。大模型应用在程序逻辑中需要访问机构的 RAG 数据库、调用第三方 API、连接云服务或其他外部系统时，所使用的 API 密钥、数据库连接字符串中的用户名密码、服务账户令牌等身份凭证管理不当，比如，硬编码在代码中以弱加密方式存储、通过不安全的渠道传输，或者这些凭证被授予了超出大模型应用完成其功能所必需的过高权限，导致一旦凭证泄露或大模型应用本身被攻破，攻击者即可利用这些权限进行横向移动、窃取大量数据或破坏外部系统。

多租户大模型应用场景下的身份与权限隔离不当的威胁。在多租户架构的 大模型应用中(即多个独立的客户或用户群体共享同一套大模型基础设施和应用实例)，由于身份验证、授权机制或数据访问控制逻辑存在缺陷或配置不当，导致一个租户的用户或该租户的大模型应用实

例可能通过身份冒用、权限提升或逻辑漏洞等方式，非法访问其他租户的数据、模型配置或计算资源，从而破坏多租户间的安全隔离性。

与机构现有身份和访问管理 IAM 系统对接失当的威胁。大模型应用（尤其是私有化部署模式）与机构现有的身份与访问管理(IAM)系统进行集成时，由于集成配置错误、协议实现缺陷、信任关系管理不当或企业 IAM 系统本身的安全策略变更未能及时与大模型应用的身份权限管理，可能导致身份认证绕过、会话劫持、权限继承不当或权限提升等安全风险。

- **智能体安全威胁**

智能体作为一种典型的大模型应用形态，在自身、组件之间以及多智能体等方面面临多种安全威胁，比较通用共性的安全威胁主要是如下几种。

提示词诱导攻击威胁。攻击者通过精心设计的提示词（Prompt），诱导智能体调用恶意工具或执行恶意行为，从而达到攻击目的。攻击者可通过直接输入或间接输入（比如从网页、文档、数据库等外部数据源）特定提示词改变智能体行为，实现预期的安全措施绕过、敏感信息提取、恶意命令执行。该攻击利用了智能体自身的功能来实现恶意意图，具有极高的隐蔽性。

拒绝服务攻击威胁。攻击者通过诱使智能体陷入计算量极大的循环或进程，大量消耗计算资源，最终致使服务无法正常运行，严重破坏智能体生态的稳定性与可用性。

通信劫持（中间人攻击）。智能体的正常运行需要实现各部件之间的数据传输和通信，包括访问数据库、外部工具、计算资源等。若数据传输或通信过程中缺乏加密，攻击者可通过中间人攻击（MITM）截获、篡改 API 请求与响应，导致敏感信息被窃取或篡改，智能体行为被恶意操纵，影响智能体正常运行及数据安全。

工具调用威胁。智能体（LLM Agent）通过调用各类外部工具实现功能的拓展，如使用 MCP、访问数据库、执行脚本、进行 API 调用、RAG 和互联网搜索等，这一过程也带来诸多安全威胁。

多智能体调度面临横向与纵向双重威胁。横向层面，因缺乏中心化控制及可靠的信任传递机制，攻击者可伪造身份或操纵信任关系，干扰任务协作链（比如窃取数据、破坏资源调度秩序）。纵向层面，依赖关系引发级联攻击风险：核心智能体被攻陷后，恶意调度指令通过层级结构迅速扩散，导致攻击呈指数级蔓延（比如操控决策引发系统级崩溃）。这两类威胁的共性在于利用了横向暴露信任机制脆弱性的系统拓扑缺陷，纵向放大了单点故障危害。

身份验证与授权绕过威胁。攻击者利用弱身份认证、令牌泄露或权限管理漏洞，通过暴力破解或凭证伪造手段冒充合法身份，将导致横向越权访问（突破系统间身份隔离）和纵向权限提升（超越功能操作边界），进而窃取敏感数据、操控智能体执行恶意活动（如篡改任务链、发起寄生攻击），最终侵蚀整个系统的访问控制体系并引发持续性安全风险。核心问题在于认证机制缺陷与权限边界失效，为攻击链形成提供了完整路径。

三、 大模型安全防御体系

由于大模型自身机制原因，导致数据、模型和应用等多层面的安全问题，仅仅通过单一手段难以有效控制风险，因此，需要针对大模型应用场景下的多样化威胁，以确保大模型技术在安全可控的前提下持续创新与发展。

需要从技术、管理、法规和生态多个层面构建综合防御体系。在技术层面，防御策略应覆盖大模型运行的各个关键环节。对于基础设施安全，核心在于强化计算、存储和网络资源的物理与环境防护。加强数据安全与隐私保护，强化提示词工程，加强 API 安全与实施身份认证与授权防护，结合传统网络安全防护技术在大模型应用场景下进行保护。大模型厂商应重点提升模型的鲁棒性以抵御对抗性攻击。

除了技术保障，应建立全面且动态的风险评估与治理体系，定期进行风险评估和审计，明确各方在开发、部署和运营中的安全职责。从更宏观的法规层面看，政府应持续推动大模型相

关法律法规的制定与完善，强化行业标准建设，制定大模型安全技术标准、评估规范和伦理准则，为行业发展提供清晰的规范指引。

此外，应积极构建开放、协作的安全生态，鼓励安全厂商、研究机构、高校以及开发者社区等各方紧密合作，共同研发大模型安全技术和工具，推动知识共享与能力互补。建立健全信息共享机制，及时共享大模型安全威胁、漏洞和最佳实践。

四、 结语

本文对大模型应用场景下的威胁框架进行了初步的探索与分析， 随着大模型技术的迅速发展和应用，其安全问题正在演变为一个复杂体系。在未来的智能世界中，我们需要以审慎的态度和前瞻的视野，逐步深入构建一个全面化、结构化、系统化的威胁框架，这不仅是理解和应对这些复杂威胁的基础，更是保障大模型健康可持续发展的关键一步。

五、 参考资料

1、《AI 就绪的大模型身份与访问管理白皮书（AI-R-IAM）》.启明星辰 2025 年 2 月.

2、《AI 安全风险评估和控制指南》Demon AI Security Handbook 2025 年 5 月.

3、《DeepSeek 等开源大模型应用 安全防护要求与实施指南》.中国移动 2025 年 4 月.

4、OWASP Top 10: LLM & Generative AI Security Risks https://genai.owasp.org.

附件：现有大模型安全框架分析

随着人工智能技术的飞速发展及其应用的日益深化，特别是随着大语言模型（LLMs）技术的崛起，全球范围内对 AI 安全和潜在风险的关注已从学术前沿走向产业实践，并上升至国家战略层面。各国政府、国内外研究机构、行业组织、技术厂商，都已认识到构建系统性风险与威胁框架的紧迫性，并开始探索和构建相关的风险与威胁框架。

本章节旨在分析当前已有的主要框架，对其特点、优势与局限性进行剖析，以期为后续提出的大模型威胁框架提供全面的背景认知和对比参照。

1. 国际和国家层面的综合性框架与标准

国家层面的框架通常是由政府机构、国际标准化组织或行业联盟发布的综合性/通用性框架与标准，适用于 AI 系统的全生命周期，并覆盖多种风险类型。一般具有宏观性、政策导向性和强制性，旨在从国家战略高度保障 AI 安全。

- **NIST 人工智能风险管理框架（AI RMF，NIST AI Risk Management Framework）**

《人工智能风险管理框架》（AI RMF 1.0）由美国国家标准与技术研究院（NIST）于 2023 年 1 月正式公布，旨在帮助组织管理人工智能生命周期中的风险。框架分为两部分六章节，概述了与 AI 相关的风险及可信 AI 系统的特征，提出了治理、映射、测量和管理四大核心功能。

该框架提供了一个高层次、通用的风险管理方法，其原则和指导方法具有较强的普适性和指导性，可广泛应用于大模型领域；但作为一个通用框架，NIST AI RMF 更多关注风险管理的流程和原则，而非详细的技术威胁类型，大模型特有的攻击与威胁相关技术细节描述较少。它需要与具体的技术威胁框架相结合，才能更好地指导大模型领域的实践。

- **欧盟人工智能法案（EU AI ACT）**

欧盟《人工智能法案》（EU AI ACT）是全球首部针对人工智能的综合性法律框架，于 2024 年 5 月通过并于同年 8 月生效、2025 年 2 月起分阶段实施。该法案采取基于风险的分级监管模式，将 AI 系统分为四个等级；对 AI 系统的全生命周期施加了明确的法律责任，对不同风险水平的 AI 系统提出不同程度的要求，包括高风险 AI 系统的严格合规性要求。

- **ISO/IEC 42001（人工智能管理系统）**

国际标准化组织（ISO）和国际电工委员会（IEC）发布的第一个 AI 管理系统国际标准。为组织建立、实施、维护和持续改进人工智能管理系统（AIMS）提供规范性指导，适用于提供或使用 AI 产品或服务的任何规模组织，以规范企业负责任地开发或使用 AI 系统。

- **IEEE AI 伦理指南（IEEE AI Ethics Guidelines）**

IEEE 全球自主和智能系统伦理倡议的一部分，提供了一系列伦理原则和指南，以促进 AI 的负责任发展。

- **中国人工智能安全治理框架/原则**

中国也在积极构建自己的人工智能治理体系，例如《新一代人工智能发展规划》中提出的“三步走”战略目标、《生成式人工智能服务管理暂行办法》等法规及《人工智能安全治理框架》等框架标准。强调“以人为本、智能向善”的发展方向，关注数据安全、算法透明、隐私保护等。

《生成式人工智能服务管理暂行办法》于 2023 年 8 月起施行，旨在促进生成式人工智能健康发展和规范应用，维护国家安全和社会公共利益，保护公民、法人和其他组织合法权益。法规提出对生成式人工智能服务实行包容审慎和分类分级监管，从技术发展与治理、服务规范、

监督检查和法律责任等方面服务提供者和使用者提出了一系列要求，其中对数据侧的要求尤为突出，以推动生成式人工智能技术发展的合法合规。

《人工智能安全治理框架 1.0》于 2024 年 9 由全国网络安全标准化技术委员会发布。该框架提出了包容审慎、风险导向、技管结合、开放合作 4 大治理原则，将人工智能安全风险分为内生安全风险和应用安全风险，涵盖模型算法、数据、系统等方面以及网络、现实、认知、伦理等领域。

2. 研究机构与行业组织层面的技术框架

网络安全组织或研究机构提出的框架往往更侧重于理论深度和攻击技术分类等方向，更深入地关注 AI 系统的具体技术性风险，包括攻击手段、漏洞利用和防御策略。

- **OWASP Top 10 for LLM Applications 2025**

由开放式 Web 应用安全项目（OWASP）发布的针对 LLM 应用的安全风险列表，探索生成式人工智能和大语言模型应用程序在开发、部署和管理生命周期中的最新十大风险、漏洞及缓解措施，以实现应用程序的开发和安全保障。该框架更多聚焦于应用层面，从应用安全的角度切入，重点关注 LLM 应用中常见的攻击手段和漏洞，主要包括提示注入、敏感信息泄露、供应链安全、数据与模型投毒、系统提示泄露等。

- **MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)**

ATLAS 是由非营利组织 MITRE 发布的针对 AI 系统的对抗性威胁框架，旨在系统地、结构化地识别和描述针对 AI 系统的攻击技术和策略，为网络安全专业人员提供威胁情报和防御指导。该框架侧重于技术性风险与细节，参照 MITRE ATT&CK，对 AI 系统面临的对抗性攻击技术进行分类组织，详细描述了可能采取的策略、技术和程序（TTPs），涵盖从侦察、资源开发到执行攻击等多个阶段。

- MIT AI Risk Repository

这是一个由麻省理工学院（MIT）维护的 AI 风险数据库，收集了大量来自现有框架和分类的风险，并提供因果分类和领域分类，为研究人员、开发者、政策制定者等提供了一个共同的参考框架。

3. 技术厂商层面的实践框架

具有丰富技术背景的厂商，侧重于基于其产品开发和实际运营经验，将安全理念融入具体技术架构、产品设计和日常运营中，以应对前沿威胁。

- **微软的 AI 安全框架**

微软作为 AI 技术的重要开发者和使用者，其内部框架着重于其在 AI 研发、部署和运营过程中遇到的实际技术风险。发布责任 AI 准则，逐步形成负责任 AI 标准和与治理框架，指导开发团队如何依据 AI 原则构建 AI 系统，确保 AI 系统的设计、开发和部署符合道德和法律要求，及 AI 模型的安全可信部署。同时构建了 AI 安全风险评估框架，明确 AI 系统的关键资产，并针对 AI 资产的独特威胁（数据泄露、模型窃取、对抗攻击和提示注入攻击等）开展威胁建模，并定期评估 AI 系统的漏洞。

- **Google 的 AI 安全框架**

SAIF（Security AI Framework）是 Google 于 2023 年发布的一个全面的、通用的 AI 安全框架，旨在帮助任何组织在 AI 系统设计、开发、部署和运营的整个生命周期中，系统性地整合安全和隐私措施。 该框架关注整个 AI 生态系统的安全性，包括数据、模型、基础设施和应用程序层面的风险，并强调如何通过风险管理和安全默认原则，构建可信赖的 AI 系统。

Google DeepMind 的前沿安全框架 FSF（Frontier Safety Framework）则是 DeepMind 专门为其最前沿的、高能力生成式 AI 模型量身定制的深度安全协议和评估方法。 与 SAIF 的通用性不同，FSF 更聚焦于应对这些高级模型可能带来的独特且更严重的潜在风险，包括误用、模型与人类价值观不对齐、以及可能导致事故或深远社会影响的结构性风险。

- **OpenAI 的 AI 安全框架**

OpenAI 作为 LLM 和生成式 AI 的头部厂商，发布了“Preparedness Framework”，专注于解决其最先进 AI 模型的潜在灾难性安全风险，包括评估、监测和缓解风险。该框架通过明确高风险能力的优先级标准，并将其分为不同能力阈值，从而在模型开发和部署前进行严格的评估和红队测试。此外，他们还提出了“模型规范（Model Spec）”等拟议框架，旨在规范 AI 模型（如 GPT-4）的未来响应方式，关注内容规范和伦理准则。

- **Anthropic 的负责任扩展政策（Responsible Scaling Policy，RSP）**

RSP 是 Anthropic 提出的一个系统性的风险治理框架，旨在管理前沿 AI 系统的灾难性风险，包括故意滥用（如制造生化武器）和模型自主行动引发的破坏。该框架基于比例保护原则，通过 AI 安全等级（ASL）实现分层管控；引入能力阈值和多层防御架构，还强调安全案例方法论和外部专家反馈。

尽管业界和学界已在大模型安全风险与威胁等方面开展探索并提出了一些初步的框架，但鉴于大模型技术的飞速发展及其带来的复杂性，这些现有框架各有侧重，在涵盖范围、系统性、针对性都各有优劣。业内对大模型风险与威胁的分类还未形成共识，难以形成统一的行业规范与标准。随着 AI 技术的发展，这些框架也在不断演进和完善，组织通常需要结合多种框架来全面评估和管理 AI 系统的风险。