大语言模型威胁分类

AI Controls Framework
Working Group

CSA GCR cloud security alliance®
GREATER CHINA REGION

人工智能控制框架工作组的永久和官方访问地址为：

https://cloudsecurityalliance.org/research/working-groups/ai-controls

联盟简介

云安全联盟（Cloud Security Alliance,CSA）是中立、权威的全球性非营利产业组织，于2009年正式成立，致力于定义和提高业界对云计算和下一代数字技术安全最佳实践的认识，推动数字安全产业全面发展。

云安全联盟大中华区（Cloud Security Alliance Greater China Region, CSA GCR）作为CSA全球四大区之一，2016年在香港独立注册，于2021年在中国登记注册，是网络安全领域首家在中国境内注册备案的国际NGO，旨在立足中国，连接全球，推动大中华区数字安全技术标准与产业的发展及国际合作。

我们的工作

联盟会刊下载地址
了解联盟更多信息

加入我们

CSA大中华区官网
(https://c-csa.cn)

点击会员

加入联盟

填写相关申请信息

成为CSA会员

JOIN US

致谢

报告中文版支持单位

中国移动云能力中心，注册名称为中移(苏州)软件技术有限公司，是中国移动通信集团于 2014 年 3 月在苏州成立的全资子公司。云能力中心主要承担中国移动云的技术研发、资源建设、业务运营、服务支撑等相关业务，自研了弹性计算、数据库、云存储、云网络、云安全等 200 余款覆盖云计算全产业链的产品，产品丰富度行业第二，为全国的工业、政务、医疗、教育、交通、金融等行业提供云计算、大数据及各类信息化解决方案。

参与本次报告的专家：

王浩硕，云能力中心安全产品部负责人

赵玲玲，云能力中心安全产品部研发管理专员

付怀勇，云能力中心安全产品部产品经理

李雨含，云能力中心安全产品部产品经理

薛四青，云能力中心安全产品部软件开发工程师

吴云飞，云能力中心安全产品部综合管理专员

中国移动云能力中心作为云安全联盟大中华区（CSA GCR）的理事单位，为该报告的翻译工作提供了必要的支持。这种支持并不涉及对 CSA 在研究内容开发和编辑方面的决策权和控制权，确保了 CSA 在这些核心领域的独立性和自主性。

报告英文版编写专家

主要作者

SiahBurke
MarcoCapotondi
DanieleCatteddu
KenHuang

贡献者

Marina Bregkou
Sanitra S. Angram
Vidya Balasubramanian
Avishay Bar
Monica Chakraborty
Ricardo Ferreir
Anton Chuvakin
Alessandro Greco
Krystal Jackson
Gian Kapoor
Kushal Kumar
Ankita Kumari
Yutao Ma
Danny Manimbo
Vishwas Manral
Jesus Luna
Michael Roza
Lars Ruddigheit
Dor Sarig
Amit Sharma
Rakesh Sharma
Kurt Seifried
Caleb Sima
Eric Tierling
Jennifer Toren
Rob van der Veer
Ashish Vashishtha
Sounil Yu
Dennis Xu

审稿人

PhilAlger
IlangoAllikuzhi
BakrAbdouh
VinayBansal
VijayBolina
BrianBrinkley
AnupamChatterjee
JasonClinton
AlanCurran
SandyDunn
DavidGee
ZackHamilton
VicHargrave
JerryHuang
RajeshKamble
GianKapoor
RicoKomenda
VaniMittal
JasonMorton
AmeyaNaik
GabrielNwajiaku
MeghanaParwate
PrabalPathak
RuchirPatwa
BrianPendleton
KunalPradhan
Dr.MattRoldan
OmarSantos
Dr.JoshuaScarpino
NataliaSemenova
BhuvaneswariSelvadurai
JamillahShakoor
TalShapira
AkramSheriff
SrinivasTatipamula
Maria(MJ)Schwenger
MahmoudZamani
RaphaelZimme

序言

在人工智能技术迅速发展的背景下，大语言模型（LLM）已成为推动技术创新和业务转型的核心力量。其在自然语言理解、生成与处理方面的强大能力，正深刻改变着人与信息、技术的互动方式。然而，随着 LLM 应用的广泛推广，伴随而来的是一系列复杂的风险和挑战，尤其是在安全性、隐私保护和合规性方面，全球范围内的应对压力愈加凸显。

为应对这些挑战，CSA 大中华区发布《大语言模型威胁分类》报告，旨在为行业提供一个全面的风险管理框架，帮助各行业识别、评估和管理 LLM 应用过程中可能遇到的风险。本报告详尽地分析了 LLM 的关键资产、服务生命周期、影响类别和威胁类别，为政策制定者、技术专家和行业决策者提供了一个清晰的理解和应对 LLM 相关风险的框架。报告的主要内容包括 LLM 资产的分类，详细描述了从数据资产到模型参数的各个方面；LLM 服务生命周期的管理，涵盖了从准备到退役的各个阶段；以及 LLM 服务的影响和威胁类别，包括数据泄露、模型操纵、供应链安全等关键领域。

希望这份报告能成为 LLM 风险管理和安全控制领域的关键参考资料，帮助各界在应对 LLM 带来的技术挑战时做出更明智的决策，为未来的研究、政策制定以及行业发展提供坚实的理论支撑。

李雨航 Yale Li
CSA 大中华区主席兼研究院院长

目录

目标与范围

本报告由云安全联盟(CSA)人工智能(AI)控制框架工作组基于 CSA AI 安全计划所撰写。它为与大语言模型(LLM)的风险场景和威胁相关的关键术语建立了通用的分类和定义。本报告的撰写目的是希望能够提供一个共享的语言和概念框架以供业界沟通交流，并指导其在 CSA AI 安全计划下进行更多研究。具体来讲，这项工作旨在协助 CSA AI 控制工作组和 CSA AI 技术风险工作组做出更多的努力。

报告的重点内容如图 1 所示：

- LLM 资产
- LLM-服务生命周期
- LLM-服务影响类别
- LLM-服务威胁类别

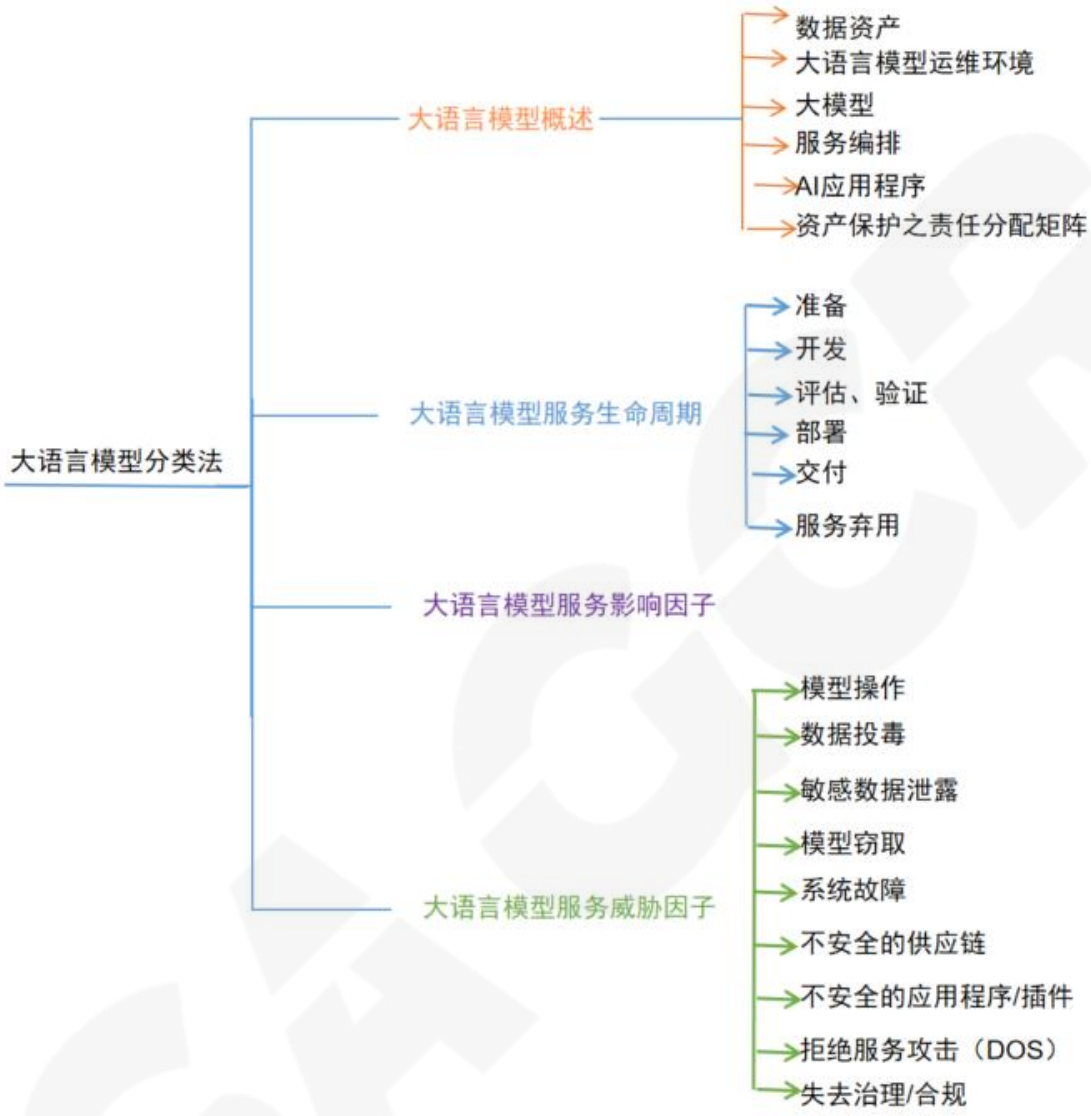

图 1：CSA LLM 威胁分类法

本报告所采纳的定义与分类体系是建立在现有文献基础之上，通过工作组成员与组长充分探讨后所形成的。通过上述过程，形成了广泛共识，建立了一套可以指导共同工作的通用术语。

报告从文档末尾引用的众多行业参考文献中汲取灵感，特别是 NIST AI 100-2 E2023 的“Adversarial Machine Learning: A Taxonomy andTerminology of Attacks and Mitigations” [Barrett et al., 2023]。

有了这些定义和分类，针对评估人工智能威胁和风险、制定适当的控制措施以及管理负责任 AI 开发，可以使不同的 CSA 工作组和项目成员之间取得更高的目标一致性。建立通用术语的目的在于其可以避免语义上的混淆，增加相关概念上的连贯性，以保证对话精确性。本报告将关键术语统一汇总起来，也是为了能够向 CSA AI 控制工作组和 AI 技术风险工作组提供对于 AI 安全计划的统一范围。

与 CSA AI 控制框架的关系

CSA AI 控制框架工作组的目标是定义一个控制目标框架，以支持各机构能够在安全、负责任的原则下对 AI 技术进行开发、管理和使用。该框架将有助于评估生成式 AI (GenAI) 特别是 LLM 相关的风险。

本报告所定义的控制目标将涵盖与网络安全有关的各方面。此外，它还涵盖了与网络安全相关的安全、隐私、透明度、问责制和可解释性等方面的内容。可回顾 CSA 前期的博客文章 AI Safety and AI Security，我们对 AI Safety 和 AI Security 之间的异同点进行了充分探讨。

CSA 控制框架着眼于 B2B 的影响，与政府在保护国家安全、公民权利和法律执行方面的努力相辅相成，倡导符合全球标准和法规的安全且道德的 AI 应用程序。

1.大语言模型概述

本章节定义了实施、管理大语言模型（LLM）系统所必需的基础组件，涵盖了从模型训练、微调过程中重要的数据到确保 AI 系统完美部署和操作的复杂运维环境。同时，本章节也对 LLM 的必要性、架构、性能和优化技术作出了进一步阐述（详见图 2）。除上述内容外，本章节还探讨了资产保护的注意事项，此外，本节还探讨了资产保护过程中较为重要的几点，利用 RACI 责任分配矩阵（谁负责、谁批准、咨询谁、通知谁）来明确了开源社区组织在实施 AI 服务时的责任划分。

图 2：大语言模型资产

1.1. 数据资产

在 LLM 服务中，多种资产内容在塑造服务的效能和功能方面起着不可或缺的作用。数据资产是这些资产中最为重要的，它是 LLM 运行不可或缺的基石与首要驱动力。以下是构成 LLM 服务的典型资产范围列表：

- 用于训练、基准测试、测试和验证的数据
- 用于微调训练的数据
- 用于检索增强生成（RAG）的数据
- 定义使用中数据的元数据的数据卡片
- 输入数据
- 用户会话数据
- 模型输出数据
- 模型参数（权重）

- 模型超参数
- LLM 系统的日志数据

以下是对这些资产的定义：

1. **用于训练、基准测试、测试和验证的数据：**这包括用于训练、基准测试、测试和验证模型的数据集，由文本来源组成，模型从中派生出对语言模式和语义的理解，这对打造模型最终的效果是至关重要的。每个数据元素都被单独创建并管理。

2. **微调训练数据：**模型经过初始训练后会用到一些额外的数据对其进行微调或进一步的预训练，这有利于调整模型的参数，使其更紧密地与特定用例或领域保持一致，从而增强模型适应性和准确性。

3. **检索增强生成（RAG）：**即将外部知识库与大语言模型集成到一起。模型在生成响应之前会检索外部知识库中的相关信息，而 RAG 便是可以将模型内部资源和外部资源有效连接起来的一种手段。RAG 可以通过内部系统、公开资源等各种渠道检索相关信息，如互联网，通过扩展输入内容及提炼模型的上下文语义，便可使模型生成更为精准的反馈。

4. **数据卡片：**数据卡片是用作管理大模型所用到的各种数据集的。它有助于 AI 数据进行管理，并记录了每个所使用的数据集的来源、流程轨迹、所有权归属、数据敏感度和合规性检查等相关的信息。随着数据内容、所有权归属以及需求的变化，对数据卡片进行适应性更新以确保数据的合规性和可见性也是十分重要的。

5. **输入数据（系统级提示）：**输入数据是指提供给大语言模型的用于确定上下文和边界的内容。在生成式对抗技术背景下，这类数据还可用于对模型主题边界范围进行设置。

6. **用户会话数据：**是指在用户与 AI 系统互动过程中所收集的信息，包括所输入的查询语句、模型最终生成的反馈以及用户在使用过程中所补充的内容，收集该类信息能够加强模型更为人性化的互动。

7. **模型输出数据：**是指模型收到输入指令后的结果反馈，包含文本型内容、用户预期的数据形式，输出数据反映了模型的理解和推理能力。

8. **模型参数（权重）：**是指模型在训练过程中调用的系统内部参数或权重，这些参数或权重的设定会对模型的行为产生重要影响，进而影响模型响应的能力。

9. **模型超参数：**是指在模型训练期间所指定的配置参数，包括学习率、批大小或架构选取等参数，这些参数在塑造模型的整体性能和行为方面至关重要。

10. **日志数据：**是指模型在运行过程中记录其各种事件响应和交互行为的数据。包括所输入的查询语句、模型最终生成的反馈、模型性能指标以及发生的任何异常行为，这对监控和完善模型的功能和性能至关重要。

1.2. 云上大语言模型运维环境

LLM 运维环境是指部署及操作大语言模型所涉及的基础设施和流程。以下是与此环境相关的专业定义：

- 运行训练环境的云
- 运行模型推理点的云
- 运行 AI 应用的云环境
- 混合和多云基础设施
- 部署环境的安全性
- 持续监控
- 云托管训练数据（存储）

以下是这些专业名词在大语言模型运维框架中的具体描述及作用：

1. **运行训练环境的云：**该环境中纳管了由云平台或云服务提供商所提供的模型训练过程中所需的计算资源、存储设施、辅助基础设施等，这些底层资源对于训练 LLM 是至关重要的。它决定了模型性能提升的上升空间。

2. **运行模型推理的云：**该环境中纳管了由云平台或云服务提供商所提供的模型训练和模型推理过程中所需的计算资源、存储设施、辅助基础设施等，该环境下运行的 LLM 能够根据用户输入生成预期响应，确保交互连贯性。

3. **运行 AI 应用的云（公有云/私有云/混合云）：**该环境中纳管了由云平台或云服务提供商所提供的运行 AI 应用程序或 AI 服务所需的基础设施，它们会使用已训练好的大语言模型的能力。它类似于一个服务中心，利用模型的推理能力向最终用户提供增值服务。

4. **部署环境的安全性：**指的是一系列简单的访问机制和政策，来规范外界对大模型各组件的访问。包括身份和访问管理(IAM)协议和网络安全措施，以保护关键资产和功能的完整性和隐私性。

5. **持续监控：**指的是对大模型运维环境的性能、安全态势、整体状况持续监测，包含了对训练环境、推理环境和应用组件的监测，确保模型在最佳状态下运行，同时也会及时识别和纠正可能出现的任何异常或问题。

6. **云托管训练数据（存储）：**指的是云平台或云服务提供商所提供的用于模型训练的大量数据集，它需要强大的存储和数据管理能力，以适应庞大和多样化的数据类型，这些数据集对于训练和完善模型性能是至关重要的。

1.3. 模型

在机器学习领域中，“模型”一词指的是能够通过训练来预测或执行特定任务的数学表示或算法。

所选取的模型架构、所采用的微调方法以及使用的开源或闭源的模型都将会大大影响大语言模型在特定领域中的性能、效能以及灵活性。

我们在以下小节中对基础模型资产进行了定义：

- 通用大模型

- 微调模型
- 开源与闭源模型
- 特定领域模型
- 模型卡片

1.通用大模型：

通用大模型是大模型发展的起点。通常指的是大型、预训练的语言模型，它们借助自我监督学习技术对大量未标记的数据进行学习，以获得对自然语言的泛化理解。通常，通用大模型为后续的微调模型和专业化模型奠定了基础，以满足特定的任务或领域的需要。对于一些先进和创新的基础模型，我们可以使用“前沿模型”这一术语来表示 AI 领域中的一个全新基础模型。从 AI 的角度来看，有时“基础模型”一词表示的是应用技术堆栈中的基础模型。

2.微调大模型：

微调模型是从通用大模型演进而来，经过改进和调整后能够适应特定任务或领域。微调过程，是利用了有监督学习技术和有标记的特定数据来调整通用模型的参数。这一迭代过程使我们的微调模型能够在保持原通用模型基础知识和能力的同时，增强其在特定任务或领域上的适用性和完整度。

3.开源与闭源大模型：

开源大模型与闭源大模型的主要区别在于对模型源代码、模型权重以及相关组件访问的许可。开源模型可能会公布它们部分或全部的训练数据、源代码、模型开发所用的数据、模型架构、权重和工具，以开源许可证的形式向大众开放，允许在特定条款和条件下免费使用。然而，闭源模型保持了私有化状态，不会向大众公开其源代码、模型权重和技术细节，这通常出于保护知识产权或商业利益的目的。允许用户访问模型进行微调或推理的闭源模型称为开放型访问模型。

以上这些模型共同构成了模型开发的支柱，促进了生成式 AI 的创新性、适应性和可访问性。

4.特定领域大模型：

特定领域大模型指的是能够在特定领域如金融、医疗、代码开发等表现出优异性能的机器学习模型。

5.模型卡片：

模型卡片是用来描述大模型特征的。它是一种维护大模型文本的文件，对于管理和确保 AI 模型的正确性是至关重要的。模型卡片包括模型文本的细节，如所有权、性能特征、模型训练的数据集、训练顺序等。这也有助于用户追溯、沿袭、理解模型的行为。模型卡片需要随着文本元数据的变化不断维护和更新。[CSA, 2024]

更多关于模型卡片的详细信息可以在 Hugging Face 平台上获取到，该平台是一个专门的机器学习社区，能够提供大模型、数据集和应用程序给学习者使用。

1.4. 服务编排

这类服务包含了一系列组件和功能，能够实现大语言模型的高效性和安全性。

以下是服务编排过程中可能涉及资产清单：

- 缓存服务
- 安全网关（LLM 网关）
- 部署服务
- 监控服务
- 优化服务
- 安全插件
- 自定义和集成插件
- 大语言模型通用代理

以下是各类编排服务资产的定义及其重要性阐述：

1.缓存服务：

缓存服务指的是一种系统或组件，它们通过缓存模型生成的响应、输入数据或其他数据来提高大模型的请求效率和性能，减少冗余计算。通过缓存服务临时存储频繁访问的数据，有助于降低请求响应时间并且减轻大语言模型的计算压力。

2.安全网关（LLM 网关）：

安全网关，也称为 LLM 网关，是作为大语言模型和外部系统交互的中间件。这些网关通过实施访问控制、输入验证、过滤恶意内容（如提示注入攻击）与个人/隐私信息的举措，防范潜在威胁或滥用，来确保 大语言模型所处理数据的完整性和机密性。

3.部署服务：

部署服务简化了大语言模型在不同环境（包括云平台和本地基础设施）中的部署和扩展。这些服务能够使部署过程自动化，促进版本管理，并优化资源分配方案，以确保大语言模型的高效和无缝部署。

4.监测服务：

监测服务在监督大语言模型的安全性、性能、安全状况和使用情况方面至关重要。这些服务使用监控工具和技术收集实时信息，来检测异常、滥用（例如提示注入攻击等）并发出警报，从而实现系统安全性，通过主动维护和及时干预，以维持大语言模型的最佳运行状态。

5.优化服务：

优化服务旨在优化大语言模型的性能和资源利用率。这些服务采用一系列技术，如模型量化、剪枝、高效的推理策略，来提高大语言模型的效率，降低计算开销，并保证在不同的部署场景下整体性能的提升。

6.安全插件：

安全插件是通过提供数据加密、访问控制机制、威胁检测能力和合规性强制措施来扩展大语言模型安全性的一种组件，能够增强网络弹性。

7.自定义和集成插件：

通过引入自定义和集成插件，大语言模型可以实现行为的个性化定制并无缝对接各类系统、应用程序和数据源。这些插件使模型更具灵活性，不仅能够根据特定场景或行业需要调整模型的功能，而且还提升了与既有架构的兼容性，增强了大语言模型部署的多功能性和实用性。

8.大语言模型通用代理：

大语言模型通用代理是指与大语言模型协作以增强模型功能和性能的智能代理或组件。这些代理可用来执行各种任务，如

- 规划
- 映射
- 功能调用
- 监测，
- 数据处理，
- 可解释性，
- 优化，
- 扩展和协作，
- 增强大语言模型部署在不同操作环境中的多功能性和适应性。

1.5. AI 应用

AI 应用已经变得无处不在，渗透到我们日常生活和商业运作的各个方面。从内容生成到语言翻译，由大语言模型驱动的 AI 应用已经彻底改变了行业，重塑了我们与信息技术交互的方式。然而，随着 AI 应用的普及，一个有效的控制框架用以指导 AI 应用开发、部署和使用成为我们迫切要解决的问题。

AI 应用使技术创新走向巅峰，提供了多种满足不同商业领域和场景的能力。这些应用程序利用大语言模型强大的功能来编译和处理自然语言输入，实现了内容生成、智能问答、情感分析、语言翻译等功能。本质上来讲，AI 应用程序是用户与底层大语言模型间实现智能交互的接口，促进了不同领域间无障碍交互和任务自动化。

而作为大语言模型的下游应用，AI 应用程序是整体 AI 控制框架中最为重要的资产之一。它们是模型与最终用户连接的直接触点，影响用户对 AI 系统的感知和交互方式。因此，AI 应用的好坏会一定程度上放大用户对大模型优劣评价。

AI 应用也可能对经济产生重大影响。随着企业日益依赖 AI 应用以推动创新、简化运营并获取竞争优势，负责任的应用开发与部署对于维护市场完整性和促进公平竞争愈发重要。鉴于上述考虑，AI 控制框架必须以 AI 应用的治理和监督为首要考虑要素。这包括了为 AI 应用开发、测试、部署、运营和维护建立指导方针和标准，确保各项环节符合法律法规，并保持在 AI 应用全生命周期的透明度和问责制。除了达成上述既定目标，AI 控制框架还应聚焦于对 AI 应用进行持续的监测和性能评估，旨在及时识别潜在风险并避免或减少意外发生。

通过在 AI 控制框架中优先考虑 AI 应用程序，组织可以积极应对与大语言模型驱动的应用相关的挑战和风险，同时发挥其变革潜力以推动创新和改善生活。

AI 应用卡片是维护 AI 应用上下文的文件，在应用程序治理方面发挥着至关重要的作用。AI 应用卡片封装了应用程序的所有相关数据，包括所使用的模型、所使用的数据集、应用程序和 AI 案例、应用归属人（见下一部分 RACI 模型中的不同所有者类型）和守护者。AI 应用卡片一种简单的方式来传达 AI 应用数据并实现数据共享，能够帮助 AI 治理组织、AI 委员会和监管机构了解这些应用及其使用的 AI 能力。而 AI 应用卡片也可以逆向关联到其背后的模型卡片和数据卡片。

2.LLM 服务的生命周期

本章节主要概述 LLM 服务的不同生命阶段，每个阶段都对服务的效率、可靠性以及整个生命周期至关重要。从构想和规划的筹备阶段到最终的归档和处置阶段，每个阶段都被纳入一个综合框架，该框架旨在改善服务交付以及保障需求和标准的一致性。组织可以通过这种结构化方法清晰有效地管理服务开发、评估、部署、交付和退出。

基于新兴的标准，例如关于 AI 系统生命周期的标准 ISO/IEC 5338，以及来自英国数据伦理与创新中心(CDEI)等机构的综述，这个生命周期涵盖了端到端的过程，从早期准备和设计到培训、评估、部署、交付，最终退出。

下面我们对 LLM 服务生命周期各个阶段进一步细分。

- 准备：
  - 数据收集
  - 数据管理
  - 数据存储
  - 资源供应
  - 团队和专业知识
- 开发：
  - 设计
  - 训练
  - 开发过程中的关键考量
  - 护栏
- 评估/确认：
  - 评估
  - 验证/红色团队

- 重新评估
- 评估/验证期间的主要考量

- 部署:
  - 编排
  - AI 服务供应链
  - AI 应用
- 交付:
  - 运营
  - 维护
  - 持续监察
  - 持续改进
- 服务退出:
  - 归档
  - 数据删除
  - 模型清除

2.1 准备

这个阶段为整个 LLM 开发过程奠定了基础，并且极大地影响了模型的质量和伦理行为。从数据开始，在本节中，我们定义以下术语:

- 数据收集
- 数据管理
- 数据存储
- 资源供应
- 团队和专业知识

在构建大语言模型（LLM）的过程中，**数据收集**应专注于识别那些多样化、规模庞大且质量上乘的数据源，包括文本和代码等。我们不仅要遵循道德采购的最佳实践，还要警惕数据中可能存在的偏见。确保我们所收集的数据不仅满足有效训练的需求，而且能够反映我们对长期管理制度的承诺，以避免产生带有偏见或歧视性的输出。

数据管理是一个提升数据质量的系统化过程，它包括数据的清洗（去除错误、不一致和不相关信息）、分类（根据逻辑主题或类别组织数据）、标注（为监督学习分配标签），以及数据的匿名化和转换（调整数据格式以确保兼容性）。

数据存储解决方案，如数据库或云存储服务，必须确保数据的可访问性，同时采取严格的安全措施来保护敏感信息并遵守隐私法规。

在准备阶段，**资源配置**需要精心策划，选择适合的计算和云资源。硬件选择应考虑处理器类型（如 CPU、GPU、TPU）以及为 LLM 优化的内存配置。软件选择则包括稳定的操作系统、丰富的库和编程环境。利用云基础设施可以显著提高系统的可扩展性、灵活性和成本效益

此外，**团队的专业能力和知识**同样至关重要。数据科学家负责收集、处理和分析数据；机器学习工程师设计并微调 LLM；软件开发人员构建必要的工具；语言学家提供深入的语言专业知识；伦理学家则评估模型的社会影响，并提出减轻潜在风险的策略。

在构建大语言模型（LLM）的准备过程中，我们首先需要明确定义模型的目标和用途。这将指导我们在数据选择和处理过程中做出负责任的决策。我们应主动识别并处理数据中的潜在偏差，确保模型的公正性和准确性。在整个数据生命周期中，实施强有力的隐私保护措施是至关重要的。这不仅包括数据的收集和存储，也涵盖了数据的管理和使用。数据保管链（data chain-of-custoday）应成为我们安全数据工作和模型开发的基石，在数据收集、管理和存储的各阶段，确保训练数据的完整性和未被篡改至关重要。

2.2 开发

在这一阶段，我们的目标是将精心准备的数据和强大的计算资源转化为一个高效、可靠的功能性大语言模型（LLM）。

主要活动包括:

- 设计
- 开发供应链
- 训练
- 开发过程中的关键考量

2.2.1 设计阶段

模型架构选择：我们首先需要根据模型的预期任务，精心挑选合适的 LLM 架构，例如基于 Transformer 的模型或循环神经网络。在这一过程中，我们将综合考虑性能需求、计算资源限制以及模型将要处理的数据类型。

超参数优化：接下来，我们将确定那些控制模型训练过程的关键超参数，包括学习率、批量大小和网络层数等。这些参数的选择将直接影响模型的训练效率、收敛速度以及最终的准确性。

评估指标设定：为了全面跟踪模型在训练期间的表现，我们将定义一系列评估指标，如准确率、困惑度和 BLEU 分数，这些指标将帮助我们识别模型性能的改进空间。

2.2.2 发展供应链

基础模型利用：我们考虑采用预先训练好的基础模型，例如 GPT-3 或 BERT，它们为我们提供了一个强大的起点。通过对这些模型进行微调，我们可以针对特定数据集获得定制化的结果。

组件评估：我们将评估不同任务的需求，如命名实体识别、情感分析或文本摘要，并决定是选择现成的开源或闭源组件，还是开发自定义组件来满足这些需求。

框架选择：为了简化模型的开发、训练和部署流程，我们将选择一个功能强大的机器学习框架，如 TensorFlow、PyTorch 或 Ray。

2.2.3 训练阶段

训练流程实施：我们将精心策划的数据输入到选定的模型架构中，并运用优化算法，如梯度下降法，迭代更新模型参数，以最小化训练数据中的误差。

训练监控：在训练过程中，我们将使用之前定义的评估指标密切监控模型的进展，及时发现过拟合或欠拟合的迹象，并相应调整训练策略或超参数。

实验迭代：通过迭代方法，我们将测试不同的模型结构、超参数和数据预处理技术，以探索最佳配置。

标记化处理：这一步骤涉及将输入文本分解为更小的单元，称为“标记”，这些可以是单词、子单词单元或单个字符。标记化的主要目的是将原始文本转换为数值格式，以便 LLM 的神经网络进行处理。通过将每个标记映射到一个唯一的整数值或嵌入向量，标记化不仅影响模型对输入文本的表示和处理方式，而且是 LLM 工作流程中的基础步骤。正确的标记化方法可以显著提升模型理解和生成自然语言的能力，同时确保计算效率。

2.2.4 开发过程中的关键考量

透明性：我们致力于记录设计决策、模型架构和训练流程，这不仅促进了项目的可重复性，也增强了结果的可靠性。透明度是构建信任和确保研究诚信的基石。

可解释性：我们优先采用能够阐明模型决策过程的技术，特别是在那些涉及高风险的应用场景中。通过增强模型的可解释性，我们能够更好地理解其输出，从而提高用户对 AI 系统的信任。

效率：我们在追求模型性能的同时，也注重计算资源的有效利用。我们探索各种优化技术，如模型量化和剪枝，旨在提升模型运行效率，同时确保其准确性不受损害。

版本控制：我们实施了一套强大的版本控制系统，用以追踪模型、标记化策略、训练数据集以及其他组件的每一次变更。这一做法不仅确保了研究的可重复性，也为必要时的版本回退提供了可能，同时促进了开发团队成员间的协作。

2.3 评估与确认

评估阶段是在部署大语言模型（LLM）之前，对其进行严格的性能、可靠性和适用性评估，以确保满足预期目标。

本节定义以下术语:

- 评估
- 验证/红队
- 重新评估
- 评估过程中的主要考量/验证

2.3.1 评估

度量：采用定量与定性指标相结合的方法，为 LLM 量身定制评估体系。定量指标涵盖准确度、精确度、召回率、F1 分数、针对语言生成任务的困惑度以及翻译任

务的 BLEU 分数。定性评估则可能包括人类评审员对输出的流畅性、连贯性和相关性的专业判断。

基准测试：对比 LLM 的表现与已建立的基准线及行业内其他先进模型，以识别其相对优势和潜在不足。

偏见和公平性检验：对模型输出进行检查，以识别可能存在于不同人群或敏感属性中的潜在偏见。通过使用公平性指标来量化这些差异。

2.3.2 验证/红队

真实世界测试：将 LLM 置于与其预期用例相似的真实环境中测试，以评估模型面对未知数据时的表现，从而衡量其泛化能力。

人在回路（Human-in-the-loop）：让专家参与到 LLM 输出的评估中，尤其在对准确性和细节要求极高的敏感领域，并收集反馈以指导未来的优化。

红色团队：组建一支专业的对抗团队，深入挖掘 LLM 的潜在漏洞、偏见和故障模式，有助于发现常规测试中可能遗漏的弱点。

2.3.3 重新评估

监控：在部署 LLM 后，对模型性能进行持续监控。建立监测机制，以识别数据和模型漂移以及性能随时间下降的问题。

数据漂移是指输入数据随着时间推移发生变化，这种变化可能导致模型性能下降。当模型真实输入数据与模型训练数据有偏离时，模型的预测准确性和可靠性就会降低。

模型漂移是指随着时间推移，输入特征和目标变量之间的统计关系发生变化，从而导致模型的预测能力下降。这种漂移可能由多种因素引起，包括但不限于：生成数据的基础过程变化、消费者行为变化、外部环境因素等。

数据漂移和模型漂移都会导致机器学习模型的性能下降，因此，对这些潜在问题进行持续监控并采取有效措施至关重要。持续监控、重新训练或用新数据更新模型等技术可以缓解这些问题。

触发再培训：设定明确标准，以判断何时需要对 LLM 进行全面或部分的再培训，以响应性能下降或数据分布的变化。

2.3.4 评估/验证过程中的主要考量

弹性：为确保 LLM 在面对不可预见的输入情况时仍能保持稳定和一致的性能，需要从对抗性输入鲁棒性、异常值、异常数据模式等方面评估 LLM。

不确定性：探索模型对其预测的信心水平，以便在实际应用中指导人类决策。

数据代表性：确保评估数据集与 LLM 在实际应用中处理的实时数据高度一致，以避免产生误导性结果。

2.4 部署

部署阶段将经过训练和验证过的 LLM 集成到提供服务的系统中。

本节定义以下术语：

- 编排
- AI 服务供应链
- 应用
- 部署过程关键因素
- Guardrails 护栏

2.4.1 编排

容器化：为了提高 LLM 的可移植性并简化其部署流程，采用容器化技术，将 LLM 及其必要的依赖项（包括库、数据等）封装进容器（如 Docker）中。

可扩展性：构建一个可根据需求进行灵活扩展的部署架构。同时使用负载均衡技术实现高效地分发传入请求。

版本控制：建立一个系统，用于跟踪模型版本、配置及性能指标的系统，有助于确保 LLM 更新是可回滚和可比较的。

IaaC：采用基础设施即代码（Infrastructure as Code）的方法，将基础设施的配置和管理过程自动化和代码化。这可以带来更改可追溯性、回滚可操作性等优点。

2.4.2 AI 服务供应链

代理：在构建大对话式人工智能系统时，确保大模型能够与自然语言理解（NLU）模块、对话管理器和知识库等关键组件进行交互。

插件：为提升大模型能力，可考虑将其与特定领域的插件或扩展进行集成，（例如，医疗保健或金融等特定领域的插件）。同时，必须考虑这些外部组件集成所带来的安全风险。

安全：在供应链中应优先考虑安全性，包括保护 API 端点、实施用户身份验证和授权机制、安全管理访问凭据，以及加密传输和存储中的数据。

2.4.3 应用

应用程序编程接口（APIs）：为便于外部系统和用户与由 LLM 驱动的应用程序进行交互，开发结构良好的 API，提供清晰的 API 文档，包括输入/输出格式、预期行为等。采用 REST（REpresentation State Transfer）等业界标准构建 API，同时实施版本控制。

检索增强生成（RAG）：考虑将检索组件集成到大模型中，以便模型能够访问并整合来自外部知识源的信息，从而提升响应的精确性。

输入提示：为了指导 LLM 更有效地执行特定任务并确保输出的安全性，探索使用提示注入技术。

不安全的输出处理：对大模型的输出进行严格审查，以预防可能引发安全漏洞的有害输出，如系统损害或数据泄露等风险。

2.4.4 部署过程关键因素

用户界面/用户体验（UI/UX）：设计用户友好的界面，确保用户能够与大模型应用程序顺畅交互。根据大模型的上下文环境，定制满足特定场景下的需求。

可观察性：为了跟踪 API 使用情况、LLM 性能和错误率，建立全面的日志记录和监控系统。记录数据有助于指导模型的调试和持续优化。

透明性：向用户清晰地阐述大模型的工作原理及其输出的潜在局限性，增强用户对模型输出的理解，建立信任感。

输入过滤：识别并防止可能对模型造成数据污染的恶意输入，以减少对模型输出的影响。

输出过滤：防止生成不恰当或有害的内容，如仇恨言论、暴力、露骨材料和其他被认为不可接受或有害内容。

隐私：为保障用户隐私安全，应实施控制措施以降低隐私风险，防止模型生成可能泄露个人或专有信息的内容。

滥用：限制 LLM 的使用，以防止其被用于生成欺诈性内容、钓鱼邮件等不当用途，或其他形式的操纵及不道德内容。

伦理准则和偏见缓解：为确保 LLM 的使用既符合伦理原则又遵循社会规范，应减少产生与种族、性别、性取向等相关的偏见和歧视性内容。

2.5 交付

交付阶段的核心在于对已部署的大模型进行持续管理，并通过不断的迭代改进以保持该模型的标准和性能。

通常认为，交付阶段涵盖了以下三个关键子阶段：

- 运营
- 维护
- 持续改进

2.5.1 运营

日志及监控：持续监控模型的关键性能指标，如准确性、延迟和资源利用率。一旦检测到安全问题或性能下降，立即通过告警系统通知相关人员。

事件响应：制定详细的应急响应计划和程序，及时处理和解决系统故障，并及时响应诸如网络攻击、漏洞或性能瓶颈等安全事件。

用户反馈：建立反馈机制，收集用户对模型输出的意见和建议。对收集到的用户反馈进行分析，以确定需要改进的领域或潜在问题。

2.5.2 维护

Bug 修复：识别并解决在模型训练、微调或部署过程中出现的代码错误或系统故障。发布补丁或更新，以确保系统的稳定性和数据的完整性。

安全更新：时刻经替新出现的安全威胁和漏洞。根据既定的漏洞管理服务水平协议（SLA），为大模型和相关系统提供安全补丁和更新。修补的过程应当全面覆盖正在被使用的第三方或公共大模型（LLM)版本。

重新训练模型：随着与大语言模型（LLM）交互数据特性的变化，可能需要更新训练数据或重新训练模型以维持最佳性能。

2.5.3 持续改进

重新训练：定期评估是否需要在新数据或更新的超参数上重新训练大模型。此举旨在解决概念漂移和性能下降问题，或以此为契机，扩展模型的能力。

持续反馈循环：将监控系统与用户反馈机制相结合，形成一个闭环反馈系统。通过监控数据和用户反馈指导模型重训练与持续优化。

实验：不断探索可能提高模型整体性能的新模型架构、算法或训练技术。

2.5.4 交付过程中的关键事项

在整个操作和维护过程中，持续监控模型，以识别模型在部署后可能的任何恶意行为或偏差。通过主动监控和及时干预，确保及时发现并解决问题，从而减轻对用户或系统的潜在负面影响。

变更管理在维护模型的稳定性和性能方面至关重要。一方面可建立全面的变更管理流程，记录所有的更新并跟踪其对性能的影响。另一方面可通过制定强有力的变更管理程序，有效管理模型的演进，减少系统中断和性能下降。此外，制定有效的回滚计划，以便在变更或修改出现问题时迅速采取应对措施。

为潜在的停机情况制定计划也是模型维护的关键内容。对可能导致服务中断的更新或维护要有一定的预见性并制定针对性工作计划。一方面要充分告知用户，确保他们了解停机窗口及伴随的服务中断。并确保利益相关方能够了解情况并为可能对其运营造成的影响做好准备。

通过主动解决停机问题，组织方在满足用户的期望和要求的同时，也可以保持大模型（LLM）的可靠性和可用性。

2.6 服务退出

此阶段的重点是当大模型被新模型取代或其继续运行会带来不可接受风险时，正确的停用大语言模型。

本节中定义如下术语：

- 归档
- 数据删除
- 模型处置

2.6.1 归档

模型保存：即对大模型及其所有相关组件，包括代码、配置文件和训练数据，进行存档。按照组织方（此处建议讨论统一）的数据保留策略，对存档内容进行安全、合规的存储。存档内容对历史分析、审计、模型迭代或复用等场景具有重要价值。

文档记录：保存大模型的所有文档，涵盖设计、开发过程、性能指标、使用限制等关键信息。同时记录使用过程中遇到的所有事件。

2.6.2 数据删除

法规：遵守数据治理法规（如 GDPR、CCPA），安全地删除在大模型运行期间所收集或训练的任何个人或敏感数据。

保留政策：制定明确的数据保留策略，规定数据的存储期限和条件，确立数据安全处置的流程和方法。

2.6.3 模型处置

再利用评估：对大模型及其组件进行全面评估，确定是否适合复用于其他应用或研究项目。从而降低开发成本和环境影响。

知识产权：针对退役的大模型，审慎处理所有相关的知识产权问题，特别是使用外部资源或许可技术开发的大模型。

安全处置：如果确定大模型无法再利用，应采取安全措施进行处置，防止未经授权的访问或潜在滥用。对于存储在物理介质（如硬盘、SSD 或可移动存储）上的模型，考虑采用物理销毁方法，确保数据无法被恢复。即可以通过消磁、粉碎或物理销毁等方式实现。

2.6.4 服务退出期间的关键考虑因素

告知：在大模型服务退役前，通知所有用户和利益相关方。如有必要，提供清晰的迁移指南，以将服务迁移到替代服务或解决方案。

影响评估：仔细评估大模型服务退役的潜在影响，特别是对敏感领域或高度依赖该服务的用户的影响。

知识转移：从退役模型的开发和运营过程中总结经验教训，并能够有效地应用到组织未来的人工智能项目中。

3.大语言模型服务影响分类

我们可以将影响类别直接对应到已经确立的 CIA 安全三要素（机密性、完整性和可用性）上。此外，根据 NIST 文件 AI 100-2 E2023，还可以增加“滥用/误用”和“隐私丧失”这两个新的影响类别。

以下是对 LLM 相关风险的概括分类：

- 机密性：存在这样一种风险，即 LLM 的数据、模型本身或其生成的输出可能会被泄露给未授权的个人，这涉及敏感信息，可能包括个人数据、商业秘密或其他机密材料。

- 完整性：存在 LLM 的数据或其生成的输出被恶意或意外地修改或损坏的风险，这可能导致结果不正确或具有误导性。
- 可用性：对 LLM 操作可能存在遭受干扰的风险，导致用户在关键时刻无法访问。这些干扰可能包括服务拒绝攻击、系统故障、意外停机、过高的计费限额或计算资源不足等情况。

4.大语言模型服务威胁分类

大语言模型服务威胁分类的初始列表涵盖一系列需要重点考虑并缓解的潜在风险和漏洞。每个类别都代表一个独特的挑战，可能会损害大语言模型服务的完整性、安全性和有效性。具体分类如下：

1. 模型操纵
2. 数据投毒
3. 敏感数据泄露
4. 模型窃取
5. 模型故障/失灵
6. 不安全的供应链
7. 不安全的应用程序/插件
8. 拒绝服务
9. 缺少治理/合规性

4.1 模型操纵

模型操纵涉及试图逃避检测或操纵大语言模型产生不正确或误导性的结果。包括直接或间接指令注入（对抗性输入）等技术，旨在利用模型学习训练和决策过程中的漏洞。

4.2 数据投毒

数据投毒是指操纵大语言模型训练数据的一种行为，攻击者可能故意向训练数据中注入虚假、误导性或无用信息，或利用数据集中已有的错误和偏差。无论哪种情况，数据投毒都可能使模型受到污染，导致模型学习到错误的模式，产生带有偏见的预测结果，并降低其可信度。

4.3 敏感数据泄露

敏感数据泄露指的是对大语言模型服务在处理或存储过程中的敏感信息进行未授权访问、披露或泄露的威胁。这类敏感信息可能涵盖个人隐私数据、商业专有数据或机密文件。一旦这些数据遭到泄露，可能会引发隐私侵犯和安全漏洞的问题。

4.4 模型窃取

模型窃取（也称为模型蒸馏）指的是恶意行为者未经授权地访问或复制大语言模型。攻击者可能会尝试对模型的架构进行逆向工程，或者提取出专有的算法和参数。这种行为可能会导致知识产权被盗用，或者模型被未经授权地复制和使用。

4.5 模型故障/失灵

模型故障/失灵指大语言模型服务中可能出现的软件错误、硬件故障、操作错误等问题。此类事件可能会破坏服务可用性、降低性能、破坏模型输出准确性和可靠性。

4.6 不安全的供应链

不安全的供应链指的是在大语言模型的生态系统中，由于第三方组件、依赖项或服务的引入而产生的安全漏洞。这些漏洞可能被恶意利用，从而损害大语言模型服务的安全性和可靠性，例如通过使用被篡改的软件库或存在缺陷的硬件组件。

4.7 不安全的应用程序/插件

不安全的应用程序/插件包括与大语言模型服务交互的插件、函数调用或扩展中引入的漏洞。不安全或恶意设计的应用程序/插件可能会引入安全漏洞、特权提升或对敏感资源进行未经授权的访问，这些都会对集成系统的输入和输出构成风险。

4.8 拒绝服务

拒绝服务攻击旨在通过大量请求或恶意流量压垮 LLM 服务，从而破坏其可用性或功能。DoS 攻击可以使服务对合法用户不可访问，导致停机、服务质量下降或信任丧失。

4.9 缺少治理/合规性

这一类别涉及不遵守监管要求、行业标准或管理 LLM 服务运营和使用的内部治理与合规政策的风险。未能遵循治理与合规标准可能导致法律责任、财务处罚或声誉损失。

采取全面的方法应对大模型服务的威胁风险，包括但不限于实施强有力的安全措施、进行持续的风险评估、集成威胁情报，以及制定针对模型独特特性的主动缓解策略。

从安全控制和风险管理的角度出发，我们需要识别与大模型系统相关的弱点和漏洞，以便采取相应的预防和修复措施。

大模型的弱点可能表现在多个方面，如训练数据的局限性、算法的偏差或模型架构的缺陷。例如，模型对训练数据中统计模式的依赖可能导致在处理语言的细微差别或识别潜在的恶意输入时存在不足。

大语言模型的漏洞是指在特定情况下，攻击者可以利用这些漏洞破坏模型的完整性、机密性、可用性或模型的输出。这些漏洞可能源自模型实践中的缺陷，例如编码错误、配置不当、训练数据被操纵以引入偏见或通过反向示例进行攻击。

从风险管理角度看，识别并减轻模型中的弱点和漏洞风险对于防范潜在威胁并降低威胁影响十分重要。这涉及评估对模型攻击的可能性及潜在影响、根据风险的严重程度确定风险的优先级，并实施适当的安全控制措施，以减轻或将这些风险转移到可接受的水平。作为安全策略的一部分，应该通过红/蓝对抗提高系统的安全性。

通过区分弱点、漏洞和攻击，人工智能控制框架可以提供一种结构化的方法来识别、评估和减轻与部署人工智能系统相关的风险。这使得组织能够制定有效的策略来防范潜在威胁，增强其人工智能系统的抗打击能力，并保证其运营的可信度。

5.参考文献

1. BARRETT, A.M., NEWMAN, J., NONNECKE, B., HENDRYCKS, D., MURPHY, E.R. and JACKSON, K. (2023). CLTC Center for Long-Term Cybersecurity UC Berkeley AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models.[online] UC Berkeley Center for Long-Term Cybersecurity, pp.1–94. Available at: https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf

2. Huang, K., Wang, Y., Goertzel, B., Li, Y., Wright, S., Ponnapalli, J. (ed.). (2024). Generative AI Security Theories and Practices. Springer. https://link.springer.com/book/9783031542510

3. CSA. (2024). AI Organizational Responsibilities Working Group. AI Organizational Responsibilities - Core Security Responsibilities. [online] Available at: https://cloudsecurityalliance.org/artifacts/ai-organizational-responsibilities-core-security-responsibilities

4. CSA. (2024). AI Technology and Risk Working Group. The AI Model Risk Management Framework. Available at: https://cloudsecurityalliance.org/research/artifacts?term=artificial-intelligence

5. IBM. IBM Watsonx (2024). AI risk atlas. AI risk atlas. [online] Available at: https://dataplatform.cloud.ibm.com/docs/content/wsj/ai-risk-atlas/ai-risk-atlas.html?context=wx&audience=wdp

6. GOV.UK. (n.d.). AI Foundation Models: initial review. [online] Available at: https://www.gov.uk/cma-cases/ai-foundation-models-initial-review.

7. Andreessen Horowitz. Radovanovic, M.B., Rajko (2023). Emerging Architectures for LLM Applications. [online]. Available at: https://a16z.com/2023/06/20/emerging-architectures-for-llm-applications

8. ENISA. (2020). Artificial Intelligence Cybersecurity Challenges. [online] Available at: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.

9. Microsoft. MicrosoftLearn. (2022).Failure Modes in Machine Learning - Security documentation. [online] Available at: https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

10. ISO/IEC 22989:2022. Information Technology – Artificial Intelligence – Concepts & Terminology. [online] ISO. Available at: https://www.iso.org/standard/74296.html.

11. ISO/IEC TR 24028:2020. (2022). Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence. Available at: https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:24028:ed-1:v1:en

12. 2912. ISO/IEC 27090. Cybersecurity – Artificial Intelligence – Guidance for addressing security threats to artificial intelligence systems. Available at: https://www.iso.org/standard/56581.html

13. ISO/IEC 27091. Cybersecurity and Privacy – Artificial Intelligence – Privacy protection – Guidance for organizations to address privacy risks in artificial intelligence (AI) systems and machine learning (ML) models. Available at: https://www.iso.org/standard/56582.html

14. ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system. Available at: https://www.iso.org/standard/81230.html

15. ISO/IEC DIS 5338. Information technology — Artificial intelligence — AI system life cycle processes. [online] ISO. Available at: https://www.iso.org/standard/81118.html.

16. ISO. Online Browsing Platform (OBP). Terms and Definitions. Available at: https://www.iso.org/obp/ui

17. The MITRE Corporation. MITRE Atlas. Atlas Matrix. Available at: https://atlas.mitre.org/matrices/ATLAS

18. NIST. NIST AI 100-2 E2023. (2024). Vassilev, A., Oprea, A., Fordyce, A. and Anderson, H. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. [online] csrc.nist.gov. Available at: https://csrc.nist.gov/pubs/ai/100/2/e2023/final

19. NIST AI RMF 1.0. AI Risk Management Framework. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). [online] doi: https://doi.org/10.6028/nist.ai.100-1.

20. NIST. U.S. Artificial Intelligence Safety Institute (USAISI). Available at: https://www.nist.gov/aisi

21. OWASP. Top 10 for LLM Applications and Generative AI. Available at: https://genai.owasp.org

22. Cornell University. (2023). ARXIV. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J. and Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. [online] https://doi.org/10.48550/arXiv.2312.10997.
23. Manral, V. LinkedIn. (2023). Shared Responsibilitu Model. Available at: https://www.linkedin.com/posts/vishwasmanral_generativeai-chatgpt-sharedresponsibility-activity-7084313628614537216-DvLj/

24. European Union. (2024). EU Artificial Intelligence Act. Available at: https://artificialintelligenceact.eu

CSA GCR cloud security alliance GREATER CHINA REGION ®

Cloud Security Alliance Greater China Region

扫码获取更多报告

云安全联盟大中华区
官网：https://c-csa.cn 电话：0755-86548359
地址：上海市浦东新区东育路255弄5号3FB30（前滩国际经济组织集聚区）/
广东省深圳市南山区粤海街道北京航空航天大厦1号楼905