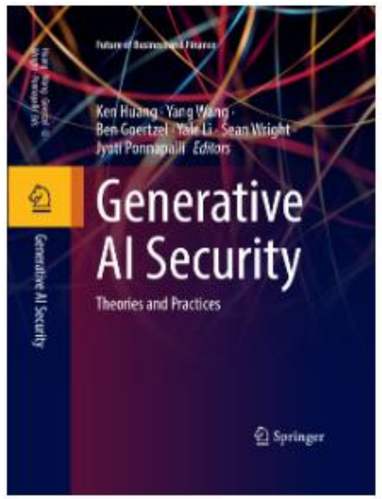
生成式 AI应用的安全挑战

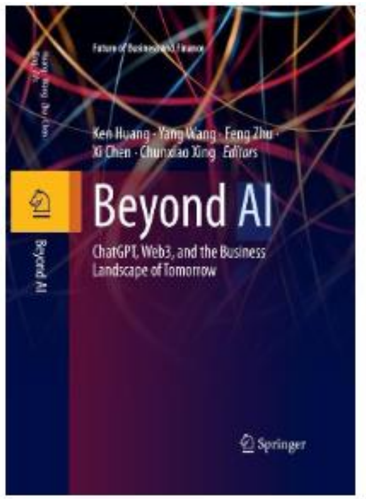
黄连金

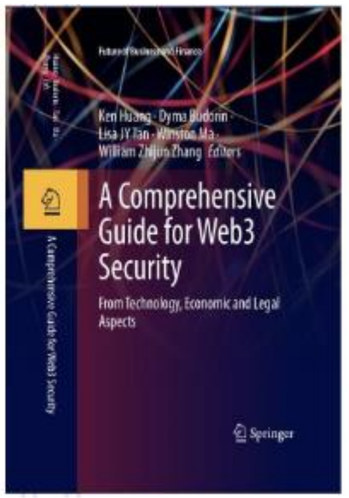
云安全联盟大中华区研究院

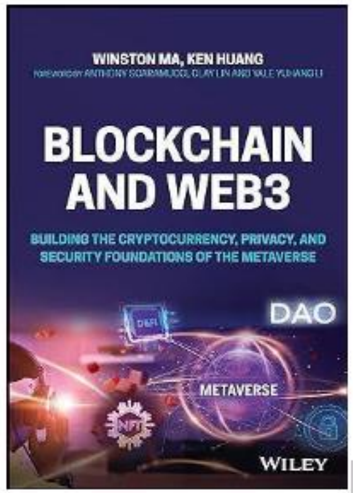
About Me: Ken Huang (黄连金)

- 过去 6 年以来撰写了 8 本有关 AI 和 Web3 的书籍
- CSA GCR 研究院副院长
- CSA全球 AI Safety 安全工作组联合主席
- WDTA AT STR 联合主席
- OWASP Top 10 for LLM 核心作者









Springer

CAMBRIDGE UNIVERSITY PRESS

WILEY

机械工业出版社
CHINA MACHINR PRESS

云安全联盟 CSA

Cloud Security Alliance

云安全联盟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。

大中华区 20+

北京、杭州、上海、深圳、苏州
西 安、大连、青岛、香港、澳门、台湾

美洲区 60+

加拿大、智利、阿根廷、墨西哥、巴拿马、秘鲁、玻利维亚、温哥华、俄亥俄州东北部、匹兹堡、旧金山、硅谷、波士顿、巴西、哥伦比亚、芝加哥.....

欧非区 40+

英国、比利时、卢森堡、丹麦、埃及、法国、德国、爱尔兰、以色列、意大利、荷兰、挪威、波兰、葡萄牙、俄罗斯、西班牙、瑞典、瑞士、土耳其、阿拉伯联合酋长国.....

亚太区 30+

澳大利亚、日本、韩国、马来西亚、孟买、尼泊尔、泰国、菲力宾、沙特、越南、新加坡、孟加拉国.....

4 大区

大中华区、美洲区、欧非区、亚太区

180+ 分会

香港、澳门、新加坡、西雅图、芝加哥、洛杉矶覆盖50多个国家和地区

2500+ 成员单位

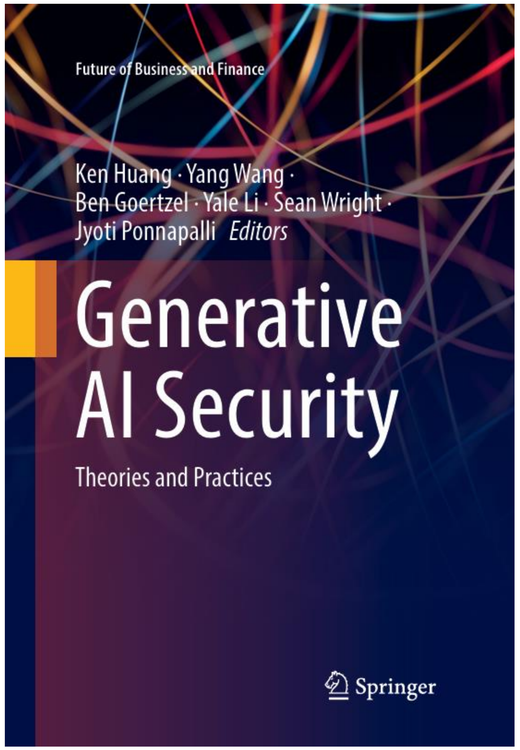
世界500强科技公司、安全厂商、中小型企业 、研究机构

180000+ 社区专业人员

研究工作组专家、社区志愿者、从业人员、CSA认证学员

生成式AI的安全挑战

- 影子AI (Shadow AI)
- 自动化和可扩展的威胁和零日漏洞 (Automated and Scalable Threats and Zero Day Vulnerabilities)
- 权限策略问题 (Entitlement Policy Issues)
- 安全工具集成问题 (Security Tools Integration Issues)
- 恶意GenAI工具的出现 (Emergence of Malicious GenAI Tools)
- 聚合导致的数据泄露 (Data Leak due to Aggregation)
- 新兴网络安全威胁 (Emerging Network Security Threats)



CSA Proposed

- 模型操纵 (Model Manipulation)
- 数据污染 (Data Poisoning)
- 敏感数据泄露 (Sensitive Data Disclosure)
- 模型盗窃 (Model Theft)
- 故障/失常 (Failure / Malfunctioning)
- 不安全的供应链 (Insecure Supply Chain)
- 不安全的应用程序/插件 (Insecure Apps/Plugins)
- 拒绝服务攻击 (Denial of Service (DoS))
- 治理/合规性丧失 (Loss of Governance / Compliance)

OWASP 大语言模型人工智能应用Top 10 安全威胁

LLM01：提示词(Prompt) 注入(Injection) 黑客通过设计过的输入（提示词）操纵大型语言模型(LLM)，从而导致 LLM 执行意外操作。提示词注入会覆盖系统提示词，而间接注入操纵外部数据源进行注入攻击。

LLM02 不安全的输出处理 当 LLM 输出未经审查而被接受时，就会出现此漏洞，从而暴露后端系统。滥用可能会导致 XSS、CSRF、SSRF、权限升级或远程代码执行等严重后果。

LLM03 训练数据中毒 当 LLM 培训数据被篡改，引入损害安全性、有效性或道德行为的漏洞或偏见时，就会发生这种情况。来源包括 Common Crawl、 WebText 、 OpenWebText和书籍。

LLM04 拒绝服务模型 攻击者对大型语言模型进行资源密集型操作，导致服务降级或高成本。由于LLM的资源密集型性质和用户输入的不可预测性，该漏洞被放大。

LLM05 供应链漏洞 LLM 应用程序生命周期可能会受到易受攻击的组件或服务的影响，从而导致安全攻击。使用第三方数据集、预先训练的模型和插件可能会增加漏洞。

LLM06 敏感信息披露 LLM可能会在其回复中泄露机密数据，从而导致未经授权的数据访问、隐私侵犯和安全漏洞。实施数据清理和严格的用户策略来缓解这种情况至关重要。

LLM07 不安全的插件设计 LLM 插件可能具有不安全的输入和不足的访问控制。缺乏应用程序控制使它们更容易被利用，并可能导致远程代码执行等后果。

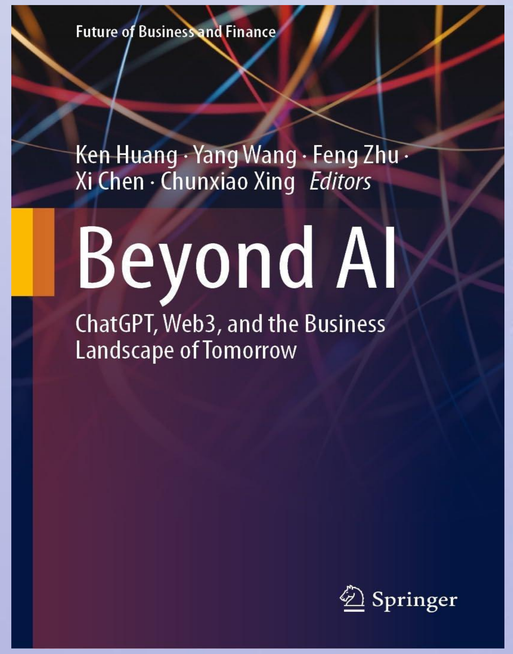
LLM08 过度代理 基于LLM的系统可能会采取导致意想不到的后果的行动。该问题源于授予基于 LLM 的系统过多的功能、权限或自主权。

LLM09 过度依赖 过度依赖LLM而不受监督的系统或人员可能会因LLM生成的不正确或不适当的内容而面临错误信息、沟通不畅、法律问题和安全漏洞。

LLM10 模型盗窃 这涉及对专有LLM模型的未经授权的访问、复制或泄露。影响包括经济损失、竞争优势受损以及敏感信息的潜在访问。

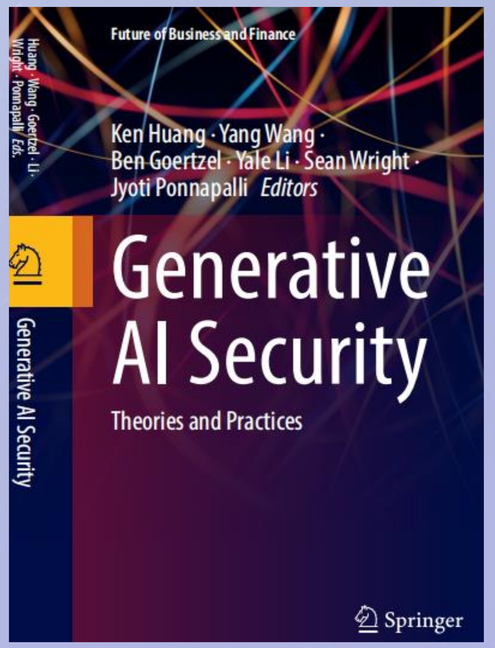
CSA/WDTA TWO Certifications and Official Study Guide of DigiBridge

Certified Chief AI Officer



Beyond AI

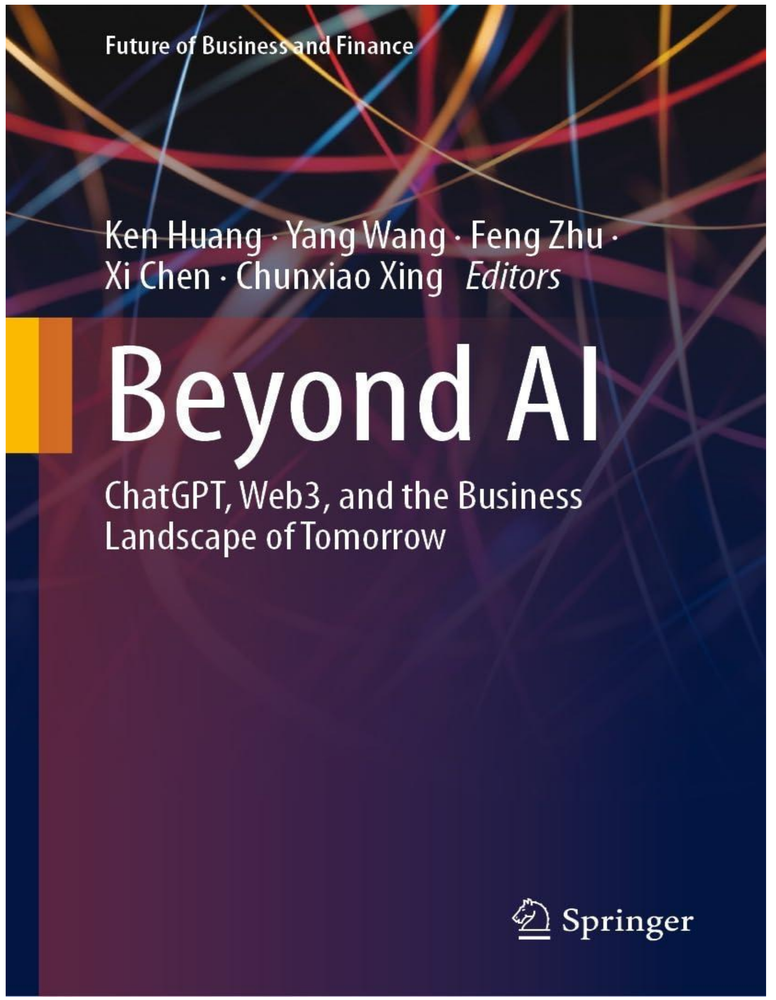
Certified AI Security Professional



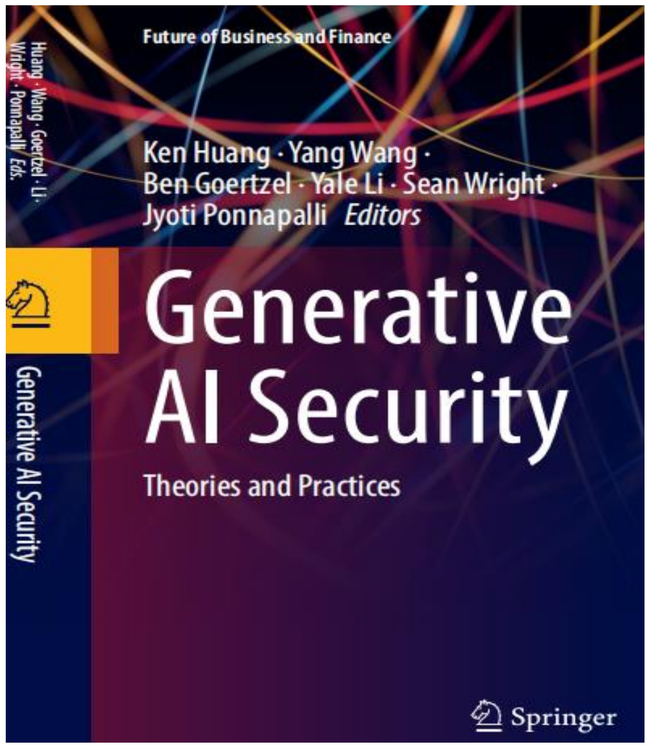
Generative AI Security

Table Of Contents

- 第一部分:第1-2章 - 概述ChatGPT、Web3及其对企业自动化的影响。
- 第二部分:第3-9章 - 探讨ChatGPT在产品管理、外包经济、营养科学、金融银行业、政府, 房地产和游戏行业的应用。
- 第三部分:10-12章 - 分析ChatGPT在安全隐私、法律和道德方面的考虑。



GenAI Security Book (39 AI and Security Experts)



- 第一部分:第一章探讨GenAI基础原理,第二章审视GenAI安全趋势。
- 第二部分:第三至七章涵盖全球AI治理、企业GenAI安全、数据，模型，应用，安全。
- 第三部分:第八章LLMOps与DevSecOps,第九章提示工程技术,第十章GenAI安全工具。

6 Editors and 19 Chapters contributors from:

OpenAI、Google、NVIDIA、Meta、JPMorgan 、Chase、Cohere、Cloud Security Alliance、OWASP、DistributedApps.ai、HKUST、Singularity Net、WDTA、UMG、Turist、Private AI、PIMCO、ISACA、Silicon Valley AI++、AI 2030 and more (Total 19 chapter co-authors)

Review, Foreword and Recommendation (by 14 KOLs)

Xuedong Huang, CTO at Zoom;Jim Reavis, CEO and Founder of Cloud Security Alliance;Seyi Feyisetan, PhD, Principal Scientist, Amazon;Anthony Scaramucci, Founder and CEO of Skybridge.;Jerry Archer, former CSO of Sallie Mae;Sunil Jain, Vice President and Chief Security Architect of SAP. Caleb Sima, Chair for AI Safety Initiative at Cloud Security Alliance;Dr. Cari Miller, Founder and Principal, AI Governance & Research at The Center for Inclusive Change;Diana Kelley, CISO, Protect AI. Tal Shapira P.hD., Co-Founder & CTO at Reco AI and Cybersecurity Group Leader at the Israeli Prime Minister's Office, and Professor Leon Derczynski, IT University of Copenhagen. Founder @ garak.ai.

CSA/WDTA AI-STR Standards

World Digital Technology Academy (WDTA)

World Digital Technology Academy Standard

WDTA AI-STR-01

Edition: 2024-04

Generative AI Application Security Testing and Validation Standard

The World Digital Technology Standard WDTA AI-STR-01 is designated as a WDTA norm.

World Digital Technology Academy (WDTA)

World Digital Technology Academy Standard

WDTA AI-STR-02

Edition: 2024-04

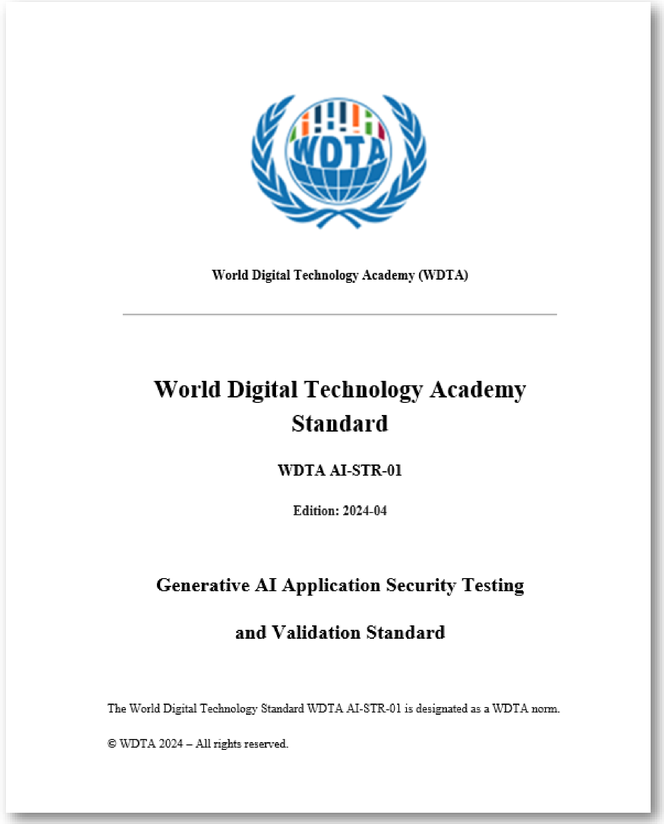
Large language model security testing method

The World Digital Technology Standard WDTA AI-STR-01 is designated as a WDTA norm.

扫码下载

CSA/WDTA AI-STR-01 Standard Overview

- **范围:**关注使用大型语言模型测试和验证下游AI应用程序的安全性。
- **重点领域:**包括基础模型选择、嵌入和向量数据库、提示执行/推理、主动行为、微调、响应处理和AI应用程序运行时安全性。
- **安全与验证**:确保AI应用程序在整个生命周期内安全和按设计运行。
- **标准框架:**为AI应用程序堆栈的每个层次提供安全和合规性指导方针。



World Digital Technology Academy (WDTA)

World Digital Technology Academy Standard

WDTA AI-STR-01

Edition: 2024-04

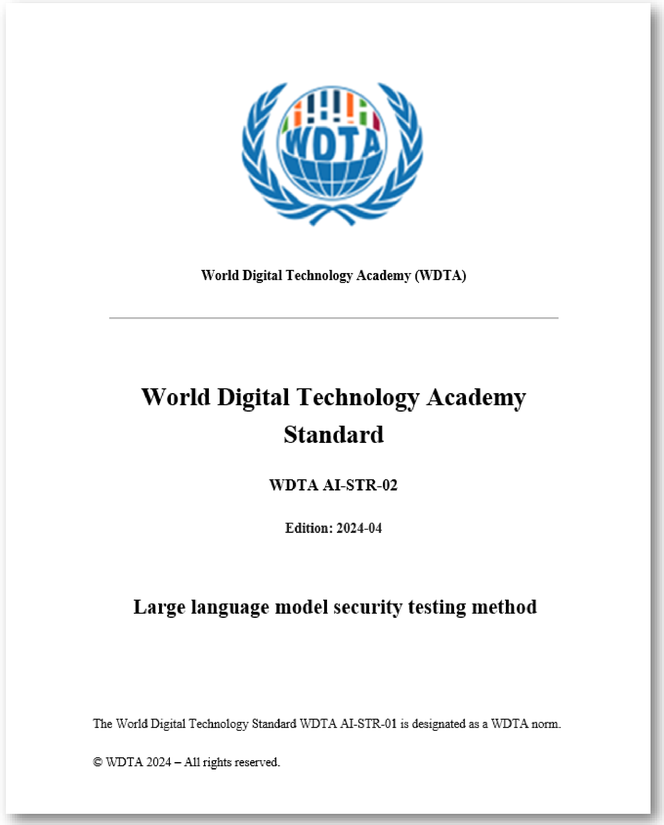
Generative AI Application Security Testing and Validation Standard

The World Digital Technology Standard WDTA AI-STR-01 is designated as a WDTA norm.

Generative AI Application Security Testing and Validation Standard

CSA/WDTA AI-STR-02 Standard Overview

- **范围**:适用于评估大型语言模型对抗攻击的能力。
- **攻击分类**:包括L1随机、L2盲盒、L3黑盒、L4白盒。
- **关键指标**:攻击成功率(R)和下降率(D)。
- **攻击方法**:详细介绍了诸如指令劫持和提示掩码等多种方法。
- **风险分析**:附录涵盖了与大型语言模型相关的伦理、安全和隐私风险。
- **测试流程**:概述了测试大型语言模型抵御不同类型攻击的方法。



World Digital Technology Academy (WDTA)

World Digital Technology Academy Standard

WDTA AI-STR-02

Edition: 2024-04

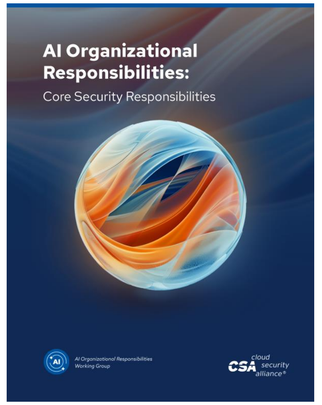
Large language model security testing method

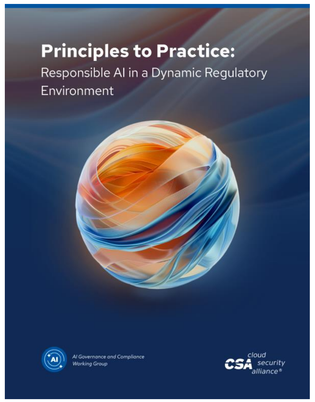
The World Digital Technology Standard WDTA AI-STR-01 is designated as a WDTA norm.

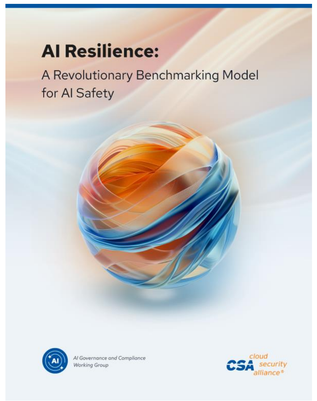
Large Language Model Security Testing Method

CSA AI安全领域其他成果

CSA大中华区成立AI安全工作组，中国电信、蚂蚁集团、安恒、电子科技大学、华为、百度安全等首批成员加入。







发布多项研究成果

2024年3月
在日内瓦万国宫举办**第27届联合国科技大会AI边会**

2024年5月
在RSA大会举办**CSA AI Summit @RSAC 2024**

CSA研讨会

2024 AI安全产业图谱

《AI安全产业图谱（2024）》由CSA大中华区发起，旨在系统收集和分析在AI安全领域活跃的各企业和组织的信息，提供一个全面的AI安全行业概览，为后续研究工作、产业发展、政策制定以及企业用户建设提供参考和支持。

调研对象：

面向AI领域的技术提供商、服务供应商、科研院所、高校及其他相关企业单位。

申报流程：

参与单位需自主填写问卷进行申报，图谱将于2024年第三季度发布。

申报通道：

扫描左图二维码添加CSA微信，获取AI安全产业图谱（2024）调研问卷

谢 谢