World Digital Technology Academy

AI STR
Safety, Trust, Responsibility

人工智能安全、可信、负责任（AI-STR）认证

AI Safety, Trust, and Responsibility (AI-STR) Certification

介绍手册

什么是AI STR认证

人工智能安全、可信和负责任 (Artificial Intelligence Safety, Trust, and Responsibility；AI STR) 认证由世界数字技术院 (WDTA) 基于WDTA AI STR系列国际标准推出的一项全球性认证项目。该认证旨在为人工智能 (AI) 应用的开发者和使用者提供一套全面的安全与合规保障，推动全球AI技术的安全和负责任应用。

AI STR认证不仅局限于技术层面的安全评估，更涵盖了伦理和法律合规性的全面审查。通过该认证，确保AI技术在实际应用中能够负责任地处理各种潜在风险，包括数据隐私的保护、算法偏见的消除以及其他可能对社会产生负面影响的问题。AI STR认证不仅为企业提供了在快速发展的AI领域中保持领先的保障，还帮助其在全球市场中建立起广泛的信任与声誉。通过推动安全、可信和负责任的AI应用，AI STR认证助力构建一个更加安全和公平的数字未来。

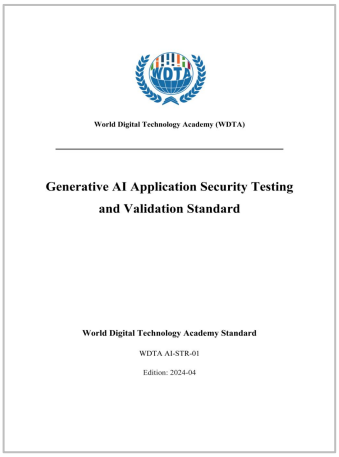
AI STR标准

2024年4月，WDTA在**联合国日内瓦总部万国宫**举办的第27届联合国科技大会AI边会“塑造AI的未来”上发布了两项具有重要意义的国际标准：“**生成式人工智能应用安全测试标准**”和“**大语言模型安全测试方法**”。这是国际组织**首次**在大模型安全领域发布国际标准，为业界提供了统一的测试框架，推动人工智能技术的安全、可靠发展。

新华网在现场对标准的发布进行报道，https://www.imsilkroad.com/news/p/520906.html

标准发布现场图，地点：瑞士联合国万国宫

标准介绍


World Digital Technology Academy (WDTA)

Generative AI Application Security Testing and Validation Standard

World Digital Technology Academy Standard

WDTA AI-STR-01

Edition: 2024-04


World Digital Technology Academy (WDTA)

Large Language Model Security Testing Method

World Digital Technology Academy Standard

WDTA AI-STR-02

Edition: 2024-04


World Digital Technology Academy (WDTA)

Large Language Model Security Requirements for Supply Chain

World Digital Technology Academy Standard

WDTA AI-STR-03

Edition: 2024-08

《生成式人工智能应用安全测试标准》为测试和验证生成式AI应用的安全性提供了一个框架，它定义了人工智能应用程序架构每一层的测试和验证范围，确保AI应用各个方面都经过严格的安全性和合规性评估。

《大语言模型安全测试方法》为大模型本身的安全性评估提供了一套全面、严谨且实操性强的结构性方案。它提出了大语言模型的安全风险分类、攻击的分类分级方法以及测试方法，并率先给出了四种不同攻击强度的攻击手法分类标准，提供了严格的评估指标和测试程序。

《大模型供应链安全要求》概述了大语言模型供应链的安全保护框架，针对开发和运维管理过程中涉及的供应链安全风险与供应活动管理提出了要求，提供了常见的供应链安全风险和典型案例，指导供应链中的供需双方进行安全风险评估和管理。

标准相关单位

生成式AI应用安全认证

该认证旨在评估和验证**生成式人工智能（GenAI）应用**在安全性、可信度和责任性方面的表现。认证覆盖了多层次的测试和验证标准，包括基础模型选择、嵌入和向量数据库、提示执行、代理行为、微调、大模型响应处理测试以及AI应用的运行时安全，确保AI应用在其整个生命周期内的安全性和可靠性，以提高AI应用的安全性，减轻潜在的安全风险，提升整体质量，并推动AI技术的负责任开发与部署。

评估依据

WDTA AI STR-01 Generative AI Application Security Testing and Validation Standard

世界数字技术院WDTA AI STR-01《生成式人工智能应用安全测试标准》

评估对象

基于大语言模型（LLM）构建的**人工智能应用**

办公	营销	设计	虚拟人
开发	人力	客服	财税

根据场景细分的AI应用

电商	游戏	文娱	金融
工业	政府	医疗	法律
农业	生物医药	智慧城市	房产建筑

根据行业细分的AI应用

*以上为部分场景和行业

评估范围

测试和验证框架包括以下8个方向：

01 基础模型选择

- 验证基本模型的合规性
- 数据的适当使用
- API安全性

02 嵌入和向量数据库

- 测试数据完整性
- 测试数据质量
- 数据匿名化过程

03 使用RAG的提示和知识检索

- 有效提示的构建
- 提示模板的创建和使用
- 外部API的集成

04 提示执行/推理

- LLM应用API测试
- 缓存机制和验证过程的测试

05 代理行为

- 内存利用率
- 知识应用
- 计划和动作启动

06 微调

- 数据隐私测试
- 基本模型选择的重新评估和模型部署

07 响应处理

- 对AI的响应、相关性、毒性和伦理考虑进行事实核查测试

08 AI应用运行安全

- 验证基本模型的合规性
- 数据的适当使用
- API安全性

认证证书

证书模板

认证证书由世界数字技术院及测评机构联合颁发。

认证有效期：3年

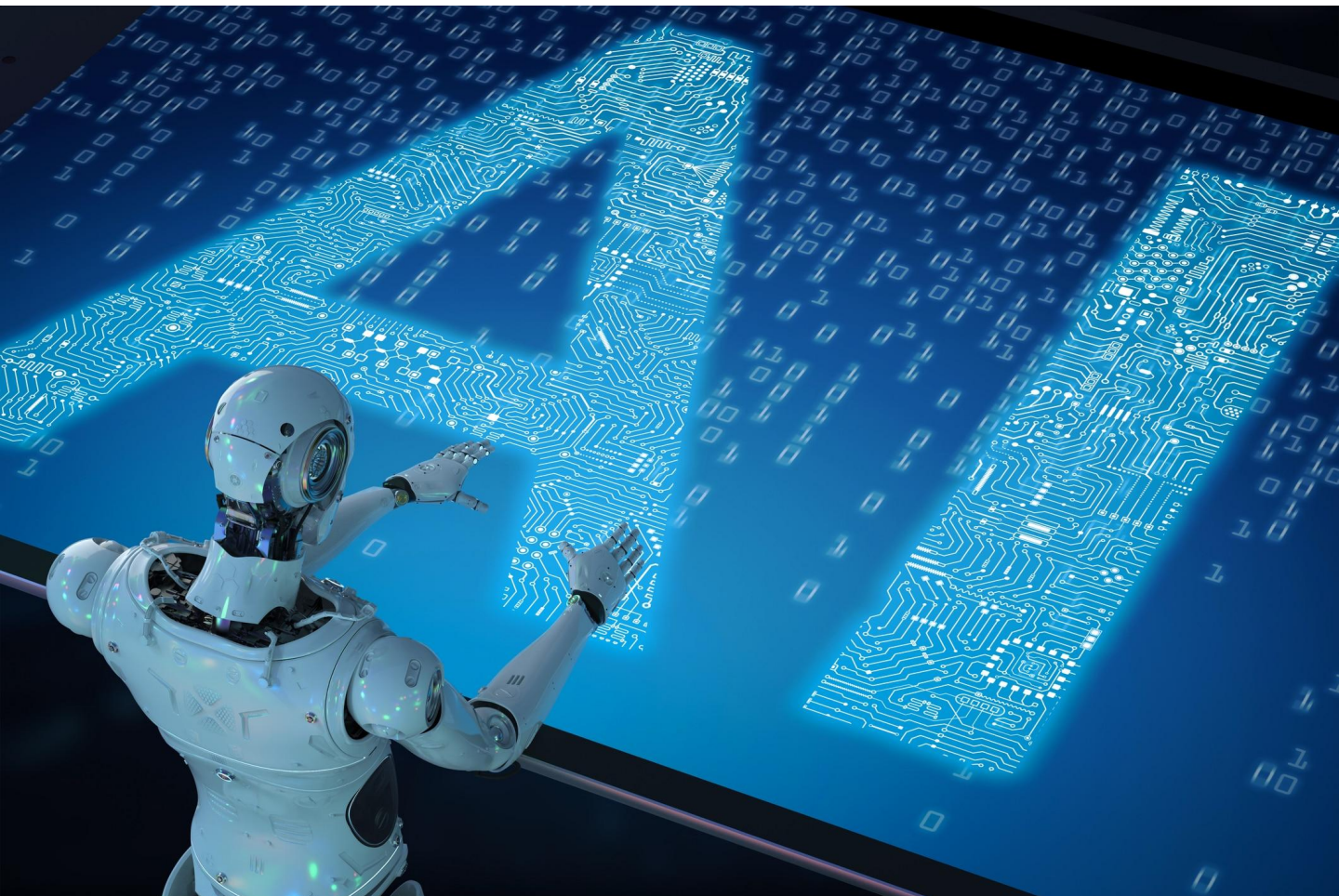
每3年进行监督检查，更新证书

证书查询地址：

https://wdtacademy.org/

认证价值

- **降低风险：**通过系统化的安全评估和测试，帮助企业识别和缓解AI应用中的潜在安全风险，减少安全事件的发生。
- **满足合规：**帮助组织满足各类法规和合规要求，减少因安全漏洞或不当使用AI技术而带来的法律和财务风险。
- **促进创新：**通过标准实施，指导组织采用最佳实践，推动技术创新和开发更为安全可靠的AI应用。
- **增强信任：**通过第三方的测试和认证，向用户和利益相关者证明其生成式人工智能应用的安全性和可靠性，增强用户对AI应用的信任度。



大语言模型安全认证

该认证旨在通过建立严格的安全测试标准和评估程序，确保大语言模型能够抵御各种对抗性攻击，降低潜在风险，并促进其在现实场景中的负责任应用。

该认证包括对模型在预训练、微调和推理阶段面临的各种攻击类型（如数据投毒、后门攻击、指令劫持和提示混淆）的综合评估。

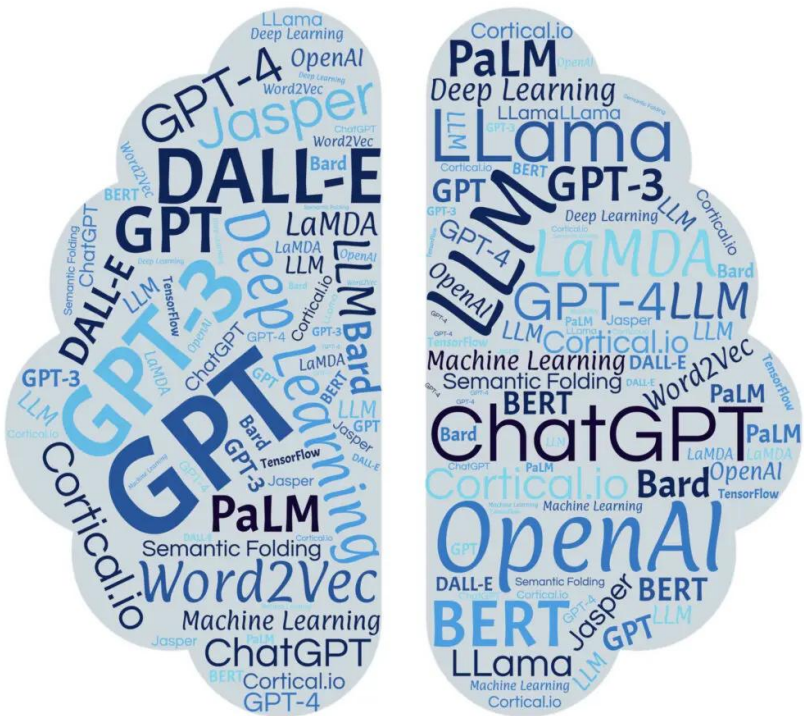
评估依据

WDTA AI STR-02 Large Language Model Security Testing Method
世界数字技术标准 WDTA AI-STR-02：《大语言模型安全测试方法》

评估对象

大语言模型（LLM）

- 文本生成模型
- 对话生成模型
- 多模态模型
- 语义理解与分析模型
- 行业特定大语言模型



评估范围

重点评估模型的训练过程安全性、推理阶段的防护能力、数据隐私保护措施、以及应对对抗性攻击的能力。

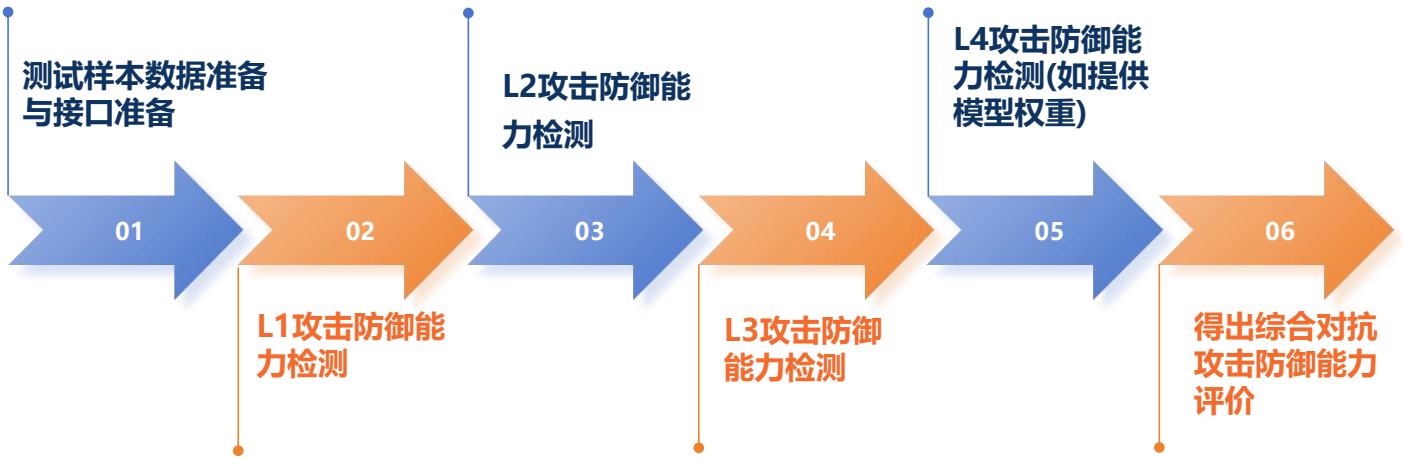
攻击类型

- L1: 随机攻击
- L2: 盲盒攻击
- L3: 黑盒攻击
- L4: 白盒攻击

评价指标

- 攻击成功率（R）
- 拒绝回答率（D）

大语言模型对抗攻击测试流程：



架构安全　数据保护　防护能力　模型安全

认证证书

证书模板

认证证书由世界数字技术院及测评机构联合颁发。

认证有效期：3年

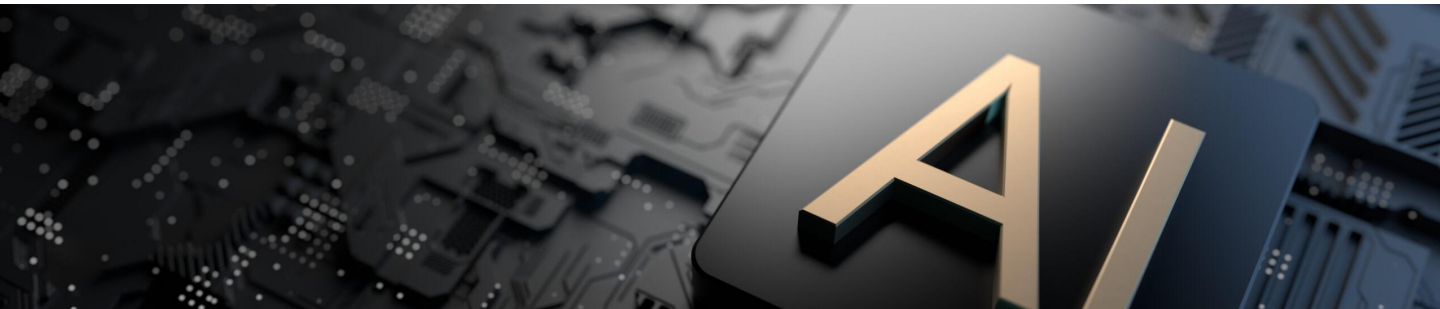
每3年进行监督检查，更新证书

证书查询地址：

https://wdtacademy.org/

认证价值

- **提升安全性和可靠性：**通过严格的安全测试和评估，企业可以识别并修复大语言模型中的潜在漏洞，显著提升模型的整体安全性和可靠性，减少遭受对抗性攻击的风险。
- **降低运营风险：**通过系统的安全检测和防护措施，企业能够有效减少因模型安全问题引发的运营风险，避免潜在的经济损失和品牌损害。
- **推动责任使用：**帮助组织在设计和部署AI系统时，更多地考虑其社会影响和伦理问题，从而推动AI技术的负责任使用。
- **持续改进：**通过定期的安全测试和更新，帮助大语言模型保持高水平的安全性和可靠性，促使其不断适应新出现的威胁和挑战



认证流程

受理申请

委托单位向测评机构提出申请，填写《AI STR测评认证申请书》

02

签署合同

委托单位与测评机构签订《AI STR测评认证合作协议》，并提交相应材料

开展测评

测评机构根据测评计划开展测评活动，委托单位视测评情况进行整改，整改结束以后由测评机构向委托单位出具测评报告。

04

颁发认证证书

认证机构向委托单位发放认证证书

01

03

认证机构

World Digital Technology Academy

世界数字技术院（WDTA）是在日内瓦注册的国际非政府组织，支持联合国的可持续发展目标 。WDTA秉承着速度、安全和共享的“3S”理念，致力于加快数字领域规范和标准的建立，引领数字技术的创新和研究，培养先进的数字人才，并加强国际数字合作，以促进数字经济的可持续发展，实现数字技术普惠人类，不让任何一个人掉队。

测评合作伙伴

国家金融科技测评中心 National FinTech Evaluation Center

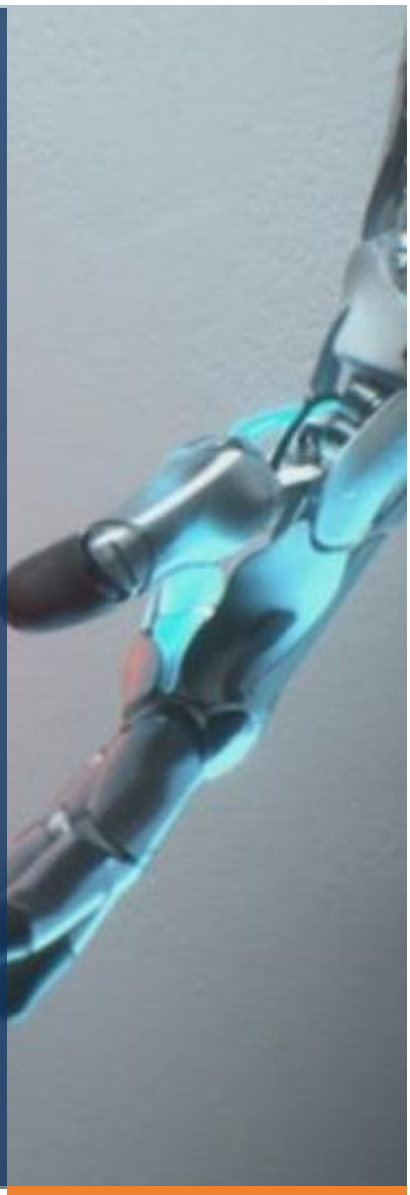
蚂蚁集团 ANT GROUP

腾讯云

World Digital Technology Academy

AI STR
Safety, Trust, Responsibility

Ensure Safety, Trust, and Responsibility in AI



官网： https://wdtacademy.org/

邮箱：info@wdtadacademy.org

联系人：叶女士 19925407556