

AI Verify 测试框架

资讯通信媒体发展局 (IMDA)

2022 年 5 月

中文翻译版

AI Verify 测试框架

概述

2022 年 5 月，新加坡资讯通信媒体发展局（Infocomm Media Development Authority, IMDA）推出 AI Verify—全球首个 AI 治理测试框架与工具包。AI Verify 将 AI 治理从“原则层面”推向“可操作层面”，帮助企业自主测试其 AI 系统的治理合规性。

核心特征

11 项可测试指标

AI Verify 围绕以下维度提供可量化的测试指标： - **透明度**: AI 决策过程的可解释程度 - **公平性**: AI 系统对不同群体的公平对待 - **安全性**: AI 系统的鲁棒性和可靠性 - **问责制**: AI 系统的治理和监督机制 - **以人为本**: 人类对 AI 决策的控制和参与

开源工具包

- 完全开源，企业可自由使用和定制
- 提供技术测试和流程检查两类工具
- 支持多种 AI 模型类型和应用场景

国际标准对齐

- 与 OECD AI 原则保持一致
- 参考 IEEE、ISO 等国际标准
- 支持跨境 AI 治理互认

AI Verify Foundation (AI Verify 基金会)

2023 年，IMDA 成立 AI Verify Foundation，推动全球协作： - 汇聚全球企业、研究机构和政策制定者 - 维护和发展 AI Verify 开源社区 - 推动 AI 治理测试标准的国际化

Global AI Assurance Sandbox (全球 AI 保证沙盒)

- AIVF 与 IMDA 联合推出
- 连接 AI 部署企业与测试服务提供商
- 在真实应用场景中测试 AI 治理方案

LLM 评估工具包

随着大语言模型 (Large Language Model, LLM) 的广泛应用, AI Verify 扩展了 LLM 评估工具包:

- 针对大模型的专项测试能力
- 评估生成内容的安全性和准确性
- 支持更快速、无缝的测试流程

使用场景

企业可以使用 AI Verify 来:

1. 评估 AI 系统是否符合治理原则
2. 识别 AI 系统中的潜在风险和偏见
3. 生成治理合规报告
4. 向利益相关方展示 AI 系统的可信度

意义

AI Verify 的创新之处在于将抽象的 AI 治理原则转化为具体的、可操作的测试工具。这一"原则→工具→实践"的路径已成为新加坡 AI 治理的标志性模式, 也为其他国家提供了可借鉴的经验。