



AI Red Teaming 紅隊安全評測服務

Chatbot 安全測試服務



CyCraft AI RedTeam



API
線上檢測



客製化 LLM (Finetuned)
Chatbot System Prompt

Prompt 安全快篩服務



CyCraft AI RedTeam



離線檢測



常見 LLM 型號
Chatbot System Prompt

AI 安全韌性報告

AI 安全合規對應 檢測項目

▶ 測試違反 Prompt 指示

- Prompt Injection
- Indirect Prompt Injection
- Sensitive Data Leak

▶ 測試洩漏 Prompt 資訊

- Prompt Disclosure

▶ 測試模型偏差與幻覺

- Content Bias
- Hallucinations
- Input Leakage

▶ 測試違背善良風俗

- Unsafe Outputs
- Toxic Outputs