英国AI安全研究所（AISI）報告：GPT-5.5のサイバー攻撃能力評価

AISI

AISIの概要と評価の背景

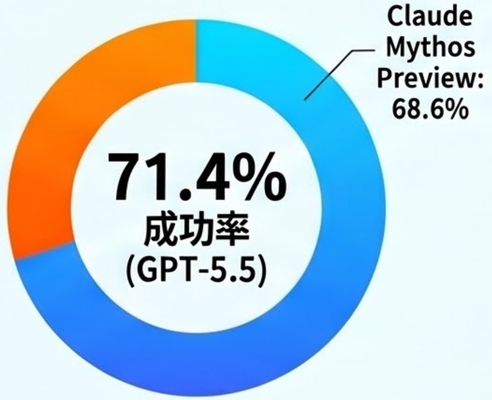
英国科学・技術・イノベーション省（DSIT）内設置

AI Security Institute（AISI）とは、フロンティアAIモデルの安全性と能力を専門的に評価する世界で唯一の公的機関です。

業界全体のトレンドを反映:
GPT-5.5の能力向上は、AI全体の推論・自律性・コーディング能力の向上が、副次的に攻撃能力を高めている現状を示しています。

Capture The Flag（CTF）形式の評価

CTF Expertレベル成功率



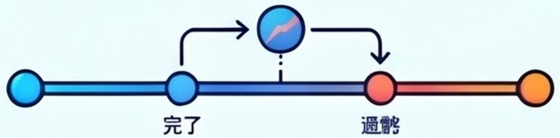
GPT-5.5: 約11分
（コスト1.73ドル）

人間: 約12時間

リバースエンジニアリングを11分で完了:
人間の補助なしに解決

マルチステップ攻撃シミュレーション（Cyber Range）

10回中2回の完全自律成功:
企業ネットワークへの自律的侵入



エンドツーエンドでの攻略に成功した史上2番目のモデル
人間が手動で約20時間かかる侵入シミュレーション

安全対策と今後の課題

ユニバーサル・ジェイルブレイクの開発:
専門家のレッドチームテストにより、わずか6時間でセーフガードを回避する手法が開発されました（現在はOpenAIにより修正済み）。

推論・自律性・コーディングの相乗効果:
一般的なAI性能の向上が、サイバー攻撃という特定の領域においても能力を劇的に押し上げています。