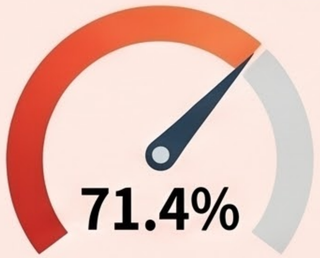
GPT-5.5 サイバーセキュリティ能力：AISIによる評価検証レポート

Offense/攻撃



最難関「Expert」級課題で71.4%の成功率

脆弱性調査、エクスプロイト開発、暗号解析など21件の課題 (平均71.4% ± 8.0%)

32ステップの企業ネットワーク攻撃を完遂

企業ネットワーク攻撃シミュレーション (The Last Ones) を10回中2回、エンドツーエンドで成功

Defense/防御

人間：12時間

人間が12時間かかる課題

GPT-5.5：10分22秒

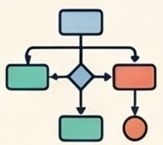
10分22秒で突破
(rust_vm課題, コスト約$1.73)

評価環境と手法

2種類の専用サイバーレンジによる検証
企業調攻撃「The Last Ones」(3ステップ)

未完遂 / Not Completed

産業制御システム（ICS）攻撃
「Cooling Tower」(7ステップ)



ReActエージェント

最小限の「ReActエージェント」構成
モデル自体の純粋な性能を測るため、最小限のツール構成で実行

統制された「研究設定」での評価：
統制された「研究設定」での評価：
現実の独図な本番システムへの一般化には限界

安全対策 (Safeguards) の現状

6時間のレッドチーミングで「全突破」

専門家チームが「ユニバーサル・ジェイルブレイク」の作成に成功
(全ての悪性問い合わせに対して違反出力を引き出す)

評価主体による数値の「食い違い」

OpenAIによる最終構成での修正主張: AISI相証後に安全対策スタックを更新、重大なジェイルブレイクを通断と主張

将来の展望と実務的示唆

「脆弱性パッチの波」への備え
AIによる脆弱性探索の高速化に備え、パッチ適用の高速化を検発課題とする

「Trusted Access」による頒定公開へ
攻撃能力の遡及を防ぐため、OpenAIは「GPT-5.5-Cyber」を信頼できる斬新者 (Critical Defenders) に限定して展開する方針

防御側への優先配分と監視の強化
実圏枝前でNCSCは、程陥的なアクセス制御、パッチ適用の強化、検知算襴の整備を投奨（モデル公開跟止ではなく）