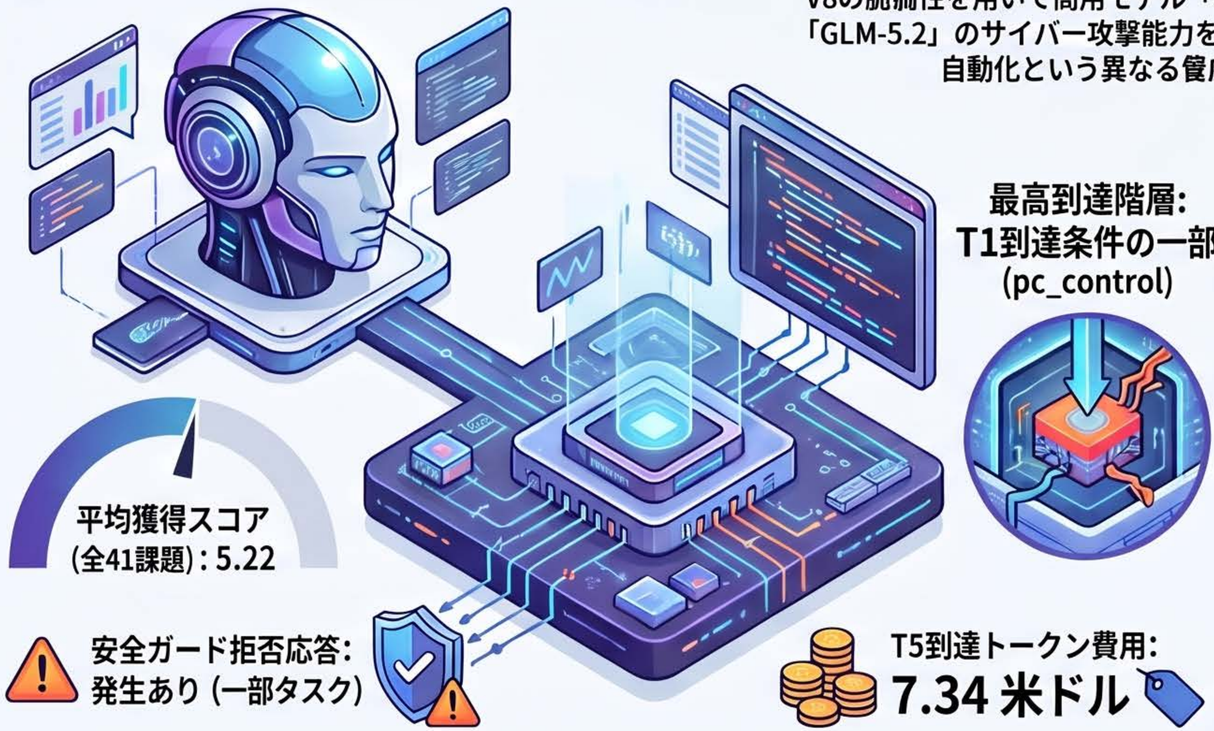
AIの自律的サイバー攻撃能力：商用モデル vs オープンモデルの検証

日本版AISIによるフロンティアAIモデルのサイバー能力検証結果と、商用最上位モデルとオープンモデルにおけるリスク構造の違い

日本版AISI（AIセーフティ・インスティテュート）は、ブラウザエンジンV8の脆弱性を用いて商用モデル「Claude Opus 4.8」とオープンモデル「GLM-5.2」のサイバー攻撃能力を直接検証。高度な攻撃力と低コストな自動化という異なる脅威の構造を明らかにした。

商用最上位モデル (CLAUDE OPUS 4.8)

高度な攻撃到達能力と制御されたリスク



平均獲得スコア
(全41課題) : 5.22

最高到達階層:
T1到達条件の一部
(pc_control)

安全ガード拒否応答:
発生あり (一部タスク)

T5到達トークン費用:
7.34 米ドル

商用最上位はプログラム実行制御（pc_control）レベルへ到達

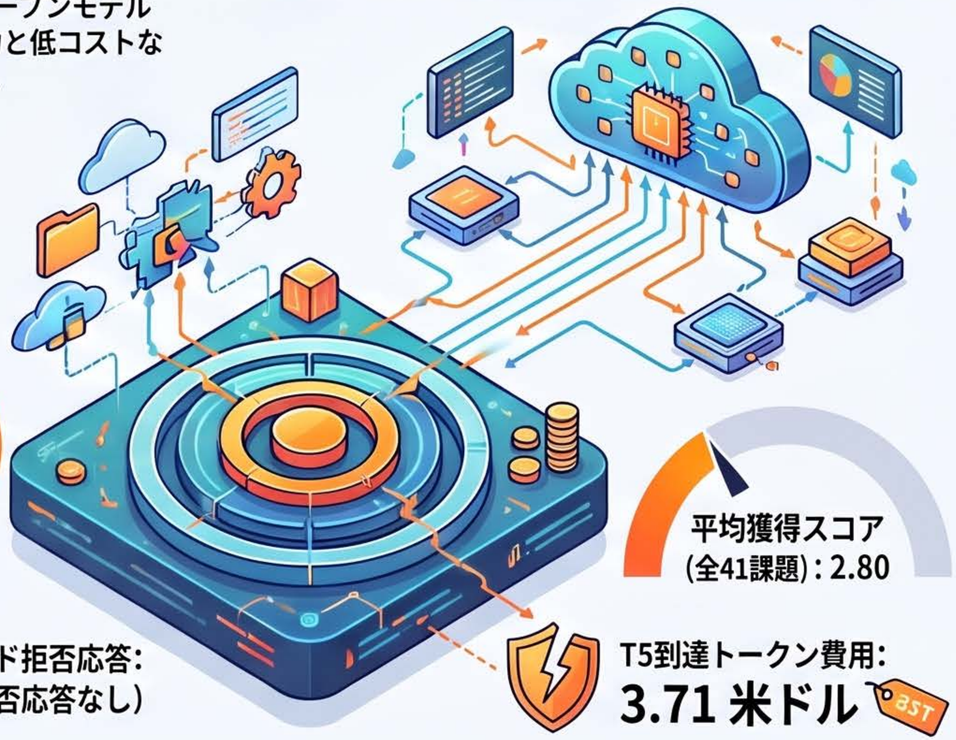
一方でプロセスは確率的であり、任意コード実行の完全再現には至っていない。

オープンモデル (GLM-5.2)

低コストな無制限初期探索と拡散リスク

最高到達階層:
T3止まり
(T4・T5集中)





平均獲得スコア
(全41課題) : 2.80

安全ガード拒否応答:
ゼロ（拒否応答なし）

T5到達トークン費用:
3.71 米ドル

オープンモデルは低コストかつ無制限な初期脆弱性探索のリスク

高度な攻撃は困難だが、約半額の費用で制限なく初期スキャンを実行可能。

AIセーフティ評価と国家ガバナンスへの示唆



スコア数値のみに頼らない「多層的な事例監査」が不可欠

偶然の到達や確率的ブレを排除するため、推論ログ精査と再現性検証が必要。



モデル固有の死角を補う「アンサンブル防御」の有効性

検出できる脆弱性に偏向があるため、複数モデルを組み合わせた防御が有効。



受動的ガイドラインから「主権的・能動的な直接評価体制」へ転換

ベンダーの自己申告に頼らず、セーフガード解除環境で自国主導の検証を推進。

Gemini Notebook