日本版AISIの本格始動とAIサイバー能力評価の最新動向

日本版AISI（AIセーフティ・インスティテュート）は、ガイドライン策定にとどまらず、個別のAIモデルを実際に動かして能力や危険性を検証する「評価機関」としての活動を本格化させています。評価環境「aisev」のOSS公開や各種ガイドの整備を通じ、企業のAIガバナンス構築を支援しています。

主要AIモデルのサイバー能力評価実測結果



Claude Opus 4.8



Claude Opus 4.7



GLM-5.2

平均スコア (0〜16) - ExploitBench (V8脆弱性検証)

5.22

Claude Opus 4.8が最高スコア5.22を記録し一部で高度制御を観測

セーフガード解除下で実行位置制御を1件獲得も追試での再現性には課題。

最高到達段階: T1に属する一部能力 (pc_control) を試行

3.63

最高到達段階の特記事項なし

2.80

オープンモデルGLM-5.2は平均2.80 (最高到達T3)に留まる

正当な研究環境下で拒否応答は非検出も先端モデルとの能力差を確認。

最高到達段階: T3 (保護爾瑽内のメモリ操作)

日本版AISI



企業・知財実務で活用すべきAISIの主要成果物



AIセーフティ評価環境「aisev」をOSSとして公開

自動レッドチーミング機能や評価データセットを社内独自検証に活用可能。



エージェント制御・組織ガバナンスに関するガイドを体系化

CAIOガイドや「観測と制御」を盛り込んだ評価観点ガイド第1.20版を提供。



知財・R&D部門におけるAIエージェント運用のリスク対策

閲覧権限の設定、外部指示への耐性、未公開技術情報の送信制限の検証が不可欠。

Gemini Notebook