NVIDIA Open Agent Safety Platform の全貌と評価

2026年9月にNVIDIAが発表した「Open Agent Safety Platform」は、AIエージェントの暴走やサンドボックス脱出を防ぐ安全基盤です。オープンソースのソフトウェアと専用ハードウェアによる帯域外監視を組み合わせた設計ですが、適用範囲やベンダーロックインなどの課題も指摘されています。

プラットフォームを構成する2つの基盤

OpenShell：オープンソースのソフトウェアランタイム

Apache 2.0で即時利用でき、カーネル隔離と外部制御を行う。



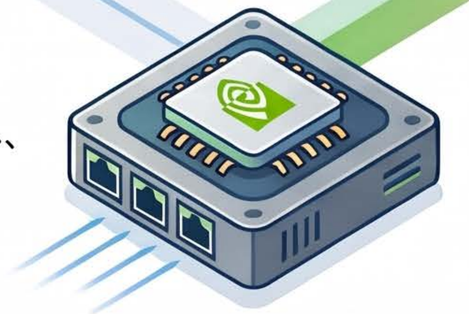
種別: ソフトウェア（ランタイム）

依存関係:
オープン（Arm/Intel対応）

提供状況:
即時利用可能（v0.1.2）

Sentry：DPU依存の帯域外リアルタイム監視

BlueField-4 DPU上で動作し、帯域外からミリ秒単位で隔離する。



種別:
リファレンス設計（監視役）

依存関係:
NVIDIA BlueField-4 DPU依存

提供状況:
未公表（時期・価格ともに未定）

アナリストの評価と現実的な限界

エージェント外部での強制制御を高く評価

プロンプト操作で解錠されないエージェント外側の強制力を評価。

企業のセキュリティ問題のカバー率は25%未満



管理外のSaaSエージェント等には効果がなく、カバー率は限定的。

ハードウェアロックインと実効性の検証不足

NVIDIAハード依存による縛りと実機事故での検証不足が懸念。

Gemini Notebook