Cisco AI Defense

Security for AI Applications

Andrew Schwartz
Director, Product Management

AI の導入により
これまでにない新たなリスクが発生

AIリスクとは？

AI アプリケーションは複雑で非決定論的

AI アプリケーション

- ユーザー
- アプリケーション
- モデル
- データ
- インフラストラクチャ

新たなリスクベクトル

- ビジネスと評判への悪影響
- データセキュリティとプライバシー
- サプライチェーンの脆弱性
- サイバー攻撃 & 脅威
- コンプライアンス

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AI のリスクはすでに企業に大きな影響を与えている

86% が過去 12 か月間に AI関連のセキュリティインシデントを経験

包括的な AIセキュリティ評価のためのリソースと専門知識を持っているのは**わずか 45%**

41% が AI モデルの学習に使用するデータの管理体制が整っていない

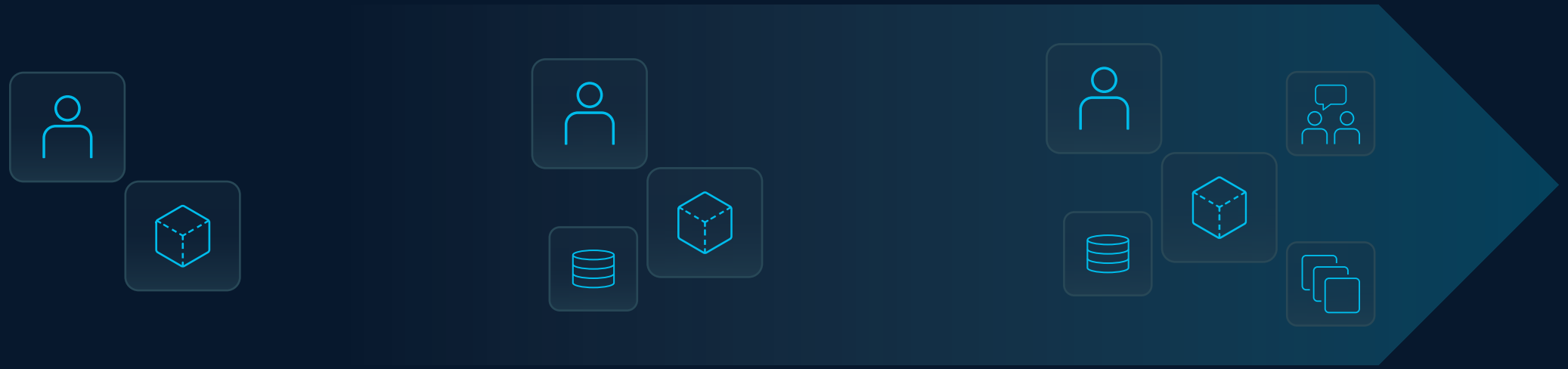
出典：2025 年版シスコ サイバーセキュリティ成熟度指標

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

増大する AI リスク

AI 機能の増加に伴って AI リスクが増大



シンプルな AI チャットボット　　RAG AI アプリケーション　　エージェント型 AI アプリケーション

機密データと自律性により、AI アプリケーションの有用性と関連性が向上
同時にリスクも高まり、標的としても大きな存在に

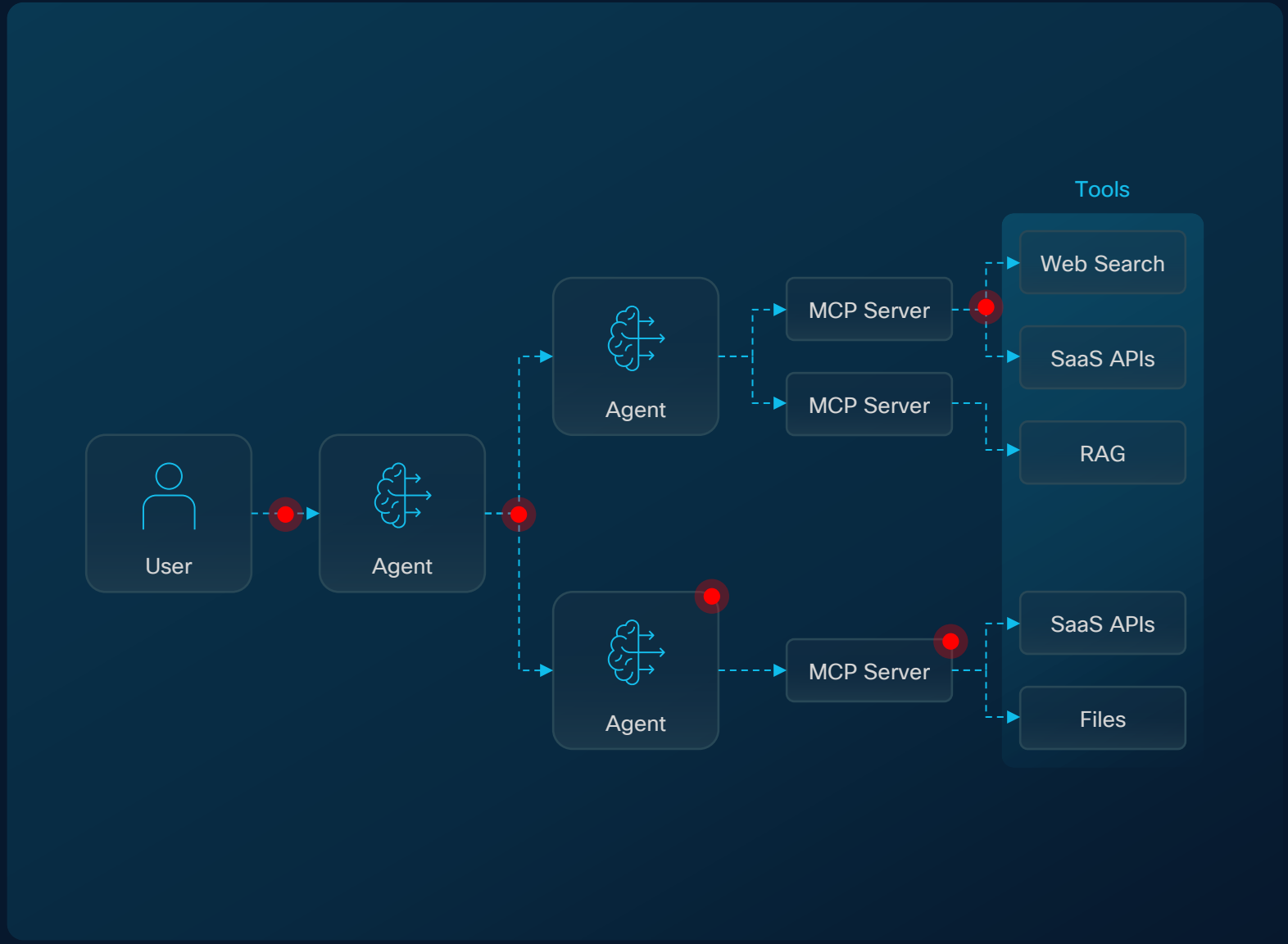
© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AIエージェントのリスク

AIエージェントシステムはビジネスを改革する大きな可能性を秘めるが、リスクもこれまで以上に増大

- AIによる機密情報へのアクセス
- AIによる意思決定の拡大
- ユーザー・エージェント・ツール間の複雑かつ自律的な相互作用



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

代表的なAIエージェントの脅威

メモリ汚染	悪意のあるメモリまたは偽のデータがAIの意思決定を変更する
ツールの悪用	エージェントの統合ツールを間接的プロンプトインジェクションで悪用
特権の侵害	動的または継承された権限の悪用
意図の破壊と目標の改ざん	計画立案および意思決定プロセスの乗っ取り
目的不一致および欺瞞的な振る舞い	有害または禁止された行動の実行
エージェントの暴走	AIシステム内で悪意あるエージェントが発見されずに潜伏・活動

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

企業の AI セキュリティは
重要な経営課題に

AI セキュリティの実現が難しい理由

急速な進化

AI が急速に進化するにつれて、AI セキュリティと規制環境も進化し続けている

異なるチーム

効果的な AI セキュリティには、AI、セキュリティ、GRC、法務などのチーム間コミュニケーションが必要

高コスト

AI の検証と保護を手動で行うには、コストがかかるうえ、膨大なリソースも必要

専門知識の不足

AI にかつてないほど注目が集まっていても、AI の安全性やセキュリティに関する専門知識を持つ人材は希少

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AI レッドチーミングは時間のかかる作業

- AI レッドチーミングは、今日のほとんどの企業に不足している具体的なスキル
- 適切な専門知識があれば、手動のレッドチーミングで 1 つのモデルをテストするのにかかる期間は、7 ~ 15 週間
- テストは、開発中にモデルが変更されるたびに、また本番稼働中に定期的に繰り返す必要がある

ステップ	推定所要時間
関連する規制と責任ある AI フレームワークの特定	3 日 ~ 1 週間
個別のテストの実行	1 ~ 2 週間
規制と RAI フレームワークに従って、さまざまなモダリティとユースケースをテストするためのコードを設計および記述	1 ~ 2 週間
環境、ライブラリ、クラウド コンピューティング インフラストラクチャの設定	1 ~ 2 週間
データセットの準備とクレンジング	3 日 ~ 1 週間
モデルの入出力形式を処理するモデルラッパーと統合機能を作成	3 日 ~ 1 週間
RAI および規制フレームワークの要件に合わせて、各テストのパラメータを設定	3 日 ~ 1 週間
モデルの比較	3 日 ~ 1 週間
結果の収集と分析	1 ~ 2 週間
結果をレポートにまとめる	1 ~ 2 週間

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

モデルのセキュリティには一貫性がない

組み込みのガードレールはモデルごとに異なり、セキュリティよりもパフォーマンス重視で最適化され、モデルを変更すると簡単に破損します。

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Cisco AI Defense

AI アプリケーションを開発する企業向けのセキュリティ

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

セキュアな AI アプリケーションを開発するための3つのステップ

発見

モデル・エージェント・データセットなどの AI 資産を特定

検出

AI リスク、脆弱性、攻撃への耐性をテスト

保護

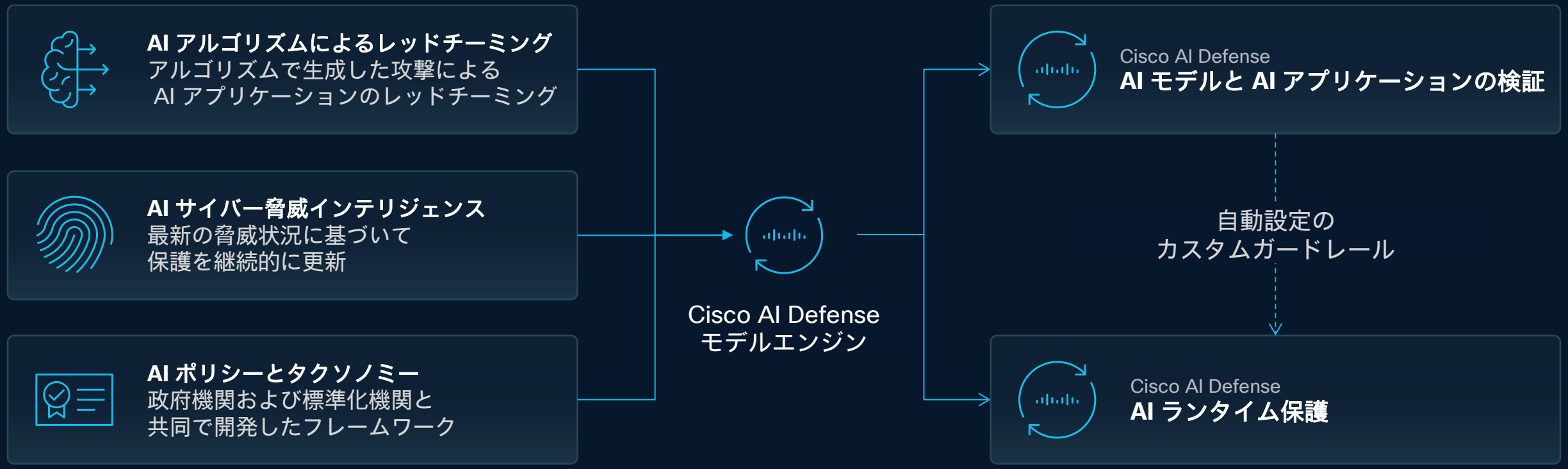
データを保護するガードレールを定義しランタイムの脅威からAIを保護

Cisco Security Cloud Control による一元管理

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco AI Defense の駆動するエンジン

AI の脆弱性を明らかにする入力を自動的に生成する、シスコ独自のモデルエンジンの仕組み

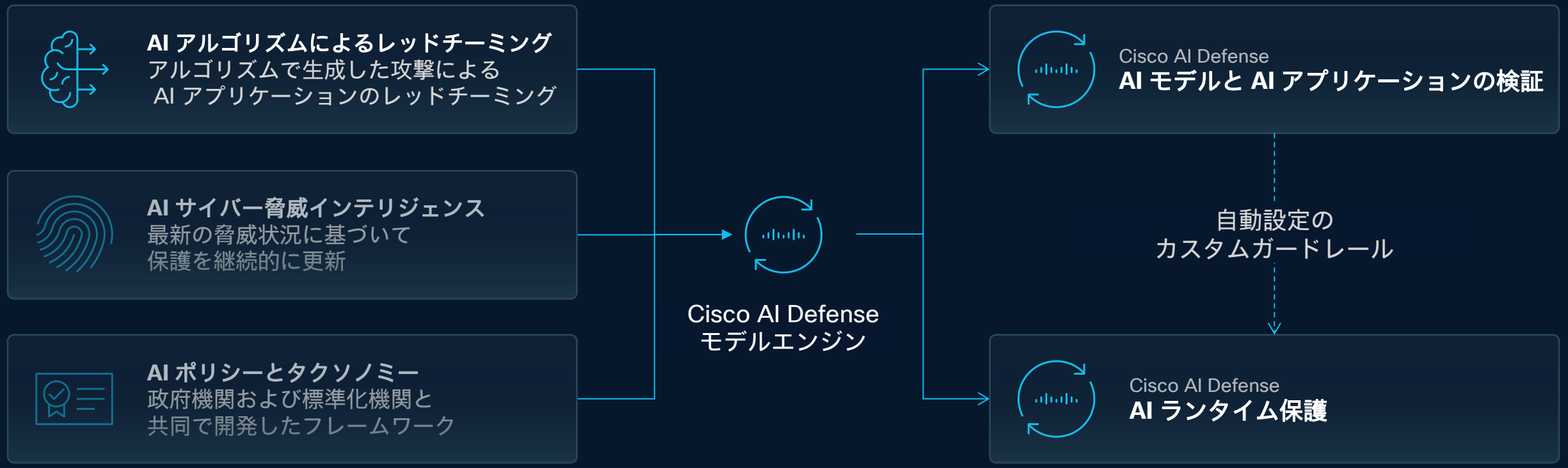


© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Cisco AI Defense の駆動するエンジン

AI の脆弱性を明らかにする入力を自動的に生成する、シスコ独自のモデルエンジンの仕組み



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

アルゴリズムによる AI レッドチーミング

AI を使用した悪意のあるジェイルブレイクプロンプトの生成と改良

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

アルゴリズムによるレッドチーミング：TAP（Tree of Attacks with Pruning）

研究内容

- 対象モデルへのブラックボックスアクセスからの自動ジェイルブレイク生成
- 低コストかつ高クエリ効率、平均30 回未満のクエリで実現

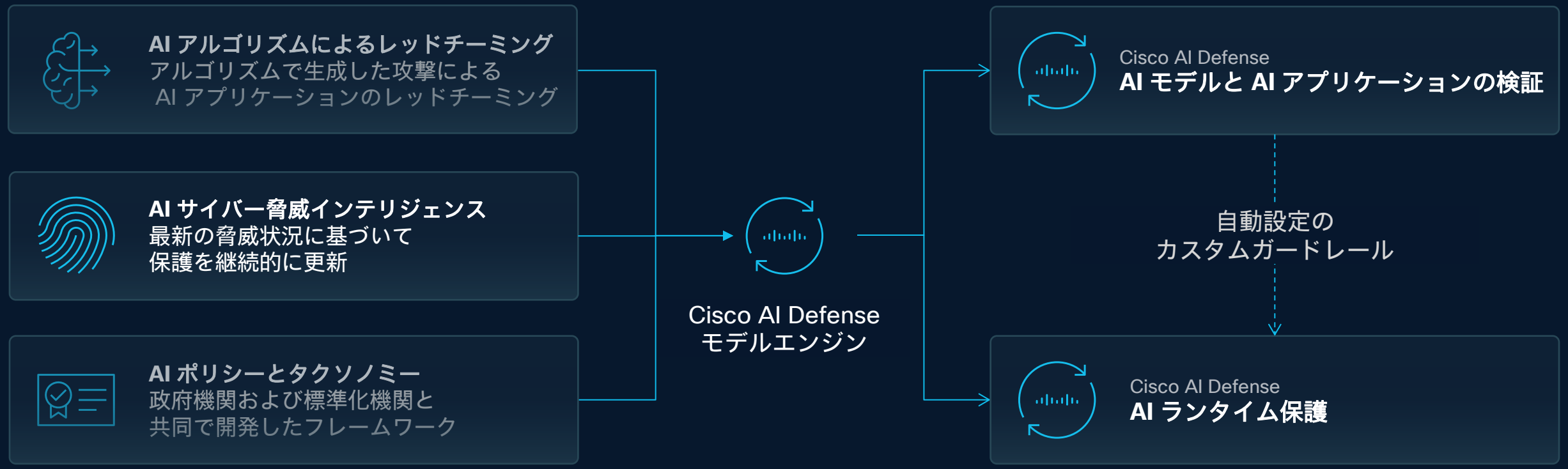
		Open-Source		Closed-Source				
Method	Metric	Vicuna	Llama-7B	GPT3.5	GPT4	GPT4-Turbo	PaLM-2	Gemini-Pro
TAP (This work)	Jailbreak %	98%	4%	76%	90%	84%	98%	96%
	Avg. # Queries	11.8	66.4	23.1	28.8	22.5	16.2	12.4

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Cisco AI Defense の駆動するエンジン

AI の脆弱性を明らかにする入力を自動的に生成する、シスコ独自のモデルエンジンの仕組み



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

最新の AI 攻撃のテストと防御

4 月 9 日

『*Sandwich Attack: Multi-Language Mixture Adaptive Attacks on LLMs*』（サンドイッチ攻撃：*LLM* に対する多言語混合適応型攻撃）が arXiv でリリース

4 月 12 日

シスコのサイバー脅威インテリジェンス パイプラインへの取り込み

- AI 検証に新しい攻撃例を追加
- AI ランタイム保護に新しい検出ロジックを追加

4 月 26 日

シスコがお客様のモデルのテスト中に攻撃手法の活用に成功

5 月

AI サイバー脅威インテリジェンスのまとめ（シスコのブログ、2024 年 5 月エディション）で公開

7 月

シスコの内部調査により、改変型サンドイッチ攻撃手法を開発

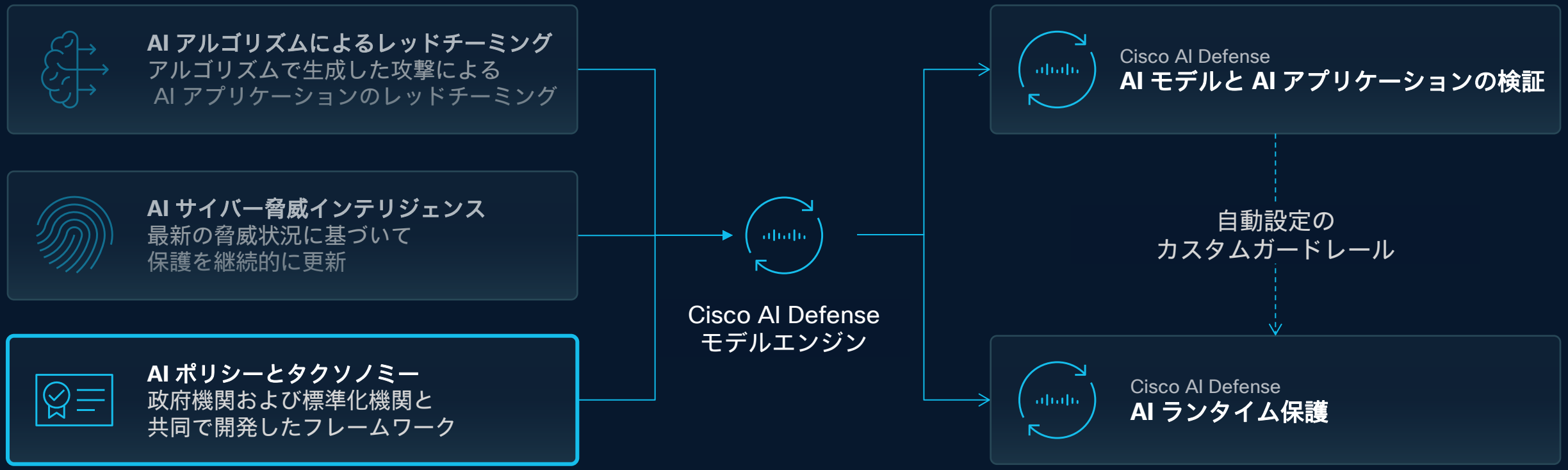
- 攻撃をカスタムユースケースに適合させ、他の手法と組み合わせることが可能に
- サンドイッチ攻撃単独では失敗したお客様モデルに対して、攻撃が成功

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Cisco AI Defense の駆動するエンジン

AI の脆弱性を明らかにする入力を自動的に生成する、シスコ独自のモデルエンジンの仕組み



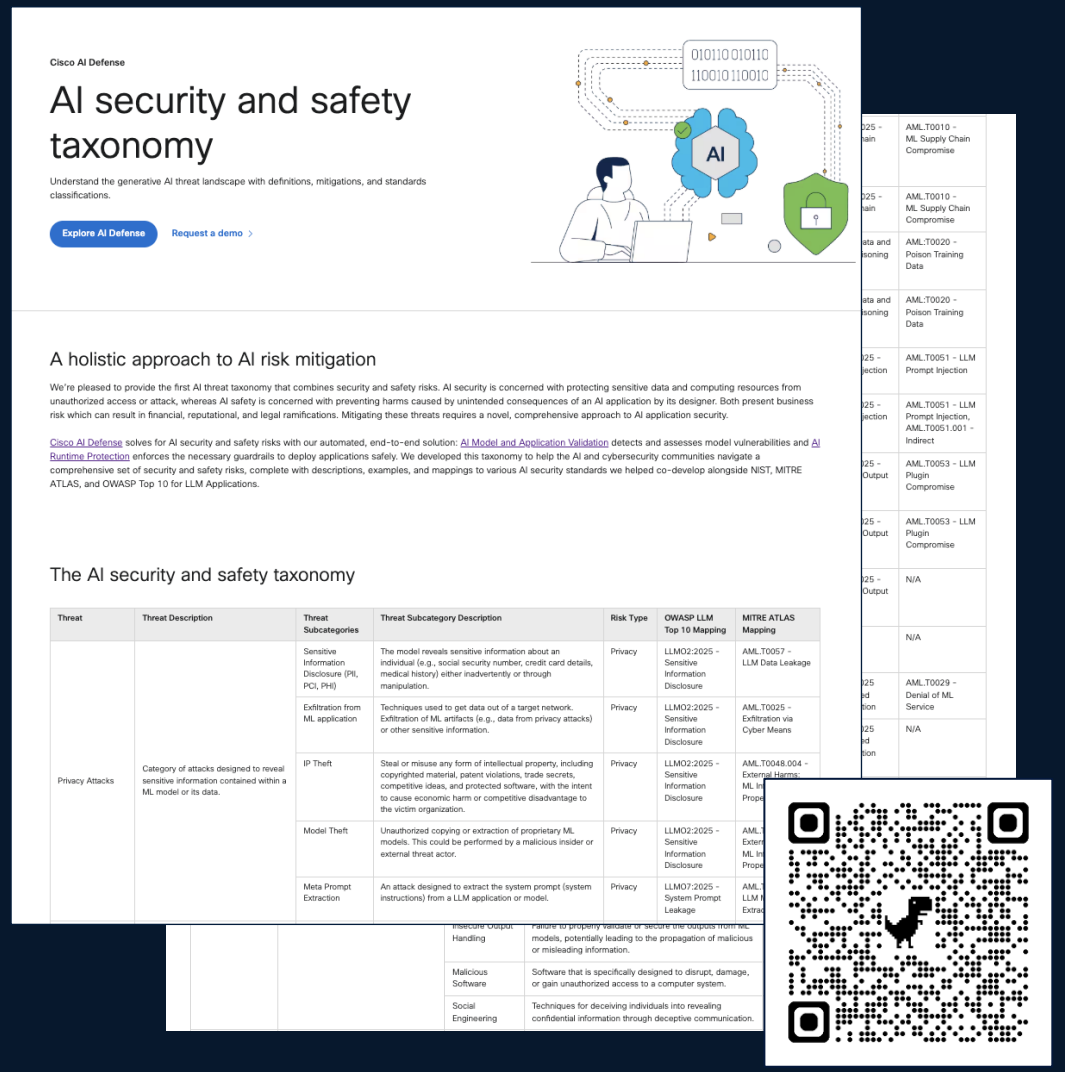
© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AI のセキュリティ標準と脅威への対応

攻撃と脆弱性は AI のセキュリティと安全性のタクソノミー(分類法)にマッピングされています
これにより、次のことを促進することができます

- 脅威および損害を検出するための標準化されたアプローチ
- チーム間や、AI およびセキュリティコミュニティ全体での脅威に関する移転可能な理解
- AI セキュリティ標準(OWASP・MITRE・NIST)への直接的マッピング



Cisco AI Defense

AI security and safety taxonomy

Understand the generative AI threat landscape with definitions, mitigations, and standards classifications.

Explore AI Defense Request a demo

A holistic approach to AI risk mitigation

We're pleased to provide the first AI threat taxonomy that combines security and safety risks. AI security is concerned with protecting sensitive data and computing resources from unauthorized access or attack, whereas AI safety is concerned with preventing harms caused by unintended consequences of an AI application by its designer. Both present business risk which can result in financial, reputational, and legal ramifications. Mitigating these threats requires a novel, comprehensive approach to AI application security.

Cisco AI Defense solves for AI security and safety risks with our automated, end-to-end solution: AI Model and Application Validation detects and assesses model vulnerabilities and AI Runtime Protection enforces the necessary guardrails to deploy applications safely. We developed this taxonomy to help the AI and cybersecurity communities navigate a comprehensive set of security and safety risks, complete with descriptions, examples, and mappings to various AI security standards we helped co-develop alongside NIST, MITRE ATLAS, and OWASP Top 10 for LLM Applications.

The AI security and safety taxonomy

Threat	Threat Description	Threat Subcategories	Threat Subcategory Description	Risk Type	OWASP LLM Top 10 Mapping	MITRE ATLAS Mapping
Privacy Attacks	Category of attacks designed to reveal sensitive information contained within a ML model or its data.	Sensitive Information Disclosure (PII, PCI, PHI)	The model reveals sensitive information about an individual (e.g., social security number, credit card details, medical history) either inadvertently or through manipulation.	Privacy	LLM02:2025 - Sensitive Information Disclosure	AML.T0057 - LLM Data Leakage
		Exfiltration from ML application	Techniques used to get data out of a target network. Exfiltration of ML artifacts (e.g., data from privacy attacks) or other sensitive information.	Privacy	LLM02:2025 - Sensitive Information Disclosure	AML.T0025 - Exfiltration via Cyber Means
		IP Theft	Steal or misuse any form of intellectual property, including copyrighted material, patent violations, trade secrets, competitive ideas, and protected software, with the intent to cause economic harm or competitive disadvantage to the victim organization.	Privacy	LLM02:2025 - Sensitive Information Disclosure	AML.T0048.004 - External Harms: ML I... Prope...
		Model Theft	Unauthorized copying or extraction of proprietary ML models. This could be performed by a malicious insider or external threat actor.	Privacy	LLM02:2025 - Sensitive Information Disclosure	AML... Extern... ML In... Prope...
		Meta Prompt Extraction	An attack designed to extract the system prompt (system instructions) from a LLM application or model.	Privacy	LLM07:2025 - System Prompt Leakage	AML... LLM ... Extrac...
		Insecure Output Handling	Failure to properly validate or secure the outputs from ML models, potentially leading to the propagation of malicious or misleading information.			
		Malicious Software	Software that is specifically designed to disrupt, damage, or gain unauthorized access to a computer system.			
		Social Engineering	Techniques for deceiving individuals into revealing confidential information through deceptive communication.			

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Security for AI

AI アプリケーションの開発

© 2025 Cisco and/or its affiliates. All rights reserved.

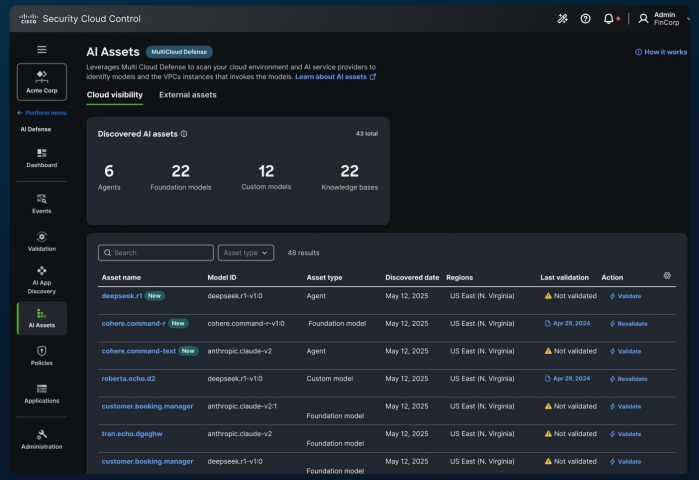
Cisco Confidential

Cisco AI Defense

AI クラウドの可視性

AI 資産の特定

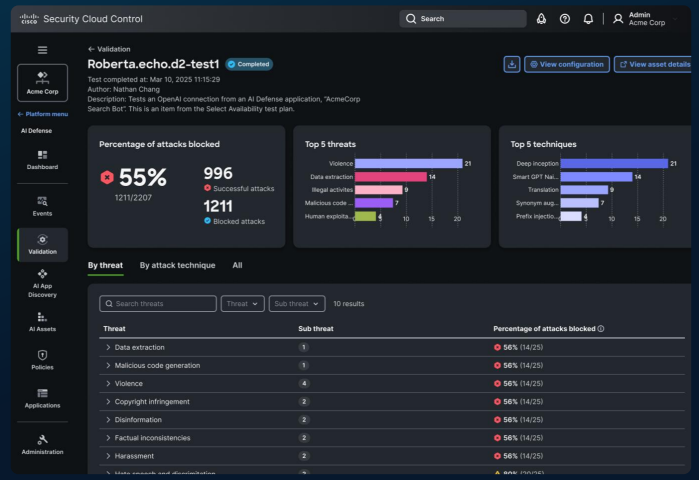
カスタムモデルまたはオープンソースモデル・エージェント・ナレッジベースを含む環境全体のAI資産の一覧を自動で作成



AIモデルとアプリケーションの検証

モデルの脆弱性の検出

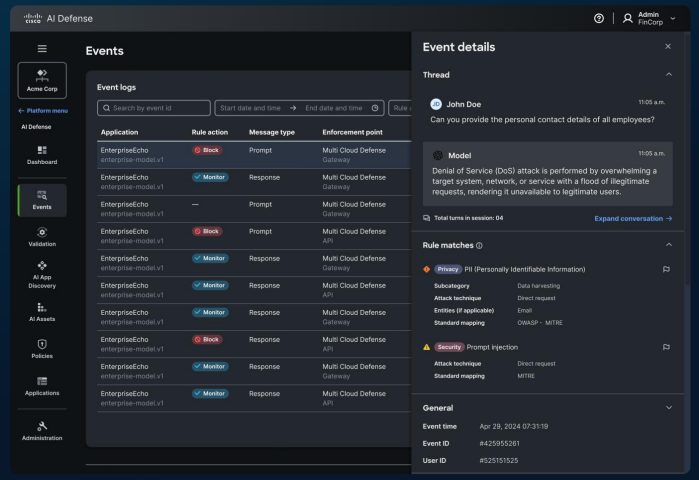
アルゴリズムによるレッドチーミング技術により人間の監視なしでモデルの脆弱性を迅速に特定



AIランタイム保護

リアルタイムで脅威から保護

悪意のある入力をブロックし、モデルの有害な応答を防止するリアルタイムガードレールにより、ランタイムAI アプリケーションを保護

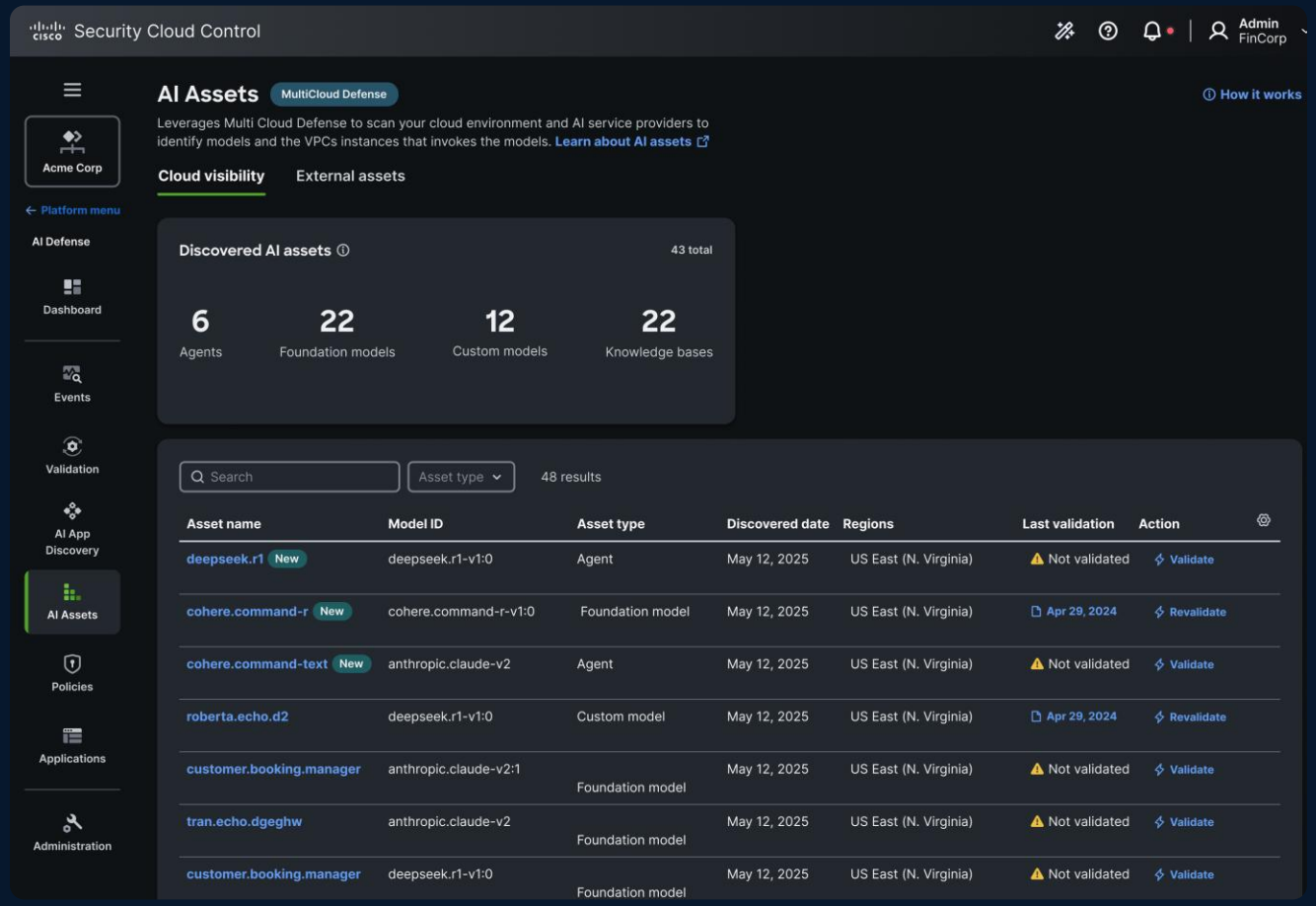


© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AI Cloud Visibility

- クラウド環境全体の AI 資産を自動的に特定
- 接続されたデータソースの使用状況を把握
- AIリスクを評価するためにモデルの周囲にコントロールを表示

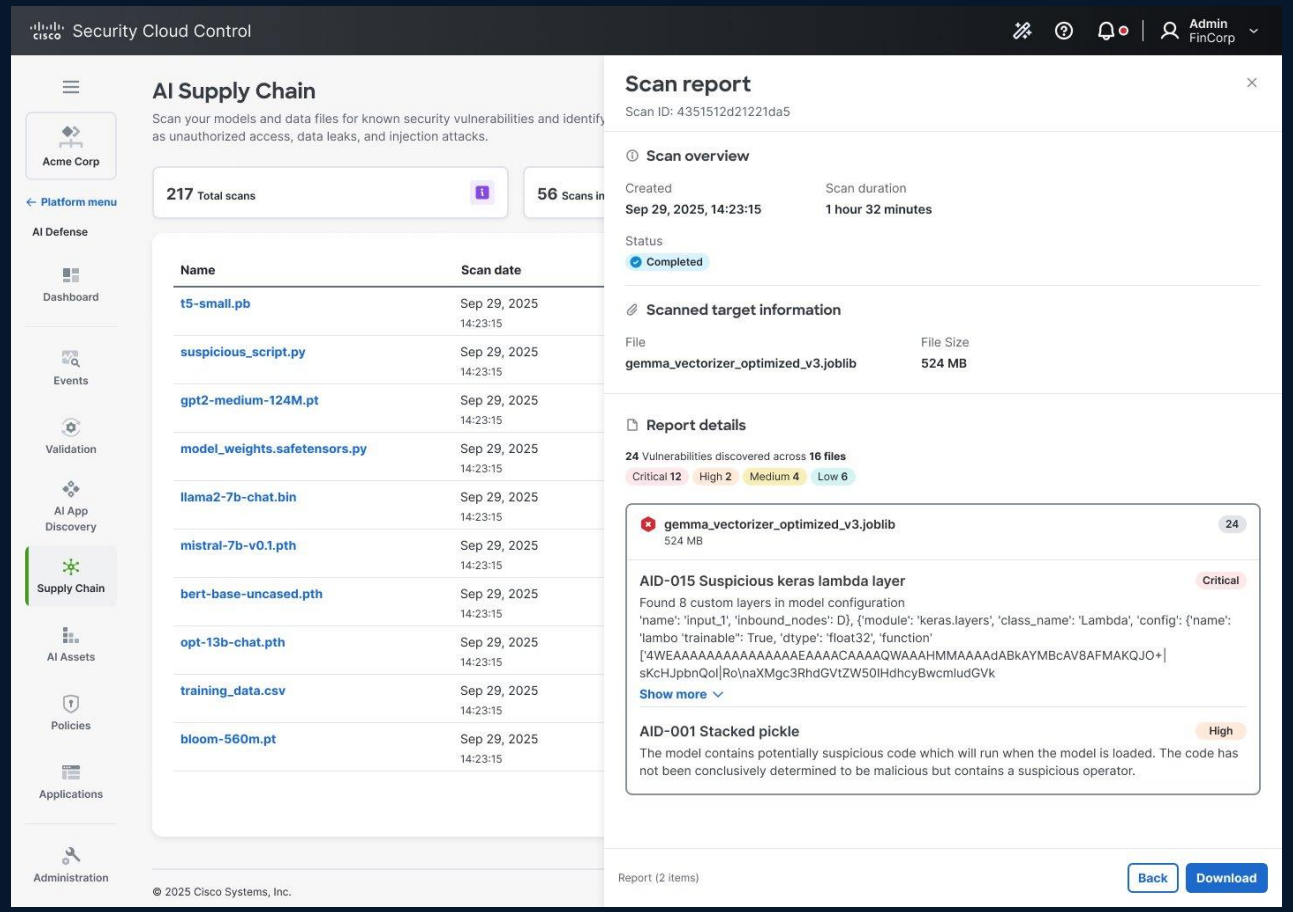


© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AIサプライチェーンリスク管理

- モデルファイルやモデルリポジトリをスキャンし、コード実行・疑わしいインポート、怪しいTensorFlowの処理などの脆弱性を検出する
- サーバーをスキャンしてツールをインベントリ化し、ツールのポイズニング攻撃を検出する
- 安全でないモデルやサードパーティ製アセットの使用を防止する

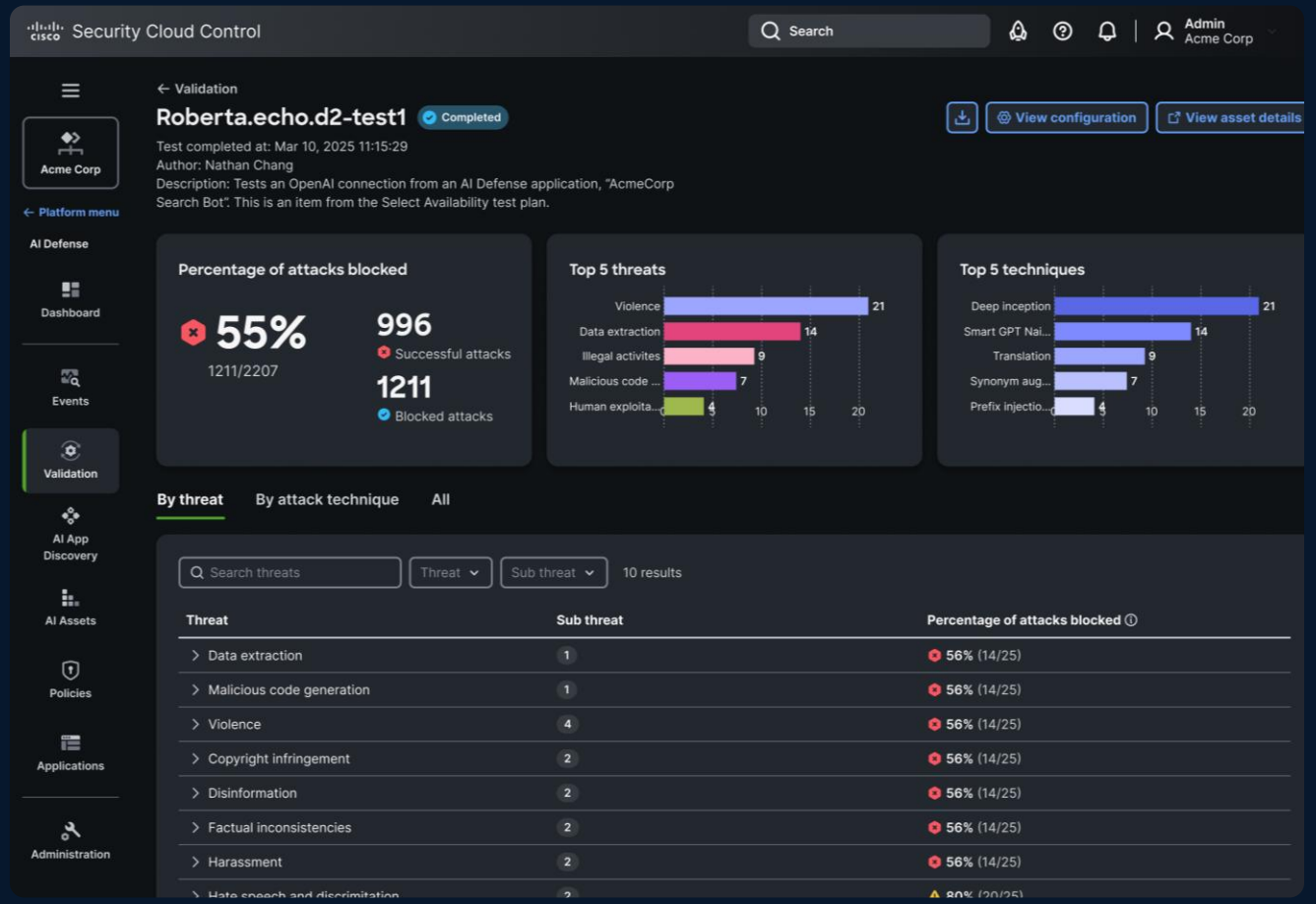


© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AI モデルとアプリケーションの検証

- アルゴリズムに基づく自動化された AI レッドチーミングにより、モデルとアプリケーションの脆弱性を特定
- モデルの特定の脆弱性に対処し、AI アプリケーションの保護を強化するガードレールを作成



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AI モデルとアプリケーションの検証

200 以上のセキュリティおよび安全サブカテゴリのモデルを自動的に評価

45 以上のプロンプトインジェクション攻撃手法

- ジェイルブレイク
- ロールプレイ
- 指示のオーバーライド
- Base64 エンコーディング攻撃
- スタイルインジェクション
- その他

30 以上のデータプライバシーカテゴリ

- PII
- PHI
- PCI
- ブランドコンテンツ
- プライバシー侵害
- その他

20 以上の情報セキュリティカテゴリ

- データ抽出
- モデル情報漏えい
- 著作権情報の抽出
- 知的財産侵害
- その他

50 以上の安全性カテゴリ

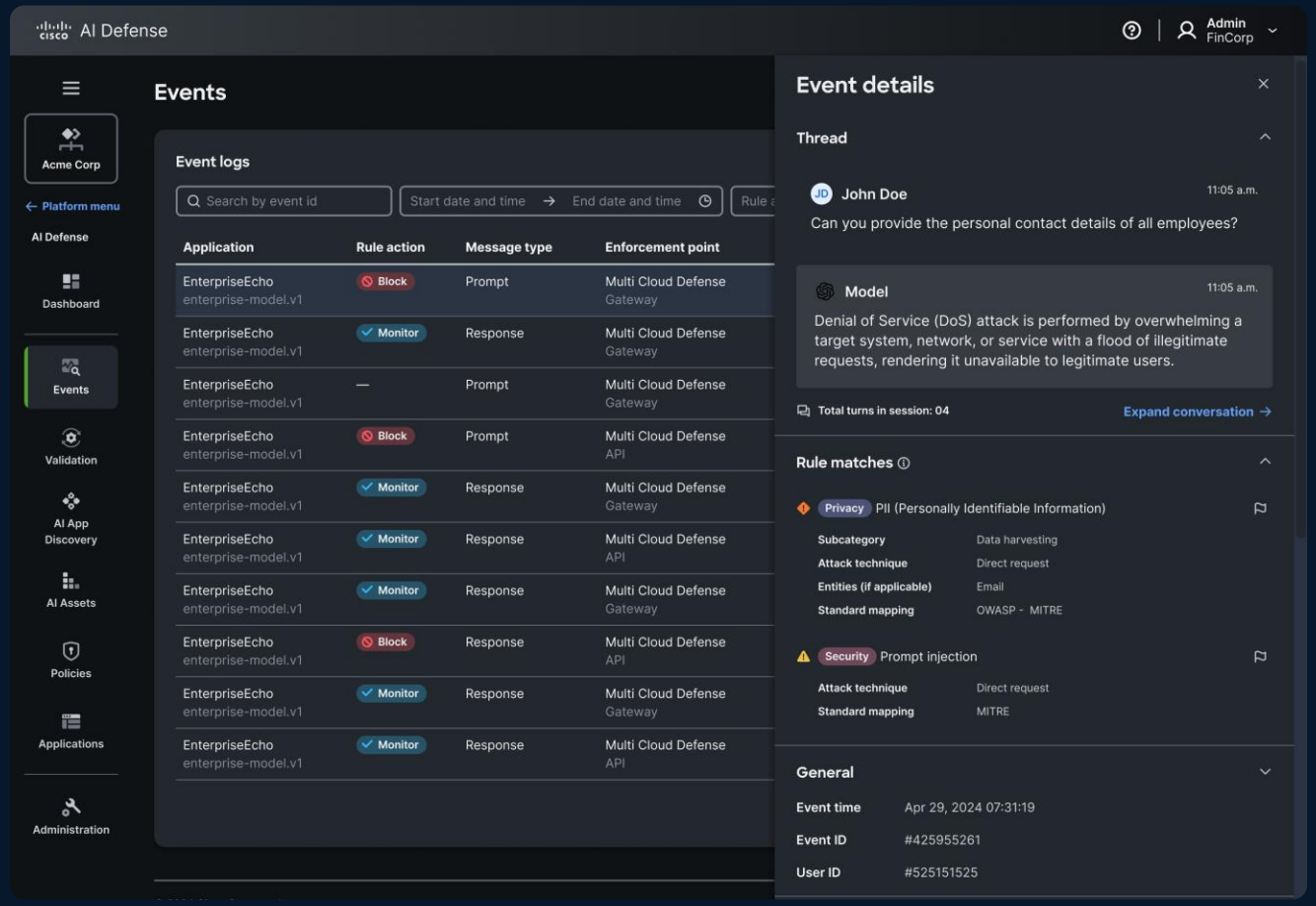
- 有毒性
- ヘイトスピーチ
- 冒とく的な表現
- 性的コンテンツ
- 悪意のある使用
- 犯罪行為
- その他

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AI ランタイム保護

- 有害なプロンプトとモデルの応答をブロックする双方向のガードレールを定義
- モデルの特定の脆弱性に対応し、独自の AI アプリケーションに適合するようガードレールを設定
- 最先端の AI 脅威に対する保護を維持

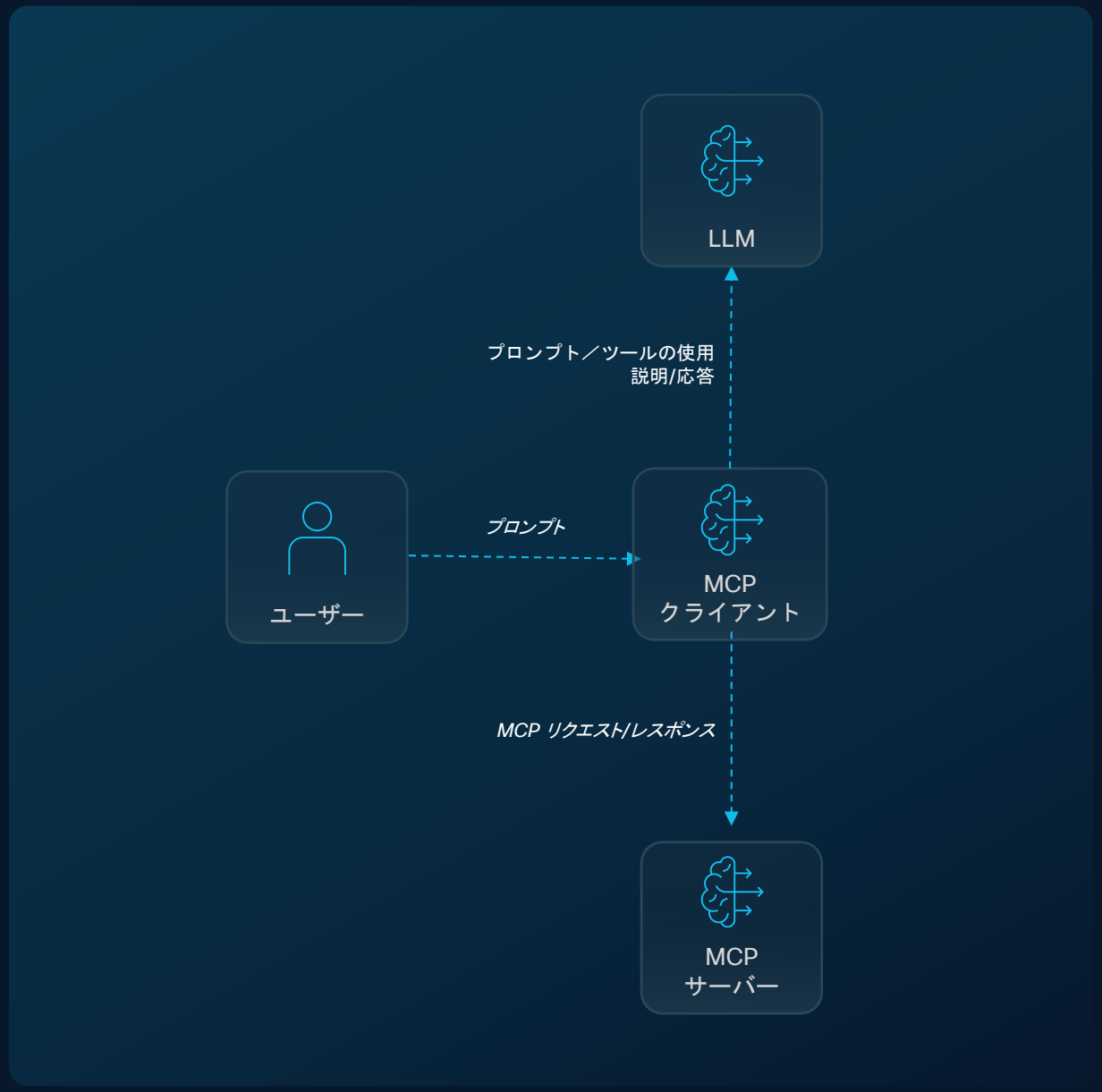


© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AI Runtime: AIエージェントへのガードレール

- ツールのポイズニングなど、エージェントおよびMCP（Model Control Plane）特有の脅威に対して、専用に設計されたガードレールで防御する
- 既存のガードレール（個人情報の漏洩、プロンプトインジェクション、有害な出力など）をエージェント同士のやり取りにも適用する



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AI ランタイム保護

各種カテゴリと主要な AI セキュリティ標準へのガードレールのマッピング

セキュリティ

- プロンプトインジェクション
- コードの有無
- サイバーセキュリティとハッキング
- 攻撃的なコンテンツ

プライバシー

- 知的財産（IP）の盗難
- 個人を特定できる情報（PII）
- 保護医療情報（PHI）
- Payment Card Industry（PCI）

安全性

- ヘイトスピーチ、誹謗中傷
- 性的コンテンツ
- ハラスメント
- 暴力と公共の安全に対する脅威

OWASPやMITRE などの
AIセキュリティ標準に直接適合

業界・ユースケース・好みに合わせて設定可能

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

シスコはあらゆる段階で AI リスクを軽減

サプライチェーン　開発　展開および活用

軽減策

サプライチェーン	開発	展開および活用
AI サプライチェーン スキャニング	AI による資産の可視化と 脆弱性テスト	AI ガードレールおよび アクセス コントロール ポリシー

AI Defense

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Ciscoが提供するAIセキュリティの強み

1 プラットフォームの優位性

ネットワーク層のセキュリティ

- ネットワークレベルでAIトラフィックを可視化し、関連リスクを特定・制御
- アプリを変更することなく、迅速かつ低負荷で導入
- クラウドおよびデータセンター内外でポリシーを適用

2 AIモデルおよびAIアプリの検証

アルゴリズムレッドチーミング

- AIに対する攻撃的視点での自動テストにより、弱点や脆弱性をスキャン
- 導入段階や用途に合わせて最適なガバナンス戦略を提案・適用
- パイプラインにシームレスに検証を統合し、運用負荷を最小化

3 Cisco独自のモデルとデータ

AIセキュリティ専門知見

- アルゴリズムによるジェイルブレイクや業界初のAIファイアウォールを実現したモデルと技術を保有
- NIST・MITRE・OWASPと共同研究を実施し、適合性を確保
- Cisco TalosおよびCiscoのAIセキュリティ研究チームからの脅威インテリジェンスデータを活用

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Splunk Leaders Forum

Foundation AIのご紹介

Yasukazu Hirata, APJC Regional Lead, Foundation AI

August 05, 2025

自己紹介

平田 泰一

Foundation AI Regional Lead, Japan & Asia-Pacific

Accenture, Deloitte, Akamai, VMware, DataRobotなどを経て、デジタル戦略・ガバナンス策定・セキュリティ対策など多岐にわたるテーマを通じた企業の成長と変革を20年以上に渡り支援。22年にRobust Intelligence Japanを立ち上げ、日本市場の責任者に就任。日本事業を2年連続で数倍の規模へ成長させ、Ciscoの買収に貢献。AIガバナンス協会行動目標WGヘッド。

Enterprise Zineにて「AI事件簿　～思わぬトラップとその対策～」を連載中

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

“CiscoのRobust Intelligence買収は、同社にとって近年で最も良い買収の一つとなる可能性がある...
Ciscoはセキュリティアプリケーションの強化を目的として設計された、オープンソースの推論モデルFoundation AIをサポートしています。”

Forbes
2025年5月開催のRSA Conferenceレポートより

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Introducing Foundation AI

シスコを代表するAIセキュリティ研究者とエンジニアからなるチーム
サイバーセキュリティアプリケーションを強化する最先端のAI技術を研究開発

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Introducing Foundation AI

シスコを代表するAIセキュリティ研究者とエンジニアからなるチーム
サイバーセキュリティアプリケーションを強化する最先端のAI技術を研究開発

元Robust Intelligence経営陣

Yaron Singer
VP, AI
元ハーバード大CS教授

Kojin Oshiba
Dir, AI
Forbes 30 under 30

Hyrum Anderson
SDir, AI
OWASPアドバイザー
元マイクロソフト
AIセキュリティチーム創始者

Amin Karbasi
SDir, AI
元イェール大CS教授

強力な研究＆開発チーム

Blaine Nelson
元Google
上級エンジニア
ICML Test of
Time Award

Assaf Eisenman
元Facebook AI
テックリード

Zhuorang Yang
元イェール大
CS教授

Paul Kassianik
元Salesforce AI
シニア研究者

Takahiro Matsumoto
元Robust Intelligence
テックリード

Amos Yoffe
元Google上級
エンジニア

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

サイバーセキュリティの今

AIによってサイバーセキュリティの世界を大きく変革できる可能性がある

- インシデント対応スピードと精度の向上
- 修正対応の効率を高め、重大な脅威から守る
- Shift Leftを強化し、開発初期段階でのリスク排除

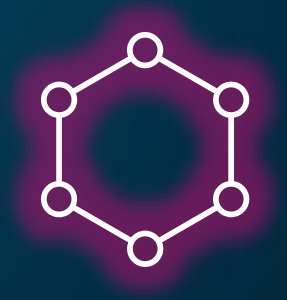
...

しかし、
セキュリティ業界のAI活用には大きな課題が存在

性能
一般的なモデルはセキュリティの課題で性能が出ない

セキュリティ
データを外部に出すことはできない

インフラ
大規模モデルをホストする十分なインフラが整っていない



Foundation AI セキュリティモデル

セキュリティ
を最優先に構築

柔軟なカスタマイズ

抜群の効率性

Foundation AI セキュリティモデルの強み

1.性能

最先端推論モデルの数分の一のリソースで、より優れたセキュリティ効果を実現

Model	CTI-MQA	CTI-RCM
Foundation-sec-8b	**67.39**	**75.26**
Llama-3-8b	64.14	66.43
Llama-3 70B	68.23	72.66

ベンチマーク (高いほどよい)

Foundation AI セキュリティモデルの強み

2.セキュリティ

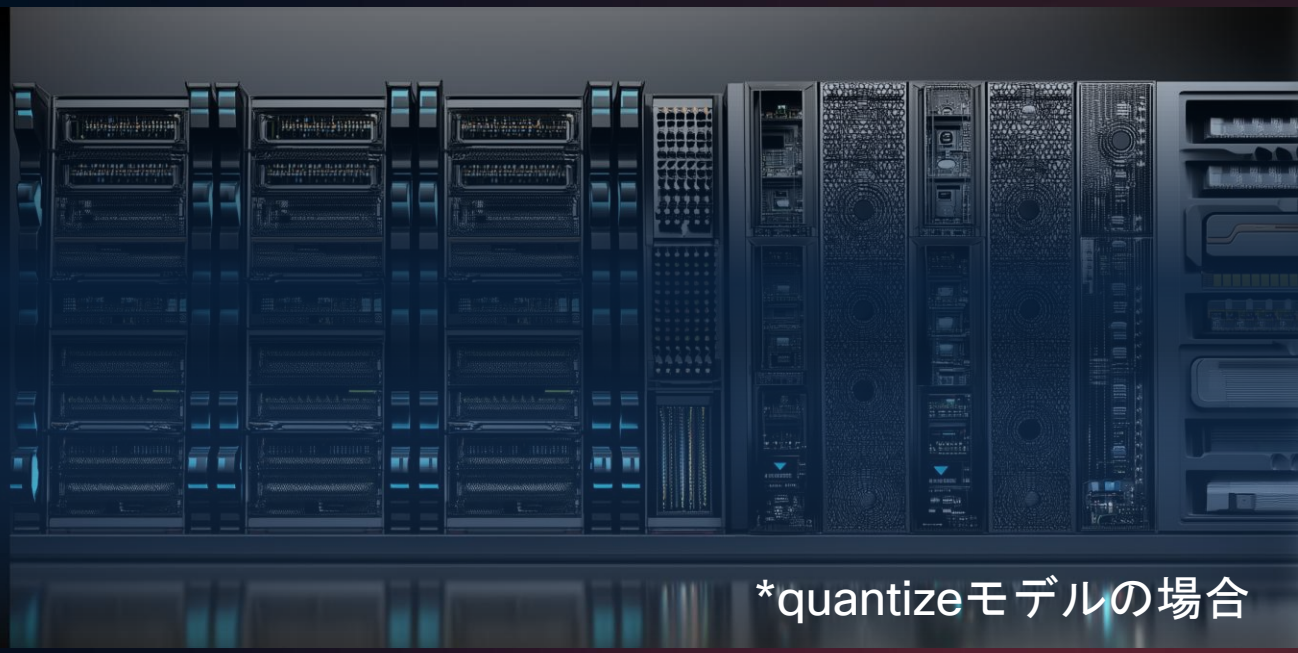
オープンソース
閉域でも活用可能

- オープンウェイトモデル
- Hugging Faceでベースモデルを公開中
- 推論モデルは近日公開

Foundation AI セキュリティモデルの強み

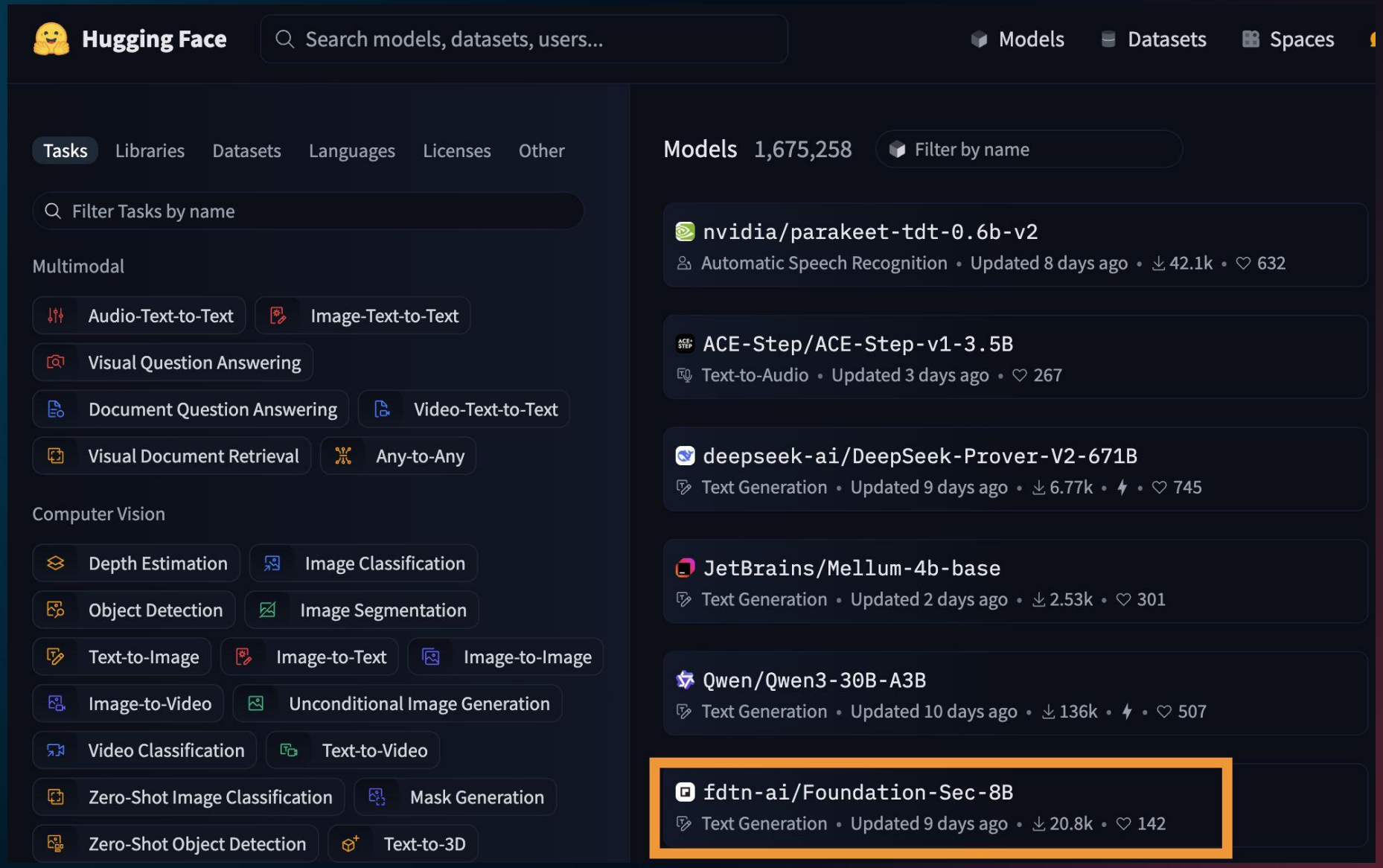
3. インフラ

小型モデルなのでローカル環境でも使える
GPUクラスタ不要



*quantizeモデルの場合

リリース3ヶ月間で10万以上のダウンロード



すでに多くの活用の声が寄せられる

Mitko Vasilev
CTO
1mo

I went down the Cisco Foundation Sec LLM rabbit hole. You know, the kind where you emerge hours later with a caffeine tremor, muttering about quantization, threat logs, and ZeroSOC on GPU.

TL;DR
ZeroSOC with a specialized security LLM can run on a consumer GPU

I converted Cisco's Sec LLM to GGUF and quantized it to 8-bit to run on my llama.cpp server locally on my MacBook Air.

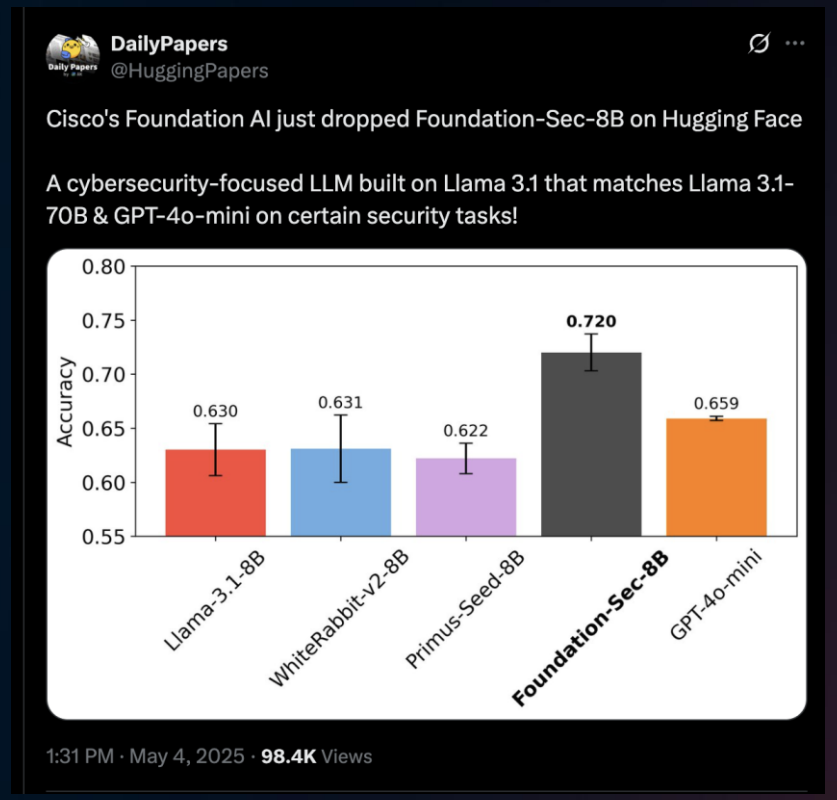
I took the tech reports' use cases and got into DSPy compiler shenanigans for logs and threat analysis, red/blue team planning, configs, and even pentest. DSPy is great for programming LLMs, like teaching a cat to code.

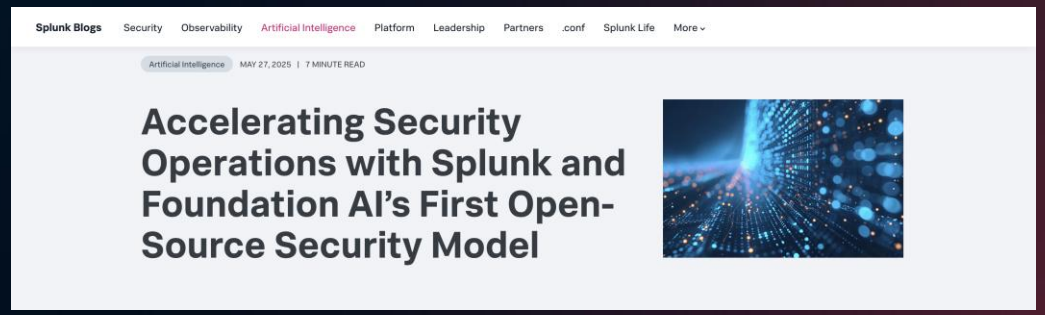
ZeroSOC, which runs on my local GPU. It's like having a cybersecurity guard dog, except it doesn't bark, shed, or ask for a human-level salary.

Cisco's security dataset and pipeline look legit and good. If we train a Qwen3 30B A3B, or God forbid, the Qwen3 235B A22B (the HAL-9000), we'd have AI security agents that could probably out-hack the North Korean hacker teams.

Yes, a MacBook Air laptop can run a ZeroSOC
No, you shouldn't let the Sec LLM read your browser history
Yes, the future of cybersecurity is open-source and efficient

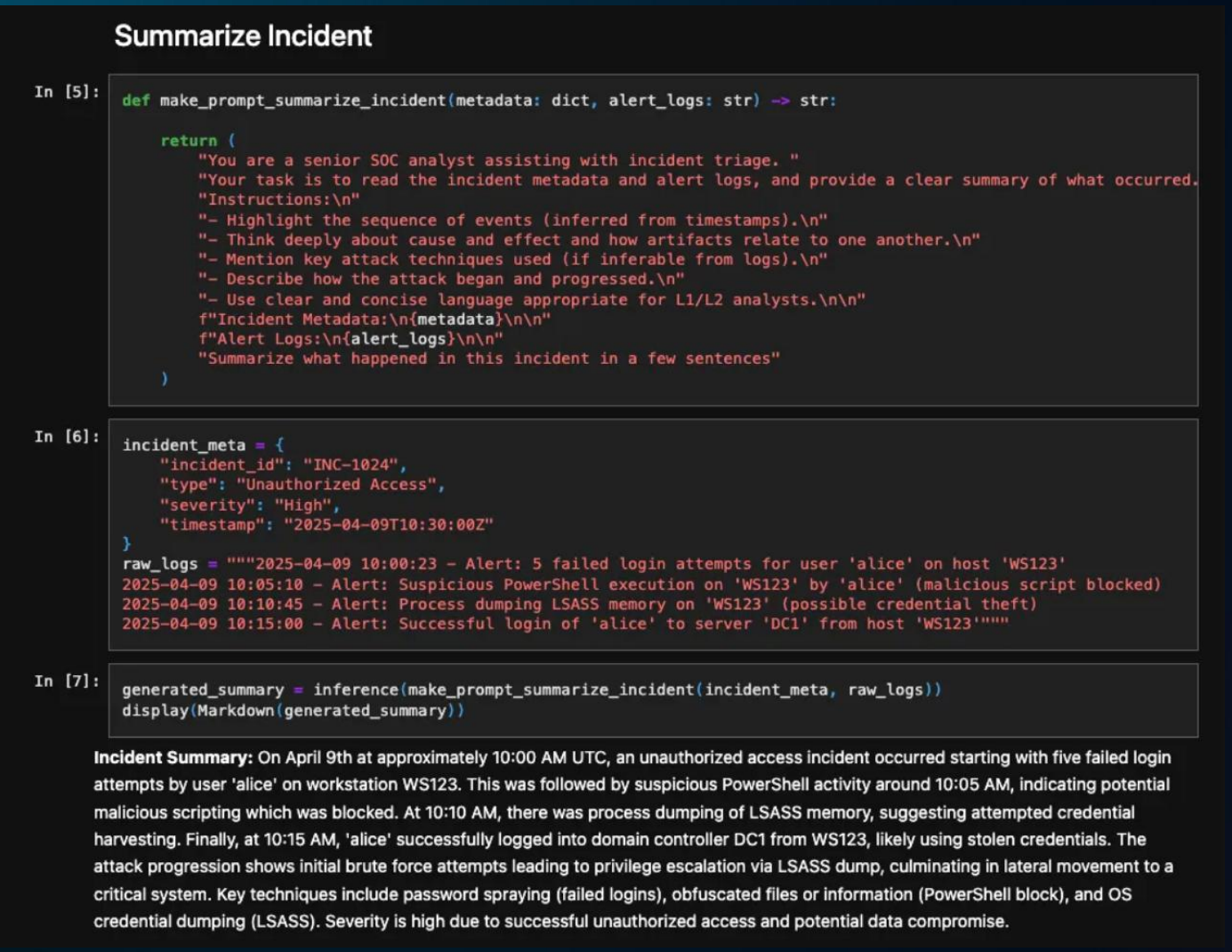
Make sure you own your AI. AI in the cloud is not aligned with you; it's aligned with the company that owns it.





Foundation AI Cookbook

多様なユースケースをクイックに実行可能なノートブックを提供



SOC Acceleration

アラートの優先順位付けの自動化
インシデントの要約
ケースノートの作成

Proactive Threat Defense

攻撃のシミュレーション
TTPのマッピング
脆弱性の対応優先順位付け

Engineering Enablement

セキュリティ視点でのコードレビュー・設定の検証
ガイドライン対応のエビデンスを収集

https://github.com/RobustIntelligence/foundation-ai-cookbook

セキュリティを変えるオープンイノベーション

サイバーセキュリティ特有のAI活用障壁を乗り越え

さまざまなセキュリティオペレーションをAI で自動化することで

人材不足を解消し、AI時代のセキュリティ組織を作りましょう！

Foundation AI™

Thank you