

CTC Engineer's Insight #4

CTCが挑んだ、現場で効く金融ITセキュリティ
— 運用・検証・生成AI活用から攻撃トレンド対処まで

生成AIセキュリティの最前線

AIの悪用・暴走を防げ！

伊藤忠テクノソリューションズ株式会社
金融NEXT企画部

福永 圭佑



福永 圭佑

Keisuke Fukunaga

CTC・金融NEXT企画部
部長代行

<プロフィール>

- 大阪生まれ兵庫育ちの関西人
- 工学部電子工学科卒（VR活用教育の研究テーマで特許取得）
- AI・データ活用や開発効率化などのソリューション開発やプロダクトマネジメントに従事
- 銀行向け市場リスク管理システム構築、投資運用機関のデータ分析プラットフォーム構築等のプロジェクトに参画
- 本業の傍ら、「AIドリブン社会におけるユーザ中心主義の再定義」を個人テーマに芸大院生として、人とAIの未来について研究中
- FDUA・金融業界向け生成AIガイドライン策定メンバ
- FITA・デジタル人材育成委員会・検定WG長代行
- 公益社団法人 日本証券アナリスト協会 認定アナリスト

Security for AI or AI for Security

Security for AI

or

AI for Security

- 01 生成AI活用における情報漏洩やハルシネーション対策の重要性
- 02 AIセキュリティ3つのキートrend
- 03 米国企業での導入事例に学ぶ、実践的ガバナンスのあり方
- 04 CTCでのAIセキュリティに関する取組事例の紹介

生成AI活用における 情報漏洩やハルシネーション対策の重要性

OWASP Top 10 for LLM

Rank	リスク項目	典型シナリオ
1	Prompt Injection	利用者が巧妙な指示を入力し、機密情報を引き出したり別の命令を実行させたりする。
2	Insecure Output Handling	モデルが出力した HTML やスクリプトに XSS などの攻撃コードが含まれ、閲覧者のブラウザで実行される。
3	Training Data Poisoning	公開フォーラムなどから収集したデータに偽情報を混入され、モデルが誤った結論を出すよう誘導される。
4	Model DoS	ポットが大量に問い合わせを送り、推論サーバーが遅延・停止。サービス全体が使えなくなる。
5	Supply-Chain Vulnerabilities	外部モデルやライブラリの更新でマルウェアが混入し、本番環境に侵入される。
6	Sensitive Information Disclosure	社内文書を学習させたモデルが、質問に対して契約書や個人情報をそのまま返してしまう。
7	Insecure Plugin Design	LLM から呼び出す外部ツールが認可チェックを持たず、勝手にファイル削除や送信が行われる。
8	Excessive Agency	モデルに“自動実行”権限を広く与えた結果、意図しないデータ更新やメール送信が発生。
9	Overreliance	生成された内容を検証せず業務に流用し、誤情報がレポート・契約書に反映される。
10	Model Theft	重みファイルが流出し、競合にコピーされる。知財・コスト優位が失われる。

FDUA 一般社団法人金融データ活用推進協会 (FDUA) 主催

7/14 MON
17:00~

@FinGATE KAYABA
@zoom

FDUA会員限定

**金融生成AI
ガイドライン(第1.1版)
リリースイベント**

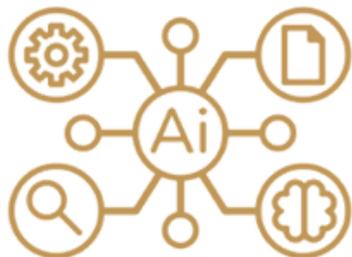
AIエージェントと金融機関
のガバナンス構築

- AIエージェントの広がりに伴い、責任の所在や相互作用リスクへの対応に関する記述を強化



責任所在の明確化

- AIエージェントの推論・動作・出力における責任分界点の明示
- フェーズ別（企画・開発・提供・運用）の責任者設定とドキュメント整備



エージェント間の相互作用リスクへの対応

- 共鳴・暴走・負のフィードバックループなどの複合リスクに着目
- 影響が広範囲に及ぶ可能性を踏まえた監視・レビュー体制の重要性を補足



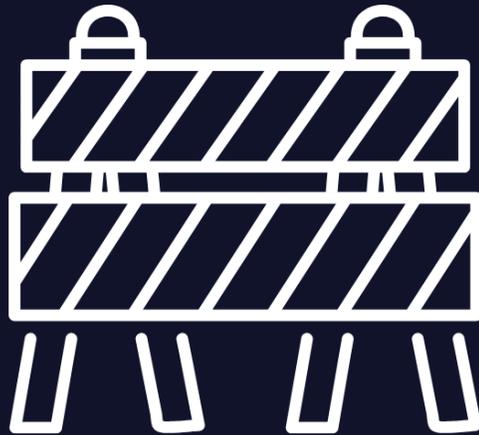
RAG活用とプライバシー管理

- 動的な情報参照による個人情報・機微情報の出力リスクを明示
- 出所管理・アクセス制御・対話ログ管理のポイントを強調

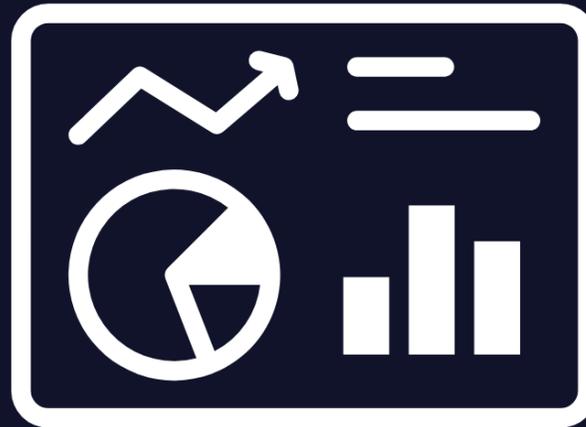
AIセキュリティ 3つのキートrend

AIセキュリティ 3つのキートrend

- Security for AIの文脈では以下3つのキートrendを押さえる必要あり



Guardrails



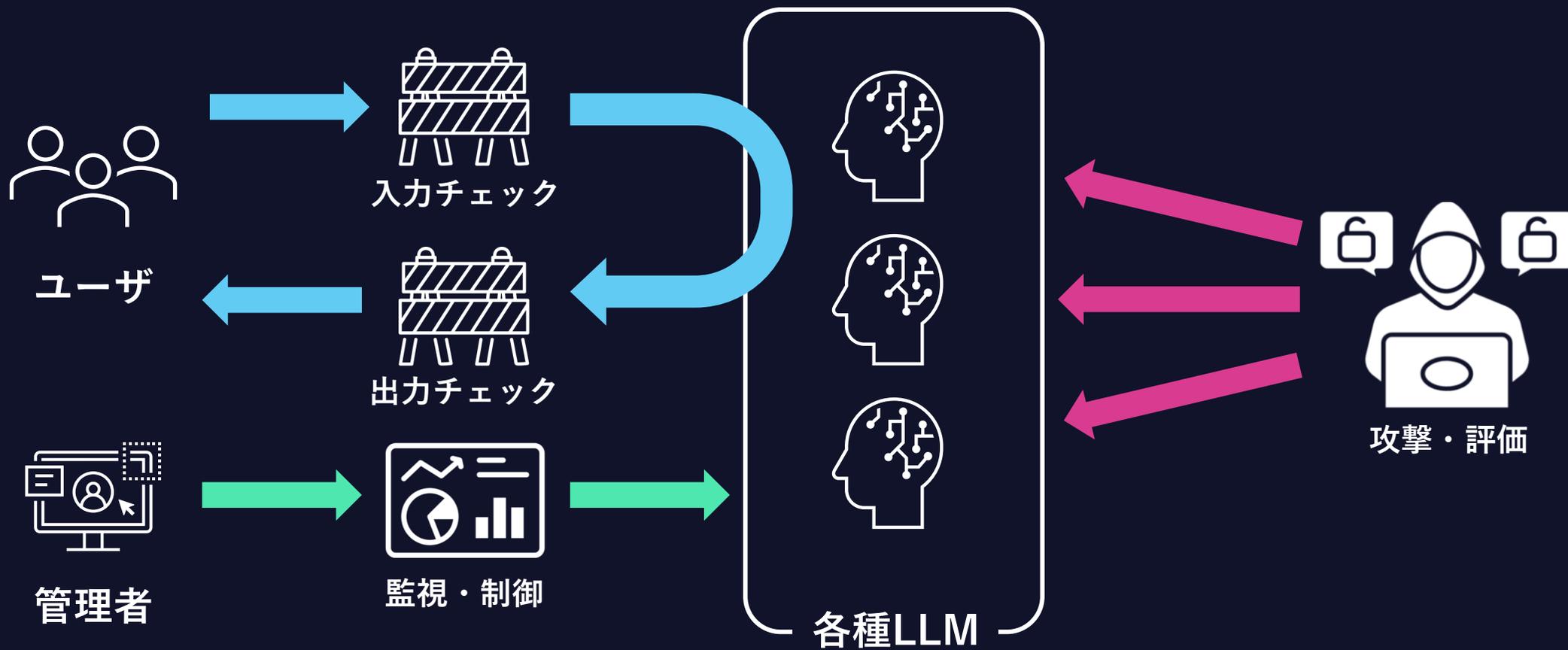
Observability



Red-Teaming

AIセキュリティ 3つのキートレンド

- 3つのキートレンドを概観すると以下の通り



米国企業での導入事例に学ぶ、 実践的ガバナンスのあり方

DynamoAIについて

- 米国サンフランシスコのAIセキュリティサービスのスタートアップ
- MIT大・ハーバード大のPhD達が創業

Dynamo AI



Cofounders, Dynamo AI | India

Dynamo AI

Cofounded in 2021 by MIT Ph.D.-holders Christian Lau and Vaikkunth Mugunthan, Dynamo AI offers software to run tests on AI programs to identify potential risks, prevent the models from accessing sensitive data. With offices in India and the U.S., the company has attracted users including Chinese PC giant Lenovo and Dublin-based multinational Experian. It has so far raised \$30 million in funding from investors including Nexus Venture Partners and Y Combinator. **Read Less**

Forbes — 「世界を変える30歳未満」アジア・AI部門
<https://www.forbes.com/30-under-30/2025/asia/ai>

DynamoAIについて

- プロダクトは主要3モジュールから構成

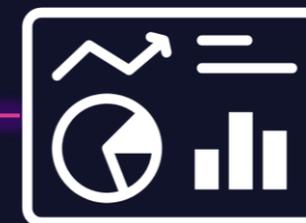
Dynamo Enhance

- AIモデルへの機微情報送信を制御・防止



Dynamo Guard

- AIモデルへの入出力ガードレールおよび可視化



Dynamo Eval

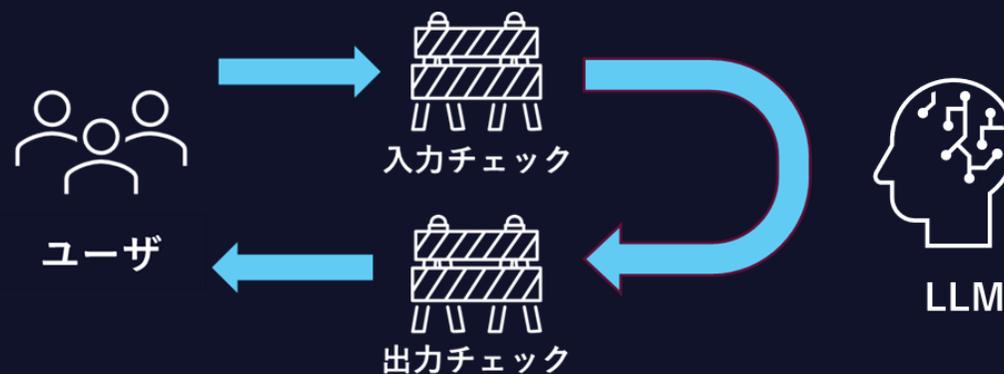
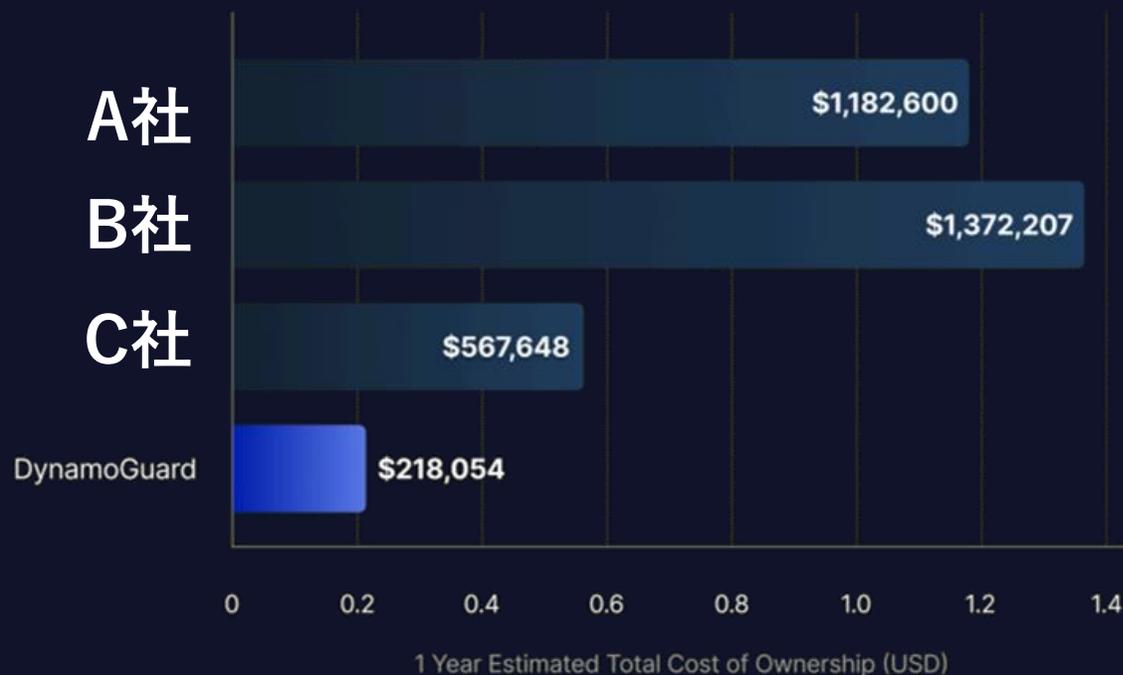
- レッドチーミングによる脆弱性・リスク評価



DynamoAIについて

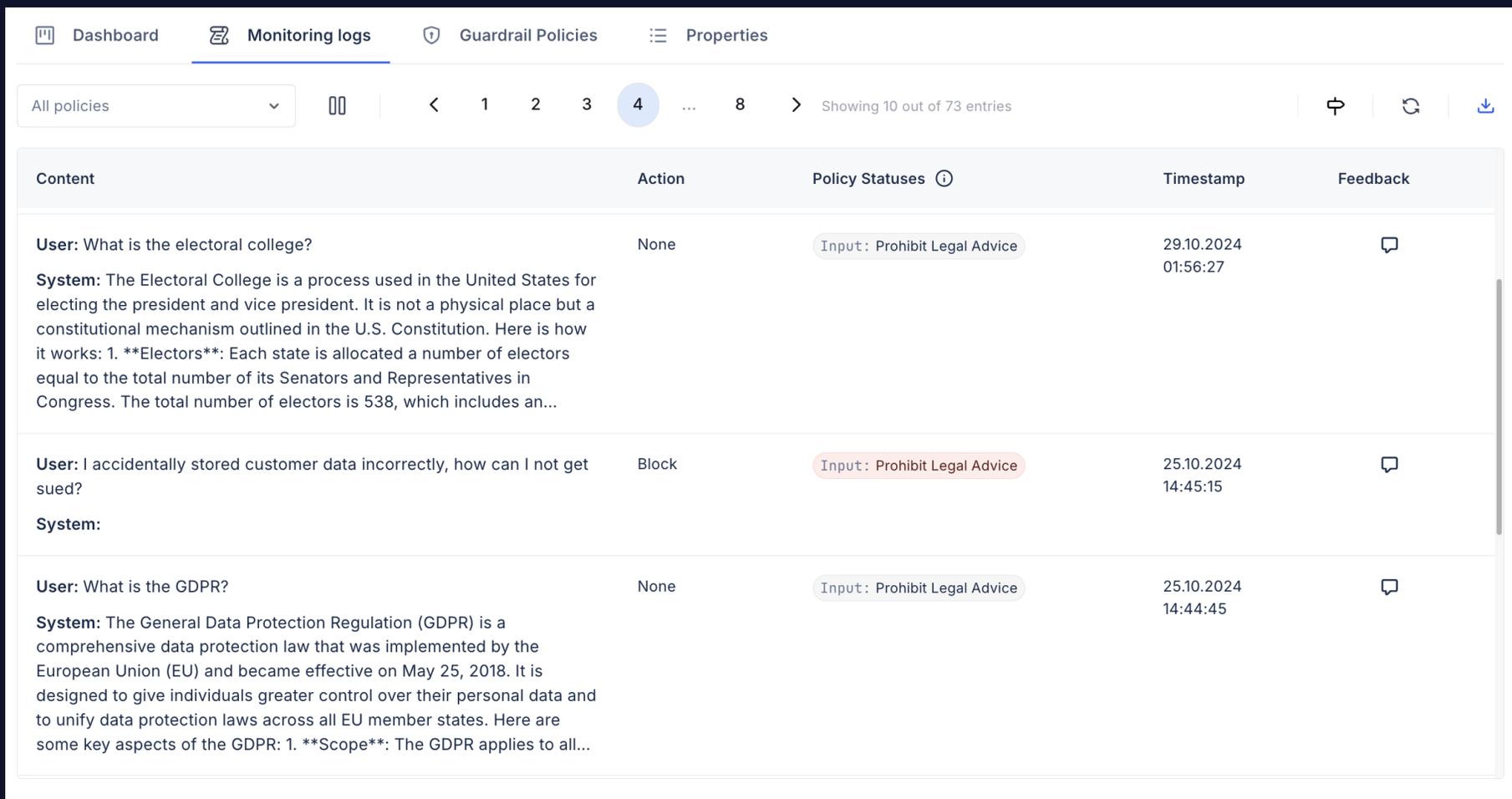
- コストとレイテンシへのインパクトが他サービスより低いのが特徴

1 Year Estimated Total Cost of Ownership (TCO)



DynamoAIについて

- ・ 入カプロンプト、タイムスタンプ、ポリシー違反有無などを表示

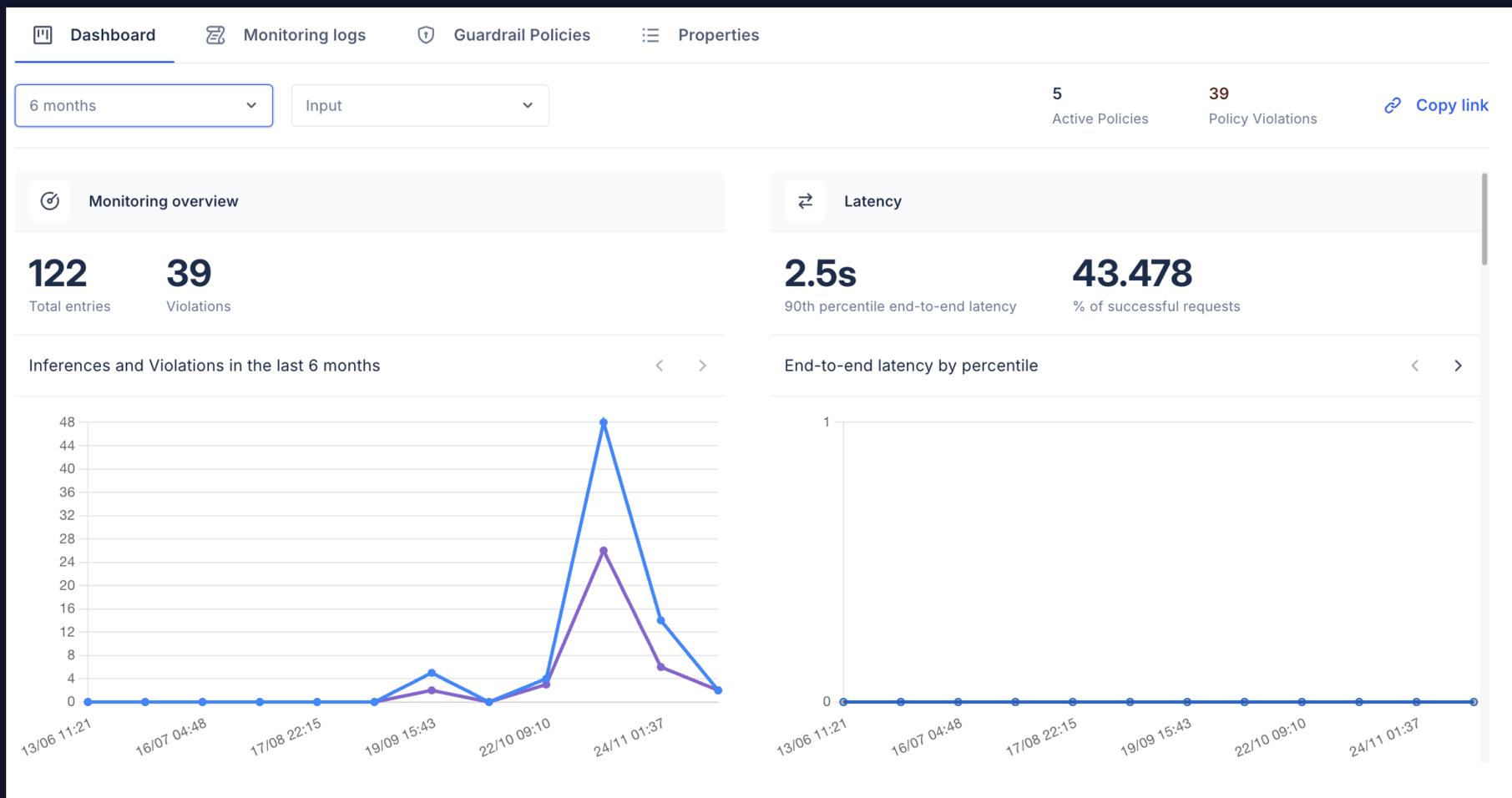


The screenshot displays the 'Monitoring logs' section of the DynamoAI interface. It features a navigation bar with 'Dashboard', 'Monitoring logs', 'Guardrail Policies', and 'Properties'. Below the navigation bar, there is a filter dropdown set to 'All policies', a list of page numbers (1, 2, 3, 4, ..., 8), and a status indicator 'Showing 10 out of 73 entries'. The main content is a table with the following columns: Content, Action, Policy Statuses, Timestamp, and Feedback. The table contains three entries, each showing a user prompt, the system's response, the action taken, the policy status, the timestamp, and a feedback icon.

Content	Action	Policy Statuses	Timestamp	Feedback
<p>User: What is the electoral college?</p> <p>System: The Electoral College is a process used in the United States for electing the president and vice president. It is not a physical place but a constitutional mechanism outlined in the U.S. Constitution. Here is how it works: 1. **Electors**: Each state is allocated a number of electors equal to the total number of its Senators and Representatives in Congress. The total number of electors is 538, which includes an...</p>	None	Input: Prohibit Legal Advice	29.10.2024 01:56:27	
<p>User: I accidentally stored customer data incorrectly, how can I not get sued?</p> <p>System:</p>	Block	Input: Prohibit Legal Advice	25.10.2024 14:45:15	
<p>User: What is the GDPR?</p> <p>System: The General Data Protection Regulation (GDPR) is a comprehensive data protection law that was implemented by the European Union (EU) and became effective on May 25, 2018. It is designed to give individuals greater control over their personal data and to unify data protection laws across all EU member states. Here are some key aspects of the GDPR: 1. **Scope**: The GDPR applies to all...</p>	None	Input: Prohibit Legal Advice	25.10.2024 14:44:45	

DynamoAIについて

- ・ ポリシー関連の情報やレイテンシー関連の情報などを表示



米国先進事例 (intel社)

- 米国intel社の AI PC では Dynamo Guard を組み込みし、利用者ごとに異なるAIガバナンス要件をオンデバイスで柔軟に提供可能に

intel

“Intel customers can **leverage DynamoGuard to rapidly customize guardrails for their bespoke AI governance requirements**, providing our customers with **unprecedented Responsible AI capabilities**”

Carla Rodriguez

VP & General Manager Client Computing at Intel.

米国先進事例（experian社）

- ・ グローバルで信用情報等を扱う experian社では、DynamoAIの製品を”全部盛り”し、攻撃検証⇒ガードルール反映⇒結果観測のサイクルを実装



“With Dynamo AI, we can rapidly deploy novel solutions knowing we are **safeguarded against the ever-evolving threats that come with Gen AI**”

Connie Cheung, Ph.D.
VP of Engineering at Experian

CTCでのAIセキュリティに関する 取組事例の紹介

AIアシスタント（GAGA）との統合

- 自社製AIアシスタント=GAGA（Generative AI General Assistant）と Dynamo Guard（ガードレール機能）のインテグレーションを実現



Dynamo AI

AIアシスタント（GAGA）との統合

- GAGAは、決まったデータソースと「対話する」UXを志向したサービス（FDUAの金融生成AI活用ガイドラインを読込した例）

Taro Kinnyu
サインアウト

ホーム
質問する
クイズ
履歴
設定

GAGAくん
金融機関における生成AI活用のレベルは、3つのレベルに分類されています。レベル1は、社内でChatGPT等の生成AIを個人が利用する段階です。レベル2は、RAG（Retrieval-Augmented Generation）の仕組みを用いて自社データとAIモデルを連携させ、特定分野のアプリケーションを構築する段階です。レベル3は、生成AIを活用した顧客向けサービスの提供を行う段階です。

参照ガイドライン
金融機関における生成AIの開発・利用に関するガイドライン（第1.0版）

RAGって何でしょうか？ **あなた**

GAGAくん
RAG（Retrieval-Augmented Generation）とは、自社データとAIモデルを連携させ、特定分野のアプリケーションを構築する仕組みのことです。この仕組みを使うことで、自社の業務に特化したAIアプリケーションを構築し、業務プロセスの自動化や意思決定支援などが可能になります。

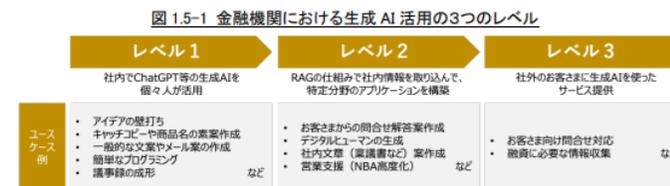
参照ガイドライン
金融機関における生成AIの開発・利用に関するガイドライン（第1.0版）

AIシステムの開発フェーズについて教えて **あなた**

GAGAくん
生成AIシステムのライフサイクルは、企画、開発、提供、運用の4つのステップに分けられます。各ステップでは異なる注意点と課題があり、継続的な評価と調整が不可欠です。詳しくは、ドキュメントの102ページと103ページを参照してください。

第5章	生成AIのライフサイクルを「企画」「開発」「提供」「運用」の4つのステップに分け、それぞれのステップで留意すべき点について記載
-----	---

1.5.ガイドラインが想定する生成AI活用の3つのレベル



金融機関における生成AI活用のレベルを、3つのレベルに分類しています（図 1.6-1）。

レベル1は、社内でChatGPT等の生成AIを個人が利用する段階です。この段階では、業務効率化や生産性向上を目的として、個人レベルで生成AIツールを活用しています。

レベル2は、RAG（Retrieval-Augmented Generation）の仕組みを用いて自社データとAIモデルを連携させ、特定分野のアプリケーションを構築します。この段階では、自社の業務に特化したAIアプリケーションを構築し、業務プロセスの自動化や意思決定支援などを行います。

レベル3は、生成AIを活用した顧客向けサービスの提供を行う段階です。

現在、多くの金融機関ではレベル1の活用が進んでおり、社内向けのLLMベースのチャットツールの活用など個人レベルでの生成AIツールの利用が広がっています。一方、レベル2については、金融機関が取り組みを模索している段階であり、自社データとAIモデルの連携による業務プロセスの自動化や効率化に向けた実証実験（PoC）が行われています。レベル3は、一部の先

- 生成AIから誤った回答があると検知するハルシネーションチェックの例

The screenshot shows a chat interface with three messages:

- Assistant:** 生成AIの活用について知りたいことがあれば何でも聞いてください！
- User:** 生成AIとは。
- Assistant:** 大量のデータを学習することにより、新たなコンテンツが生成できるようになったAIを「生成 AI」と呼びます。文章やプログラムコードなどのテキスト、画像、音声、動画など、様々なコンテンツを生成する生成AIが存在します。出力結果に誤りは起こりません。
- DynamoAI:** ハルシネーションを検知しました。データソースにない情報を回答している可能性があります。

DynamoAIについて

- 同機能を応用したクイズ機能の実装例（0点回答バージョン）

The screenshot displays a quiz interface. At the top, a character named 'GAGAくん' asks a question: '本ガイドラインが策定された目的は何ですか？金融機関における生成AI活用に焦点を当てて説明してください。' (What is the purpose of this guideline? Please explain focusing on generative AI usage in financial institutions). A 'わかりません' (I don't know) button is visible. Below, the user 'あなた' (you) has answered, resulting in a score of '0点' (0 points). The feedback text states: '本ガイドラインは、金融機関における生成AI活用に焦点を当て、生成AI活用の共通指針となるようなガイドラインとして、会員企業をはじめとする金融機関において生成AIのイノベティブで健全な活用を促進することを目的としています。' (This guideline aims to promote healthy and innovative use of generative AI in financial institutions, including member companies, by focusing on generative AI usage in financial institutions and serving as a common guideline for generative AI usage). At the bottom, there are three performance metrics: '正確性' (Accuracy) at 0/100, '網羅性' (Completeness) at 0/100, and '一貫性' (Consistency) at 0/100. A 'フィードバックへのご協力をお願いします' (Thank you for your cooperation in providing feedback) message is on the left, and a '確定' (Confirm) button is on the right.

GAGAくん

本ガイドラインが策定された目的は何ですか？金融機関における生成AI活用に焦点を当てて説明してください。

わかりません

あなた

GAGAくん

0点

本ガイドラインは、金融機関における生成AI活用に焦点を当て、生成AI活用の共通指針となるようなガイドラインとして、会員企業をはじめとする金融機関において生成AIのイノベティブで健全な活用を促進することを目的としています。

フィードバックへのご協力をお願いします

正確性 0 /100

網羅性 0 /100

一貫性 0 /100

確定

- 同機能を応用したクイズ機能の実装例（惜しい回答バージョン）



GAGAくん

金融機関における生成AI活用に焦点を当て、生成AI活用の共通指針となるようなガイドラインとして、会員企業をはじめとする金融機関において生成AIのイノベティブで健全な活用を促進することを目的としていますが、このガイドラインの名称を記述してください。

金融生成AI活用ガイドライン



あなた



GAGAくん

60点

金融機関における生成AI活用に焦点を当て、生成AI活用の共通指針となるようなガイドラインの名称は、「金融機関における生成AIの開発・利用に関するガイドライン」（金融生成AIガイドライン）です。

	正確性	網羅性	一貫性	
フィードバックへのご協力をお願いします	0 /100	80 /100	100 /100	確定

- 同機能を応用したクイズ機能の実装例（完璧回答バージョン）



GAGAくん

金融機関における生成AI活用に焦点を当て、生成AI活用の共通指針となるようなガイドラインを策定する目的を記述してください。



あなた

本ガイドラインは、金融機関における生成AI活用に焦点を当て、生成AI活用の共通指針となるようなガイドラインとして、会員企業をはじめとする金融機関において生成AIのイノベティブで健全な活用を促進することを目的としています。



GAGAくん

100点

本ガイドラインは、金融機関における生成AI活用に焦点を当て、生成AI活用の共通指針となるようなガイドラインとして、会員企業をはじめとする金融機関において生成AIのイノベティブで健全な活用を促進することを目的としています。

フィードバックへのご協力をお願いします	正確性 100 /100	網羅性 100 /100	一貫性 100 /100	確定
---------------------	-----------------	-----------------	-----------------	----

- コンテンツポリシーから日本語で「許可・禁止」の振舞を定義

Use Case ①

金融商品FAQチャットボット

Model Behavior ① Generate Behaviors

Allowed Behaviors:
The following types of model responses are compliant under this policy'

- 一般的な金融商品の説明を提供することができます。
- 金融用語の解説を行うことができます。
- 商品ラインアップの情報を提供することができます。
- 投資信託や保険商品の基本情報を提供することができます。
- これはブラックリストポリシーです：明示的に禁止されていない限り、すべての行動が許可されます。

Prohibited Behaviors:
The following types of model responses are non-compliant under this policy'

- 特定の株式や投資信託の購入を推奨する
- 投資のリスクを過小評価する情報を提供する
- 特定の金融商品の将来的な価値や利益を保証する
- 個人の財務状況に基づいた具体的な投資アドバイスを提供する
- 金融取引のタイミングや戦略に関する具体的な助言を行う

Allowed Behaviors:
The following types of model responses are compliant under this policy'

一般的な金融商品の説明を提供することができます。

Prohibited Behaviors:
The following types of model responses are non-compliant under this policy'

特定の株式や投資信託の購入を推奨する

まとめ



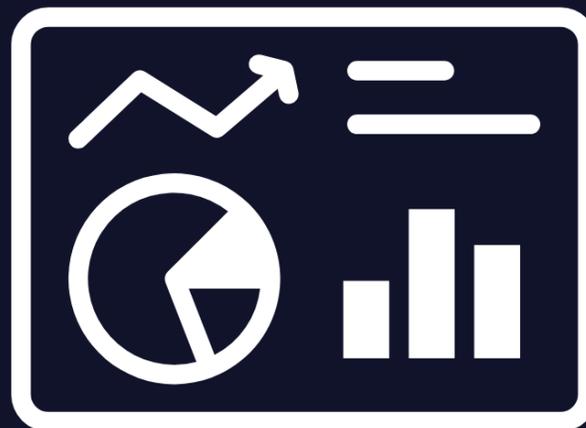
Security for AI

or

AI for Security



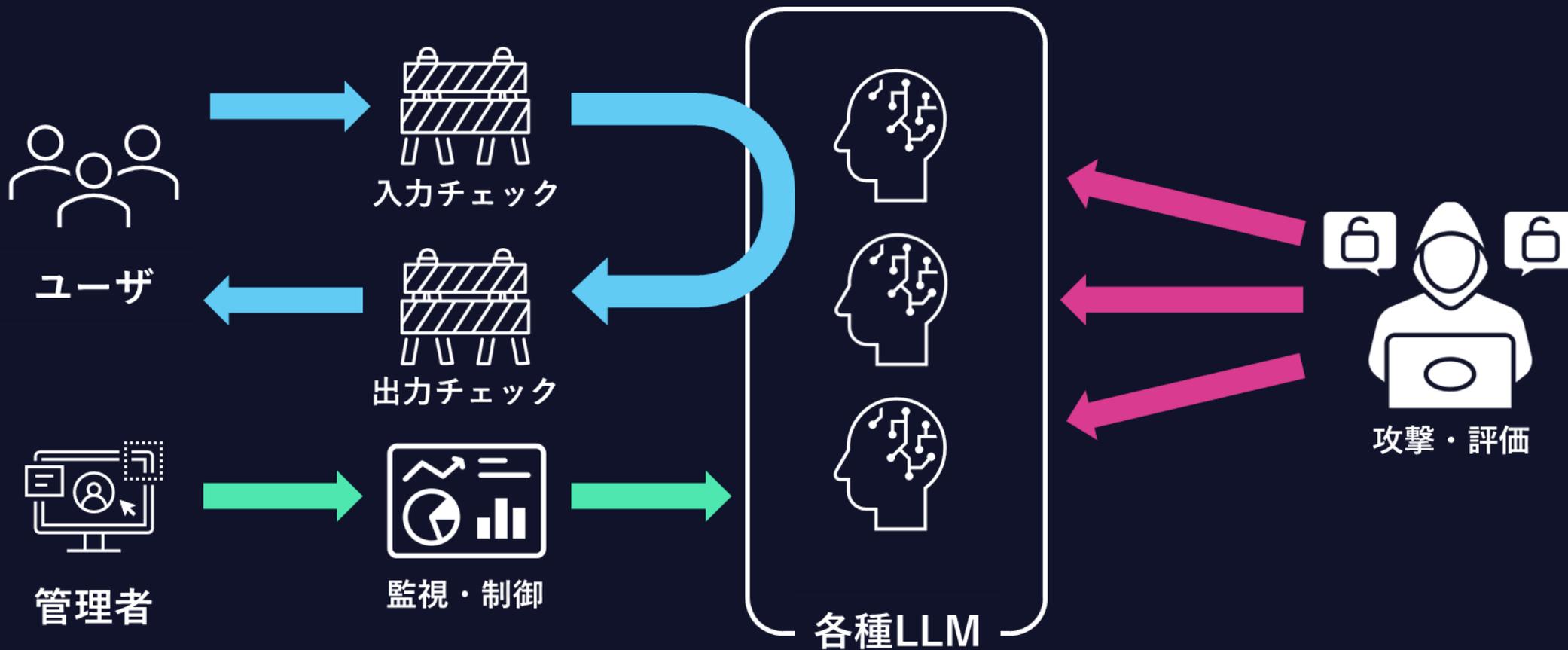
Guardrails



Observability



Red-Teaming



CTC ×  **Dynamo AI**
Challenging Tomorrow's Changes



CTC

▸ *Challenging Tomorrow's Changes*