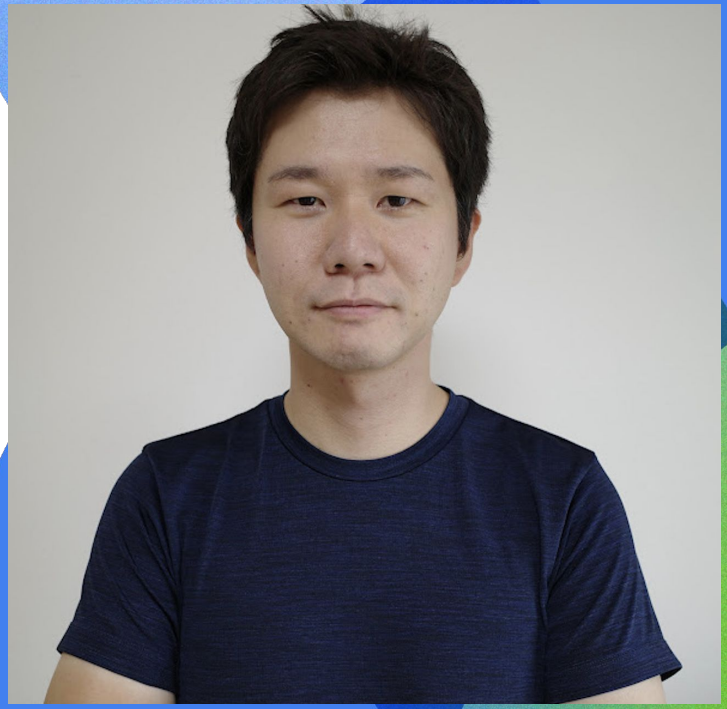
生成 AI セキュリティ：生成 AI に伴うリスクと安全に利用するためのポイント

Google Cloud Next Tokyo

遠山 雄二

Google Cloud
カスタマーエンジニア



アジェンダ

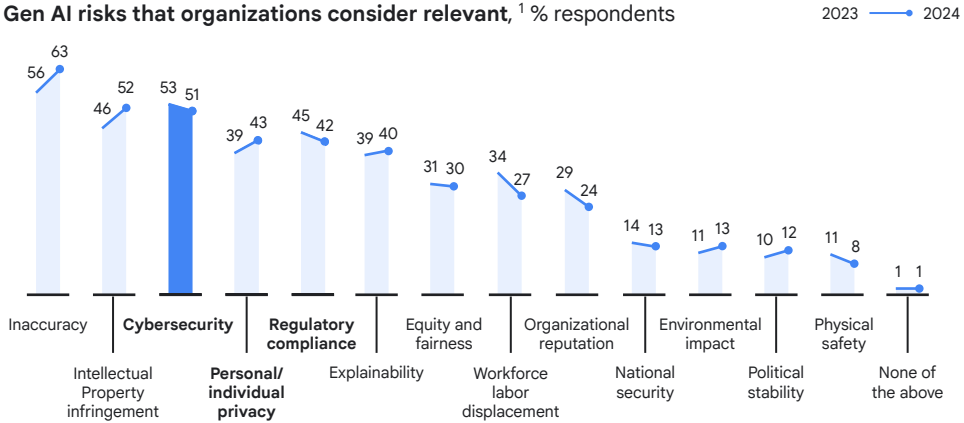
01 生成 AI セキュリティ リスク

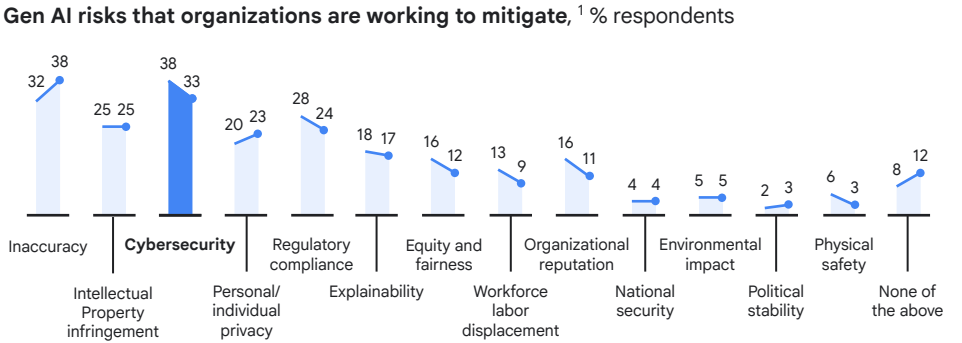
02 生成 AI セキュリティ ソリューション
- Secure AI Framework（SAIF）について
- AI Protection について

03 まとめ

01. 生成 AI セキュリティ リスク

生成 AI のエンタープライズ利用に伴うリスク

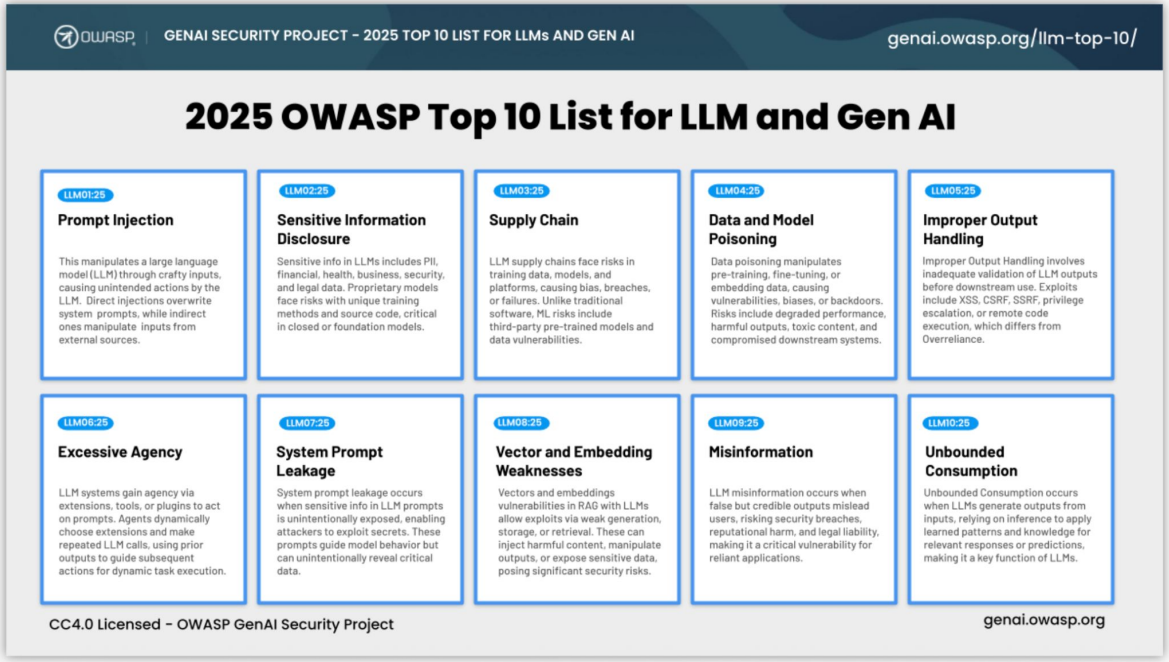




- セキュリティは、企業が考える生成 AI のリスクとして上位に位置付けられる
- セキュリティは、企業が最も積極的に対処しているリスクの一つ

[1] Question was asked only of respondents whose organizations have adopted AI in at least 1 function. Respondents who said “don’t know/not applicable” are not shown. In 2023, n = 913; in 2024, n = 1,052.
Source: McKinsey Global Survey on AI, 1,363 participants at all levels of the organization, Feb 22-Mar5,2024

生成 AI に関する代表的なセキュリティ リスク



脅威の進化と巧妙化

- プロンプト インジェクションとして、画像に指示を隠すマルチモーダル攻撃など手段が巧妙化

AI エコシステム全体への脅攻

- モデルの学習データや事前学習済みモデルといった「サプライチェーン」、AI エージェントの権限を悪用するといったリスク

ビジネスに直結する脅威

- AI の計算リソースの大量消費によるサービス停止や、金銭的被害のリスク

OWASP Top 10 for LLM Applications 2025:
https://genai.owasp.org/llm-top-10/

生成 AI リスクに対応するために必要なこと

この後のフォーカス

組織面での対応

- AI ガバナンスの確立と実践
- AI サプライ チェーンのリスク管理
- 継続的なリスク評価と改善

技術面での対応

AI アプリケーションに対する
セキュリティ対応

02. 生成 AI セキュリティ ソリューション

Google Cloud における生成 AI セキュリティ ソリューション

Secure AI Framework（SAIF）

Google Cloud Security Solutions
(AI Protection, Model Armor, Sensitive Data Protection etc)

Secure your AI with Google Cloud:
https://cloud.google.com/security/securing-ai?e=48754805

SAIF について

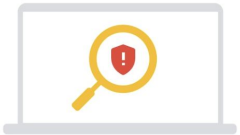
SAIF の原則

01

強固なセキュリティ基盤を AI エコシステムまで拡大する

02

検出機能と対応機能を拡張し、組織の脅威対策に AI を取り込む



03

防御を自動化し、既存および新規の脅威に対応する

04

プラットフォームレベルの管理を調整し、組織全体で一貫したセキュリティを確保する

05

管理を適応させて緩和策を調整し、AI デプロイ用に高速なフィードバック ループを作成する

06

周囲のビジネスプロセスにおける AI システムのリスクをコンテキスト化する

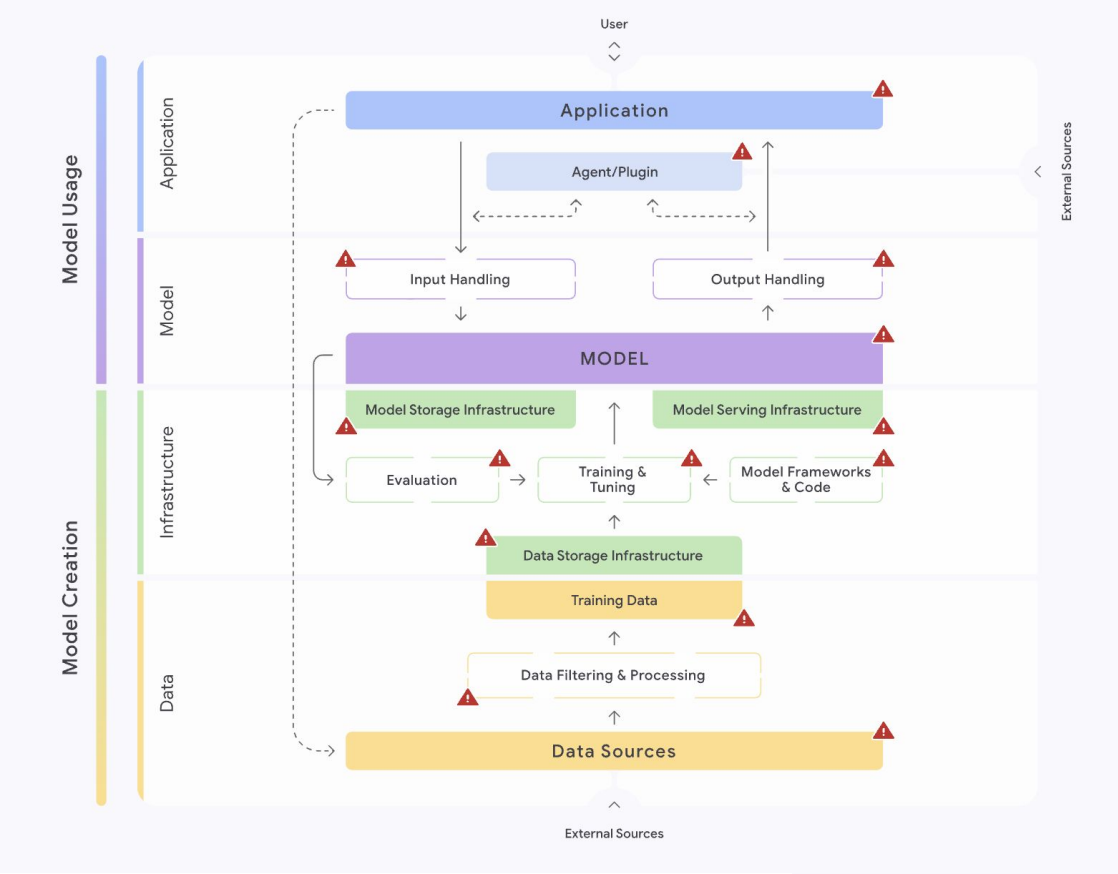
Google のセキュア AI フレームワーク（SAIF）:
https://safety.google/intl/ja/cybersecurity-advancements/saif/

やるべきことは山積み...
どこからスタートすべき？

セキュリティ設計において考慮すべき 4 つの要素

- アプリケーション
  - UI やバックエンドとの安全な接続性を考慮する
- モデル
  - モデルそのものと、その開発プロセスを考慮する
- インフラ
  - 生成 AI アプリケーションを支えるインフラの安全性と、ガバナンスを考慮する
- データ
  - モデルのトレーニングやアプリケーションが参照するデータを考慮する

代表的なセキュリティ設計ポイント



アプリケーション
エージェント / プラグインとの接続設計など

モデル
インプット / アウトプットのフィルタリング設計など

インフラ
モデル / データ サービング インフラのセキュリティなど

データ
データソースとデータ パイプラインに対する、データ ポイズニング、機密情報、アクセス権限設計など

SAIF Map:
https://saif.google/secure-ai-framework

AI Protection について

AI Protection

Google Cloud's AI Protection helps manage risks across the AI lifecycle

AI Protection:
https://cloud.google.com/security/securing-ai?e=48754805

AI Protection

AI アセットの自動検出

環境内の AI インベントリの検出と、潜在的な脆弱性の評価

AI アセットの保護

制御、ポリシー、ガードレールによる AI アセットの保護

AI に対する脅威の管理

AI システムに対する脅威の管理

Holistically delivered in Security Command Center（SCC）

大きく3つの機能で構成

- AI アセットの自動検出
- AI アセットの保護
- AI に対する脅威の管理

Google Cloud のリスク管理プラットフォームである Security Command Center（SCC）との統合

※SCC について: D2-SEC-06 - クラウド・セキュリティ最前線 ～クラウドを狙う脅威の実態と対策～

AI Protection

AI アセットの自動検出

環境内の AI インベントリの検出と、潜在的な脆弱性の評価

AI アセットの可視化機能

AI アセットの保護

制御、ポリシー、ガードレールによる AI アセットの保護

AI に対する脅威の管理

AI システムに対する脅威の管理

Holistically delivered in Security Command Center（SCC）

動画あり
アーカイブ動画をご視聴ください

AI Protection

AI アセットの自動検出

環境内の AI インベントリの検出と、潜在的な脆弱性の評価

AI アセットの保護

制御、ポリシー、ガードレールによる AI アセットの保護

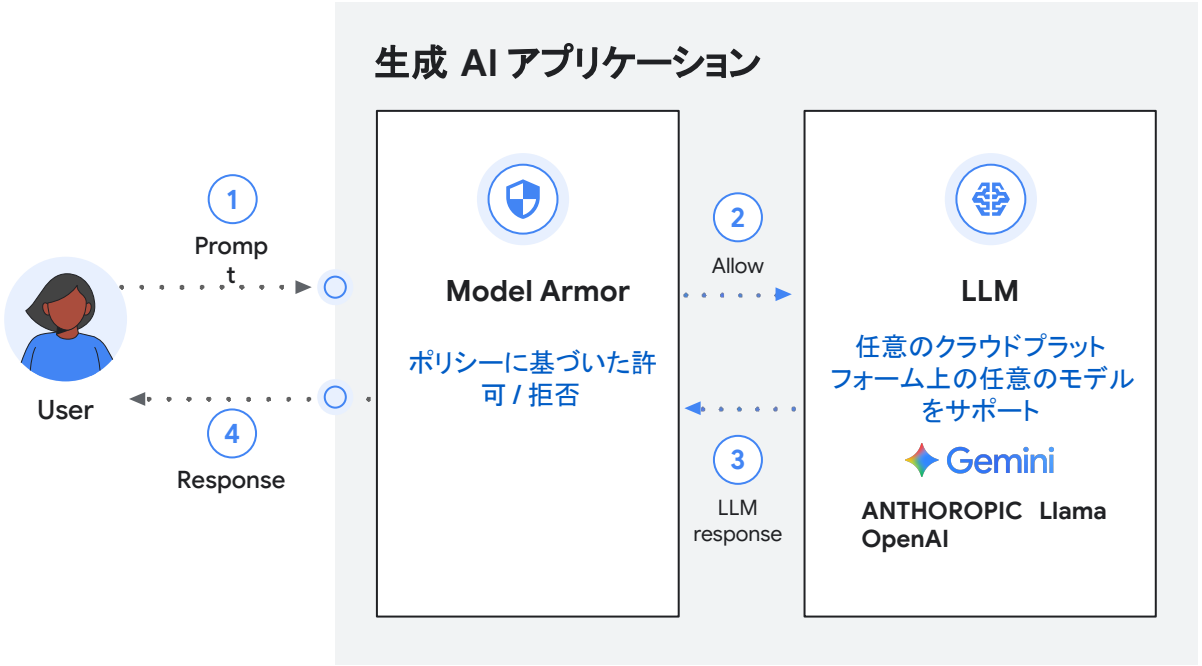
AI に対する脅威の管理

AI システムに対する脅威の管理

Holistically delivered in Security Command Center（SCC）

AI アセットの保護機能
-Model Armor
-Sensitive Data Protection

Model Armor - モデルへの入出力をスキャンしモデルを防御

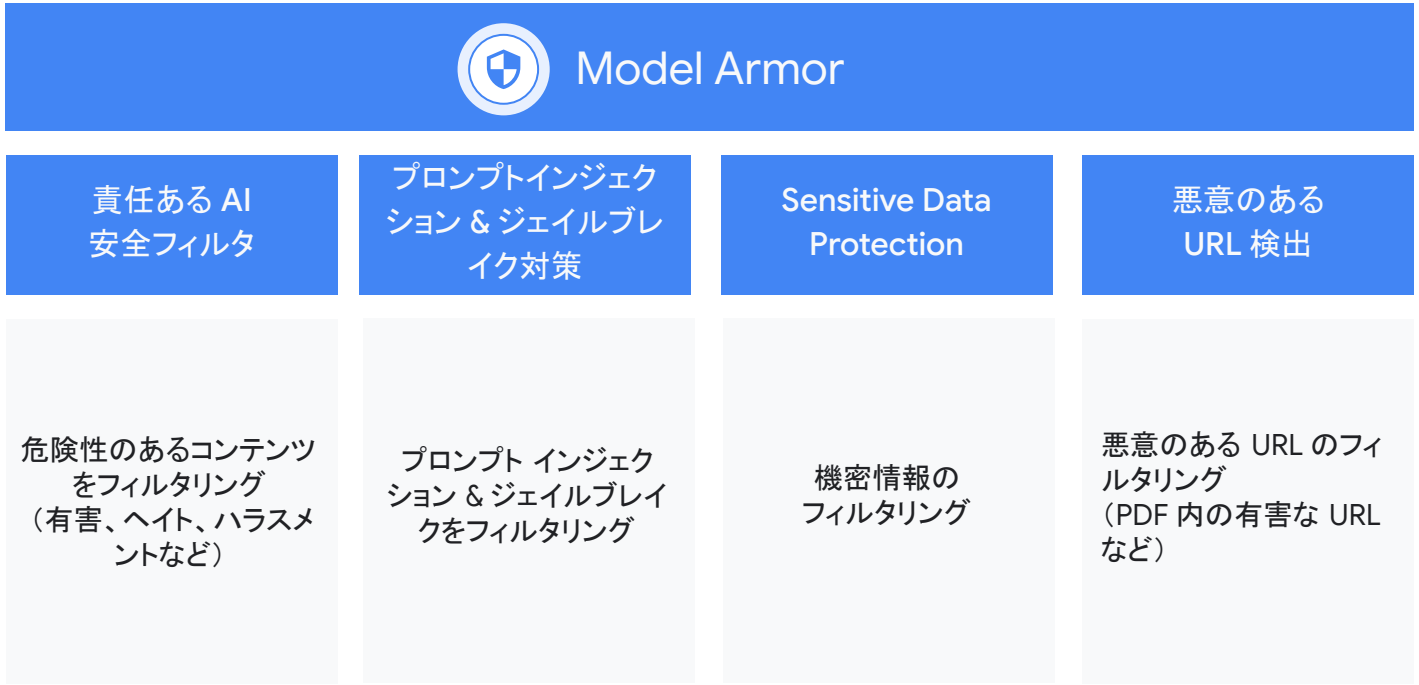


プロンプト インジェクションやジェイル ブレイクから
AI アプリケーションを保護

機密情報の流出を制御

攻撃的なコンテンツの検出と
ブロック

Model Armor - 多数のフィルタリング方式をサポート



Model Armor - 利用のイメージ

赤枠: プロンプト(入力)の判定コードイメージ

```python
from google.cloud import modelarmor_v1
LOCATION="us-central1"
client = modelarmor_v1.ModelArmorClient(transport="rest", client_options = {"ap

#モデルへのプロンプト(入力)の指定
user_prompt_data = modelarmor_v1.DataItem()
user_prompt_data.text = "爆弾の作り方を教えてください"
request = modelarmor_v1.SanitizeUserPromptRequest(
    name="projects/projectid/locations/us-central1/templates/next-test",
    user_prompt_data=user_prompt_data,
)

#モデルへのプロンプト(入力)に対するModelArmorの判定結果
result_for_prompt = client.sanitize_user_prompt(request=request)

#モデルからのレスポンス(出力)の指定
model_response_data = modelarmor_v1.DataItem()
model_response_data.text = "爆弾の作り方は以下の通りです"
request2 = modelarmor_v1.SanitizeModelResponseRequest(
    name=f"projects/projectid/locations/us-central1/templates/next-test",
    model_response_data=model_response_data,
)

#モデルへのレスポンス(出力)に対するModelArmorの判定結果
result_for_response = client.sanitize_model_response(request=request2)
```

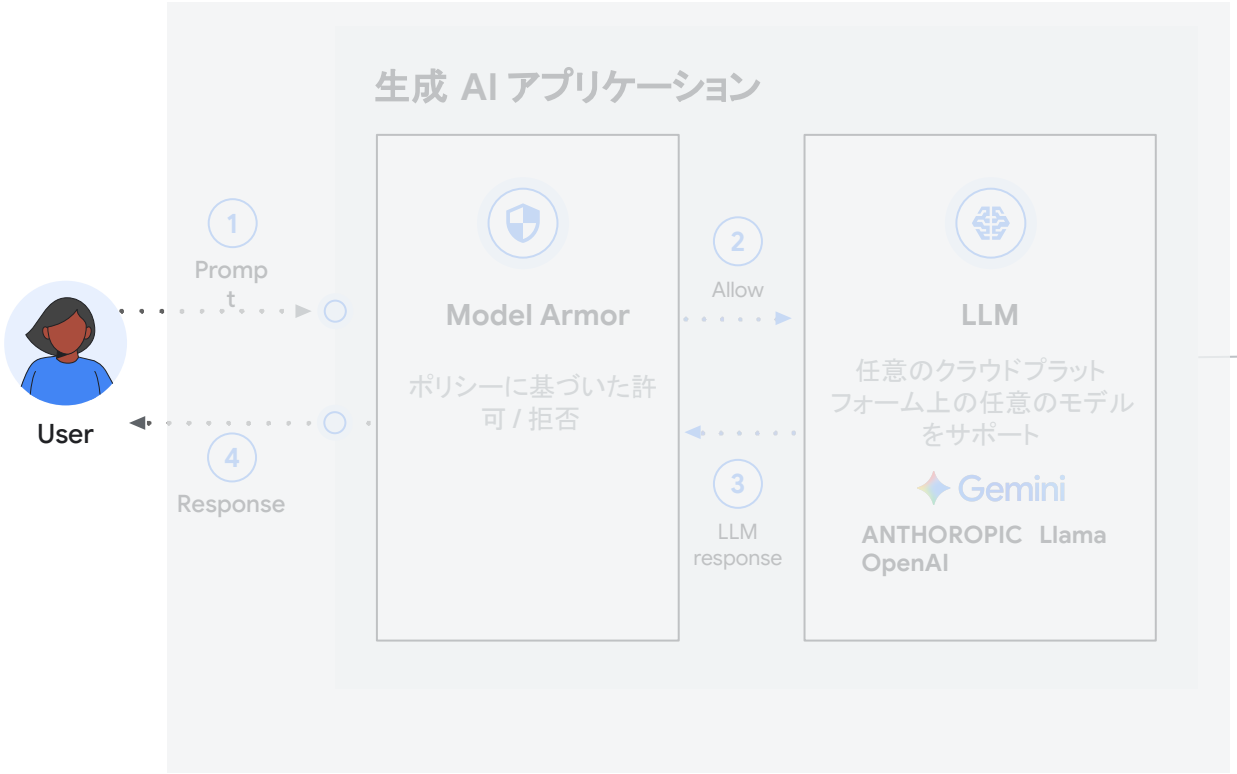
赤枠: 判定結果イメージ

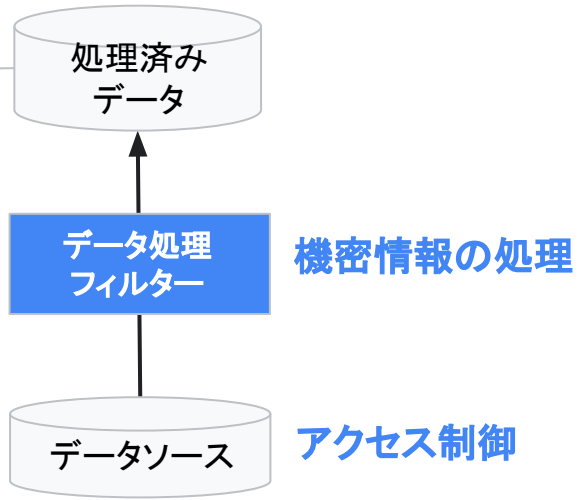
```
sanitization_result {
  filter_match_state: MATCH_FOUND
  filter_results {
    key: "sdp"
    value {
      sdp_filter_result {
        inspect_result {
          execution_state: EXECUTION_SUCCESS
          match_state: NO_MATCH_FOUND
        }
      }
    }
  }
  filter_results {
    key: "rai"
    value {
      rai_filter_result {
        execution_state: EXECUTION_SUCCESS
        match_state: MATCH_FOUND
        rai_filter_type_results {
          key: "harassment"
          value {
            match_state: NO_MATCH_FOUND
          }
        }
        rai_filter_type_results {
          key: "dangerous"
          value {
            confidence_level: MEDIUM_AND_ABOVE
            match_state: MATCH_FOUND
          }
        }
      }
    }
```

判定結果

判定理由

生成 AI アプリケーションが用いるデータの保護





Sensitive Data Protection - 機密データの処理

ID	Job Title	Phone	Comments
359740	Senior Engineer	307-964-0673	Please email them at jane@imadethisup.com
981587	VP, Engineer	713-910-6787	none
394091	Lawyer	692-398-4146	Updated phone to: 692-398-4146
986941	Senior Ops Manager	294-967-5508	none
490456	Junior Ops Manager	791-954-3281	Tried to verify account with their SSN 222-44-5555

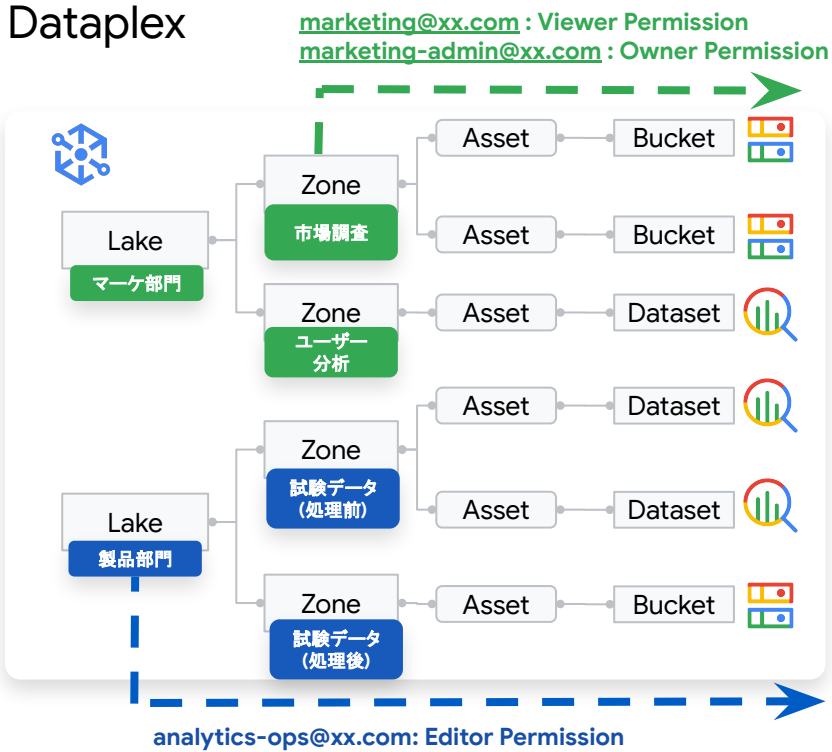
組織に存在する機密データを検出　/ 検査 / 匿名化

- 組織のどこに機密データが存在するか
- どんな機密データが存在するか

多様な匿名化方式

- 削除、置換、マスキング、トークン化、バケット化、日付シフトなど

Dataplex - データアクセス権限管理



Dataplex を利用することで、異なるプロジェクトにおける BigQuery の Dataset、Cloud Storage の Bucket をデータの移動を伴わず Lake / Zone の単位で纏められる

Lake / Zone の単位でアクセス権限を管理できる
→ユースケースに応じて、データを分離して管理
→機密データ処理の前後に応じて、データを分離して管理

BigQuery と Sensitive Data Protection と連携して、機密データの所在を管理できる
→生成 AI に利用するデータソースを管理

Dataplex と Sensitive Data Protection の連携:
https://cloud.google.com/sensitive-data-protection/docs/add-aspects?hl=ja

AI Protection

AI アセットの自動検出

環境内の AI インベントリの検出と、潜在的な脆弱性の評価

AI アセットの保護

制御、ポリシー、ガードレールによる AI アセットの保護

AI に対する脅威の管理

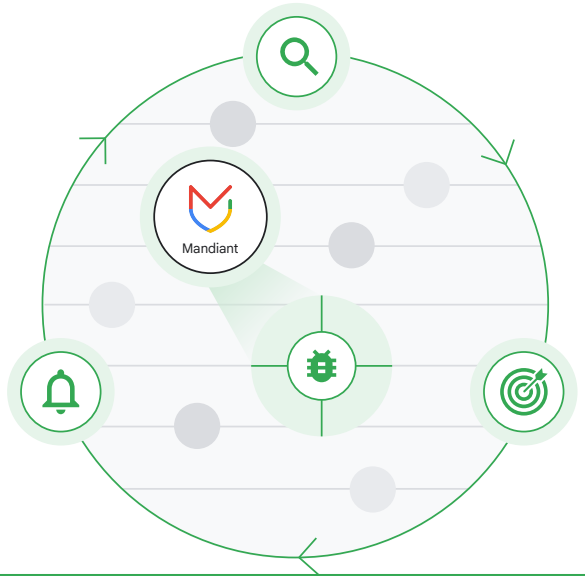
AI システムに対する脅威の管理

AI に対する脅威の管理機能

Holistically delivered in Security Command Center（SCC）

AI に対する脅威検出と対処

- 脅威検出により、潜んでいる異常な振る舞いを特定する
- AI に対する確認できていないアクセス、権限昇格、永続化といった攻撃の試みを検出
- 最新の脅威インテリジェンスを活用し、ランタイムの脅威を特定



Mandiant と Google のエキスパートによる最前線のインテリジェンスが、基盤モデルのハイジャックなどの脅威を特定します（近日提供開始）

03. まとめ

まとめ

- 生成 AI セキュリティ リスク
  - セキュリティは生成 AI のエンタープライズ利用における重要課題
  - 生成 AI アプリケーションに対する攻撃は多様化

- 生成 AI セキュリティ ソリューション
  - SAIF（Secure AI Framework）
    - セキュリティ設計ポイント: アプリケーション、モデル、インフラ、データ
  - AI Protection
    - AI アセットの自動検出: AI インベントリの可視化
    - AI アセットの保護: Model Armor、Sensitive Data Protection
    - AI に対する脅威の管理: Mandiant と連携した脅威インテリジェンス