資料４-４

AWSの生成AIサービスとセキュリティ対策

（総務省）AIセキュリティ分科会第４回会合

アマゾン ウェブ サービス ジャパン 合同会社

執行役員 パブリックセクター技術統括本部長
瀧澤 与一 / Yoichi Takizawa

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Agenda

1. AWSの生成AIサービス

2. AI提供段階における、AIシステムの機密性、完全性、可用性を損なう脅威とAWSが提供している対策

3. まとめ

Agenda

1. AWSの生成AIサービス

2. AI提供段階における、AIシステムの機密性、完全性、可用性を損なう脅威とAWSが提供している対策

3. まとめ

AIによるイノベーションに必要なすべてを提供

すぐに使えるソリューション

満足度の向上
Amazon Connect

ソフトウェア開発
Kiro

マイグレーション モダナイゼーション
AWS Transform

ビジネス生産性
Amazon Q

AWS MARKETPLACE
パートナーソリューション | エージェント

ビルダーのためのツールとインフラ

AIエージェントとアプリの構築
Amazon Bedrock | Amazon Nova
Strands Agents | MCP

データ準備・モデルトレーニング
Amazon SageMaker
Amazon S3 | データベース等

AI計算資源
Trainium | Inferentia
NVIDIA GPU

高度な生成AI/AIMLに関する専門知識

GEN AI INNOVATION CENTER
概念検証 | エージェントの導入

PARTNER NETWORK
140,000以上の パートナー | 生成AIコンピテンシー

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock

基盤モデルを活用した
生成 AI アプリケーションを
簡単に構築、拡張できる方法

API を介して基盤モデルを利用することで
生成 AI アプリ開発を加速
インフラ管理は不要

お客様の業務用途に適した基盤モデルを選択
Amazon, AI21 Labs, Anthropic, Cohere, Meta,
Mistral AI, Stability AI

データセキュリティ、プライバシー、安全性
に配慮した構築

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon Bedrock

AIをリードする企業の多様なモデルをフルマネージで幅広く利用可能

AI21labs

Highly efficient processing & grounded generation for long context lengths

JAMBA

MISTRAL AI_

Specialized expert models for agentic reasoning and multimodal tasks

MISTRAL

MIXTRAL

PIXTRAL

amazon

Frontier intelligence and industry leading price performance

NOVA

OpenAI

Automate tasks, enhance creativity, and solve complex problems efficiently

GPT-OSS

ANTHROPIC

Excels at complex reasoning, code generation, and instruction following

CLAUDE

poolside

Software engineering AI for large enterprises

Coming soon

cohere

Powering efficient, multilingual AI agents with advanced search & retrieval

COMMAND

EMBED

RERANK

Qwen

Advanced reasoning with agentic intelligence

QWEN

deepseek

Advanced reasoning models that solve complex problems step-by-step

DEEPSEEK

stability.ai

Professional-grade images with creative control, deployable at scale

STABLE DIFFUSION

STABLE IMAGE

Luma

High-quality video generation with natural, coherent motion & ultra-realistic details

RAY

TwelveLabs

CTRL + F for video data: unlock the full potential of enterprise video assets

MARENGO

PEGASUS

Meta

Advanced image and language reasoning

LLAMA

WRITER

Purpose-built models for building and scaling AI agents across the enterprise

PALMYRA

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrcok
国内にデータを閉じて推論が可能な基盤モデル

- Amazon Nova Lite, Sonic, Canvas, Reel
- Amazon Titan Text G1 – Express, Titan Embeddings G1 – Text, Titan Text Embeddings V2, Rerank 1.0
- Anthropic Claude Haiku 3, Sonnet 3.5, Haiku 4.5, Sonnet 4.5 (JP-CRIS)
- Amazon Cohere Embed v4, Embed English, Embed Multilingual, Rerank 3.5
- Open AI gpt-oss-120b, gpt-oss-20b
- DeepSeek DeepSeek-V3.1
- Qwen Qwen3 Coder 480B A35B Instruct, Qwen3-Coder-30B-A3B-Instruct, Qwen3 32B (dense), Qwen3 235B A22B 2507

https://docs.aws.amazon.com/bedrock/latest/userguide/models-regions.html

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

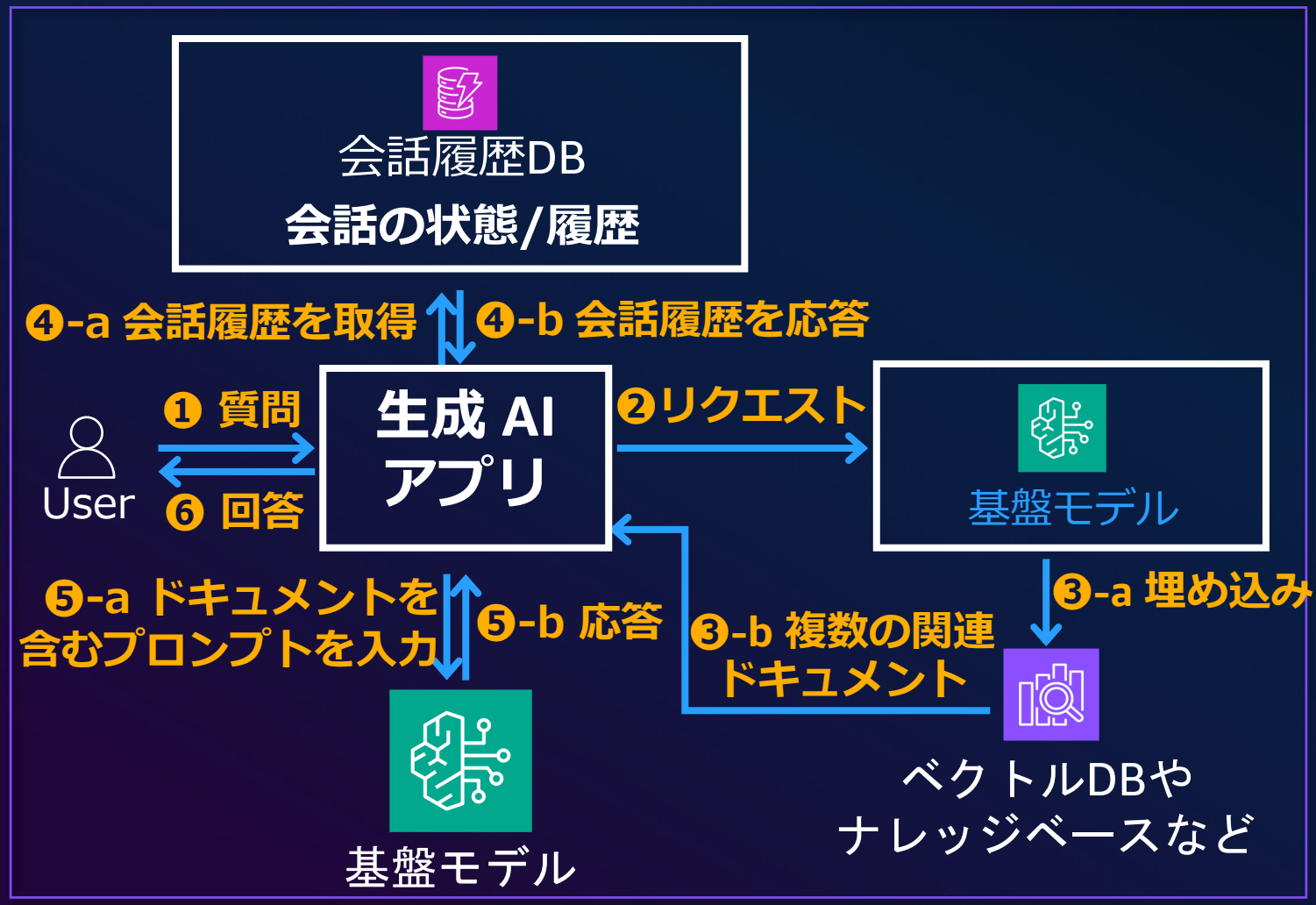
1. AWSの生成AIサービス

2. AI提供段階における、AIシステムの機密性、完全性、可用性を損なう脅威とAWSが提供している対策

3. まとめ

生成 AI アシスタントの一般的なワークフロー

RAG - 非構造化データ（ドキュメントなど）の場合

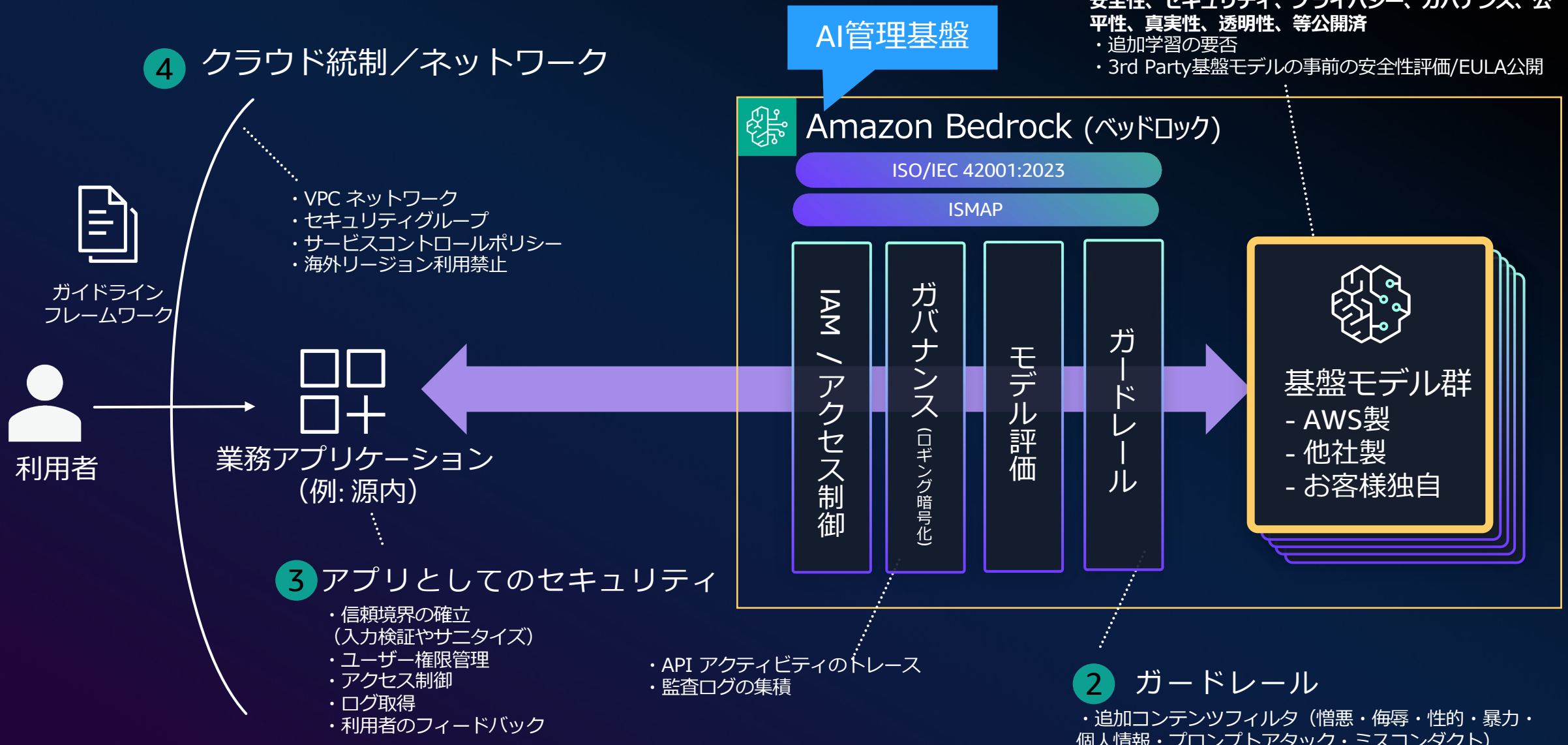


生成 AI アシスタントの一般的な RAG ワークフロー

1. ユーザーが質問
2. 質問に対する埋め込みを生成
3. ベクトル DB から関連性の高いドキュメントを複数個取得
4. (Optional) 会話履歴を取得
5. ドキュメントの内容含めたプロンプトを LLM に入力
6. 最終的な回答を送信

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

生成 AI アプリケーションにおける リスクの考慮点と、その多層防御



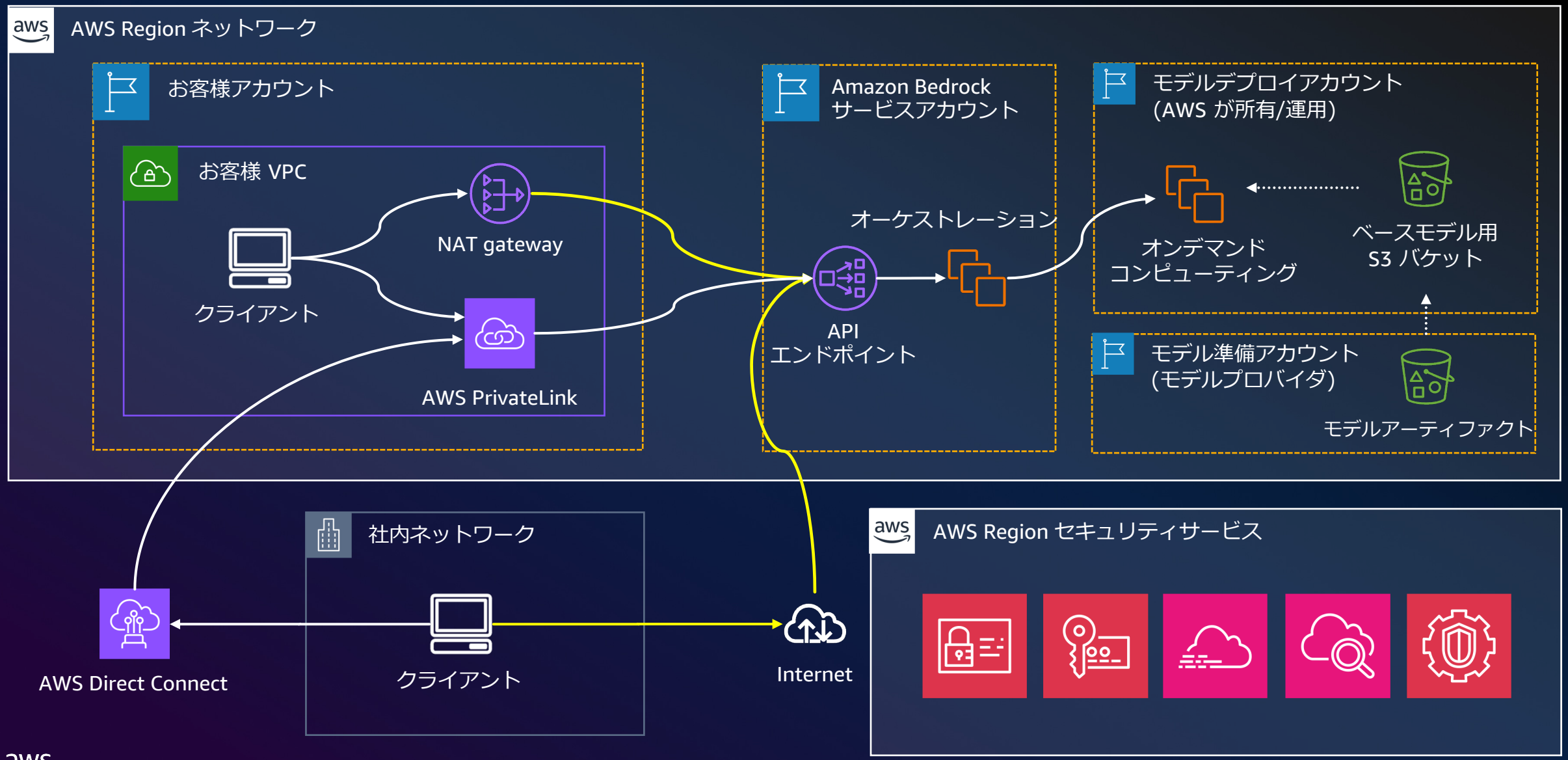
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

推論における分離

1. 顧客データは一切記録されません
2. デプロイされたモデルへの変更なし
3. ホストからのアウトバウンド接続なし
4. 分離型推論処理
5. 処理中にデータをキャッシュしない

aws
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock のデータフロー



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

2 ガードレール

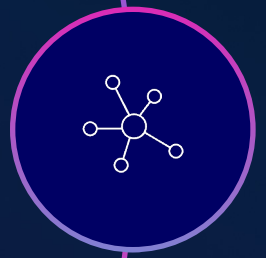
Amazon Bedrock Guardrails

アプリケーション要件に合わせてカスタムしてセーフガードを実装し、責任ある AI ポリシーに準拠

Amazon Bedrock 上の一部の基盤モデルでネイティブで提供されている保護と比較して、有害コンテンツを最大 85% 多くブロックし、RAG や要約でのユースケースにおけるハルシネーション回答の 75% 以上をフィルタリング

エージェント、ナレッジベース、Amazon Bedrock の基盤モデル、カスタムまたはサードパーティの基盤モデルのプロンプトとレスポンスを評価

有害コンテンツ、保護機能回避、プロンプトインジェクション攻撃をフィルタリングするしきい値を設定



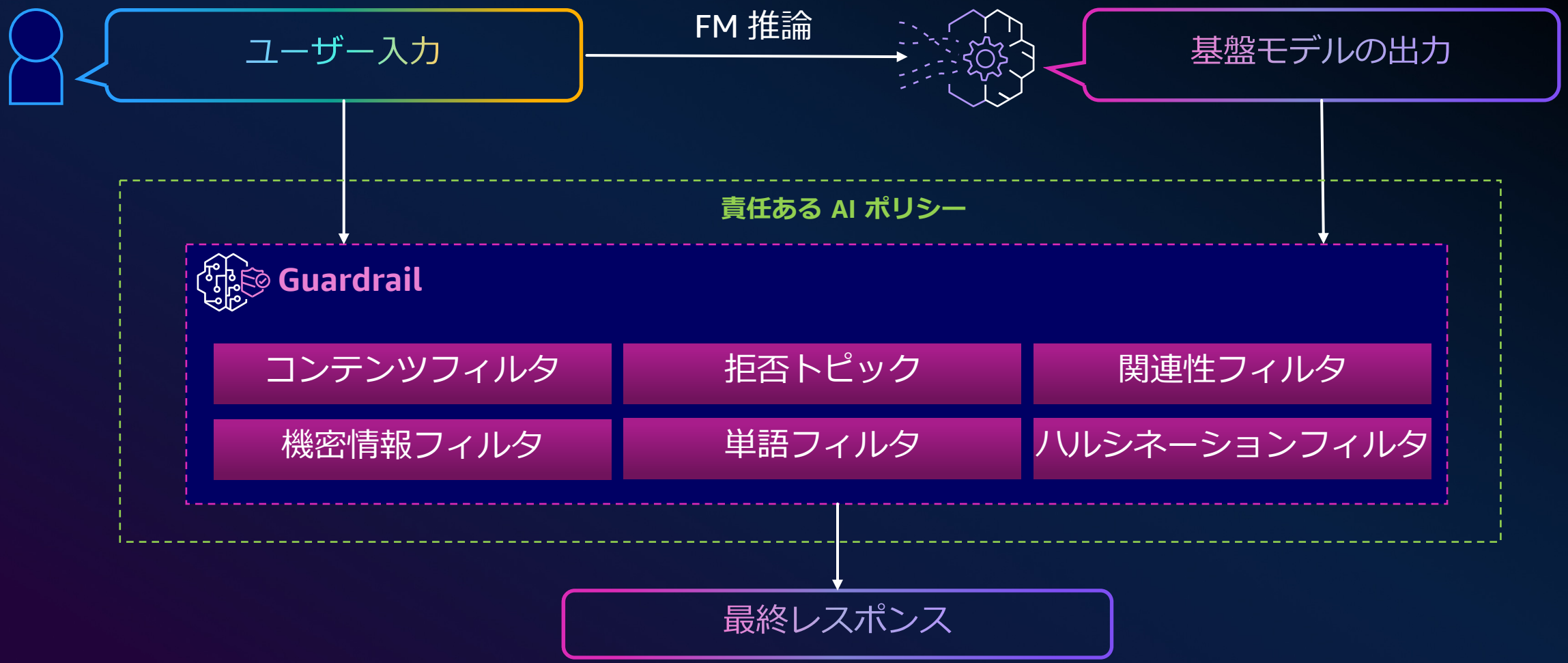
簡潔な自然言語の記述で禁止トピックを定義し拒否

生成 AI アプリケーションから個人を特定できる情報（PII）や機密情報を除去

コンテキストに基づいてモデルレスポンスの根拠と関連性を検出しハルシネーションをフィルタリング

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock Guardrails の仕組み



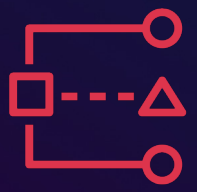
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

データ、モデル、出力へのアクセスを制限

生成 AI ワークロードのデータアクセス制御のための ID 活用とゼロトラスト

AWS IAM
Identity Center

Amazon Verified
Permissions



IAM
Access Analyzer

AWS Verified
Access

- IAM Identity Center と IAM Access Analyzer を使用して、トレーニングデータ、モデル、およびアプリケーションに最小権限権のポリシーを適用
- AWS Verified Access と Amazon Verified Permissions を使用してきめ細やかなアクセス制御を実現するゼロトラスト機能
- AWS Verified Access を使用してコスト、複雑さ、パフォーマンスの課題を解消

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ガバナンスと監査のサポート

包括的なモニタリングとログ機能

- Amazon CloudWatch を使用して**利用状況メトリクス**を追跡し、カスタムダッシュボードを構築
- AWS CloudTrail を使用して **API アクティビティをモニタリング**し、他のシステムをアプリケーションへ統合する際の**問題をトラブルシューティング**
- 準拠している規格：C5, CISPE, DoD CC SRG IL2, ENS High, FINMA, **FedRAMP** (Moderate/High), **GDPR**, **HIPAA BAA**, ISMAP, **ISO** and CSA STAR, MTCS, OSPAR, **PCI**, Pinakes, PiTuKri, **SOC 1, 2, and 3**

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

OWASP TOP 10 for Large Language Model Applications - 2025

UPDATED TOP 10 THREAT LIST

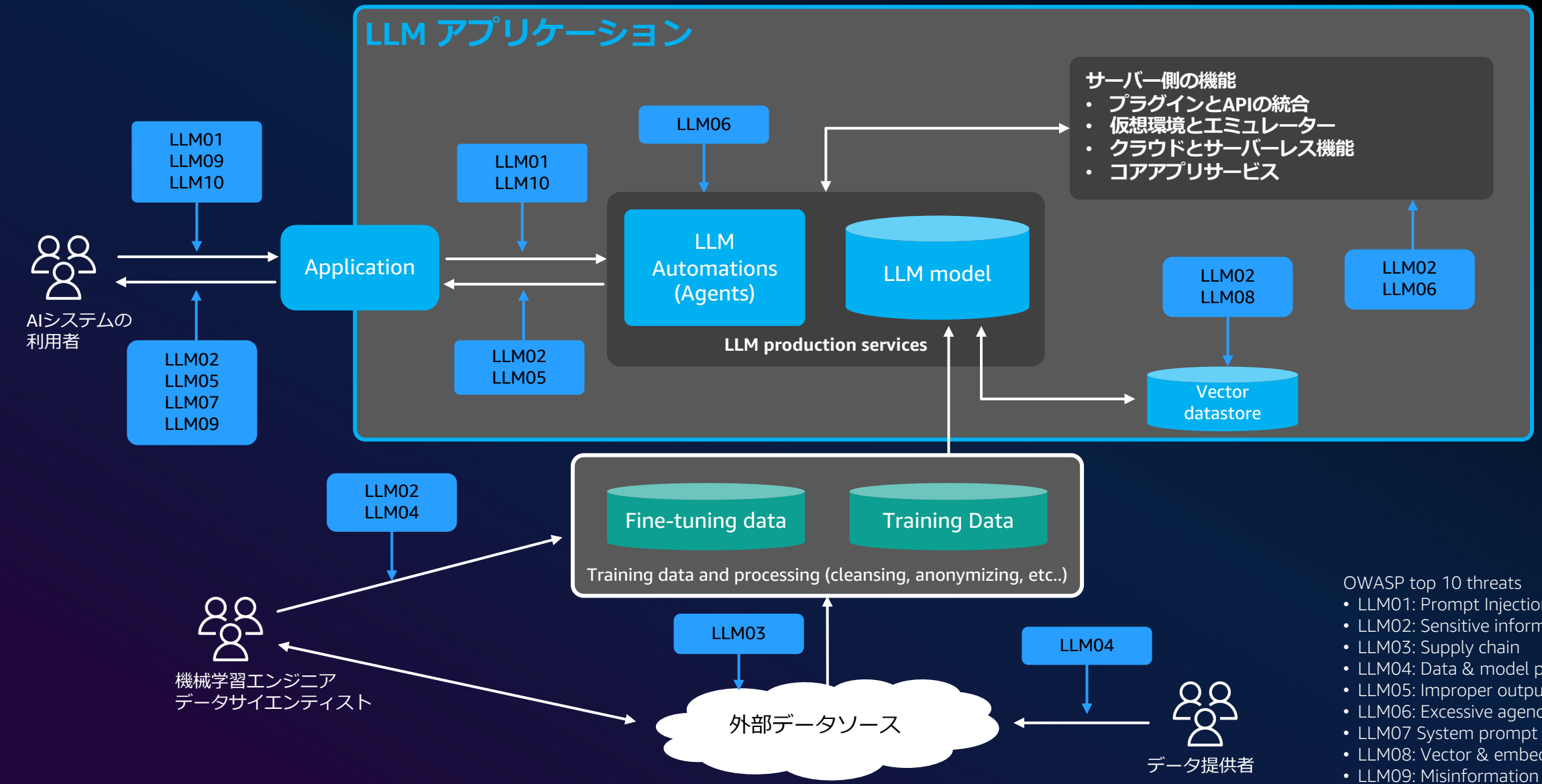
LLM01 Prompt Injection	LLM02 Sensitive Information Disclosure	LLM03 Supply chain	LLM04 Data and model poisoning	LLM05 Improper output handling
LLM06 Excessive Agency	LLM07 System prompt leakage	LLM08 Vector and embedding weaknesses	LLM09 Misinformation	LLM10 Unbounded consumption

https://genai.owasp.org/llm-top-10/

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

典型的な生成AIアプリケーション



OWASP top 10 threats
- LLM01: Prompt Injection
- LLM02: Sensitive information disclosure
- LLM03: Supply chain
- LLM04: Data & model poisoning
- LLM05: Improper output handling
- LLM06: Excessive agency
- LLM07 System prompt leakage
- LLM08: Vector & embeddings weaknesses
- LLM09: Misinformation
- LLM10: Unbounded consumption

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

OWASPで記載されているリスクと その対応を検討するお客様に提供するAWSサービスの対応表

OWASP Threats	Service / Features to help mitigate OWASP threats
LLM01 (Prompt Injection)	**Amazon Bedrock Guardrails** (prompt attack)
LLM02 (Sensitive information disclosure)	**Amazon Bedrock Guardrails** (sensitive data detection) **Amazon Bedrock Guardrails** (Denied topics) **Amazon Bedrock Guardrails** (custom words) **Amazon Bedrock Guardrails** (regex) **IAM** (controlled access to KB & source data) **Macie** (detecting sensitive data)
LLM03 (Supply Chain)	Amazon Bedrockで提供するモデルプロバイダーアカウントの制限*1
LLM04 (Data and model poisoning)	**IAM**
LLM05 (Improper output handling)	**Amazon Bedrock Guardrails** (profanity) **Amazon Bedrock Guardrails** (sensitive data detection) **CodeGuru Security** **Amazon Inspector***
LLM06 (Excessive Agency)	**IAM**
LLM07 (System prompt leakage)	Amazon Bedrockで提供するモデルプロバイダーアカウントの制限*1
LLM08 (Vector and embedding weaknesses)	**IAM**
LLM09 (Misinformation)	**Amazon Bedrock Guardrails** (Contextual grounding checks) **Amazon Bedrock Guardrails** (Automated reasoning checks)
LLM10 (Unbounded consumption)	**WAF & Shield**

*1
3rd Partyモデルにおいて、モデルプロバイダーが管理する「モデル準備アカウント」を使ってモデルをAWSに準備し、AWSはモデルプロバイダーがアクセス権限を持たない「モデルデプロイアカウント」でテスト・評価をして、ユーザが利用できる環境を準備します。

aws
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

1. AWSの生成AIサービス

2. AI提供段階における、AIシステムの機密性、完全性、可用性を損なう脅威とAWSが提供している対策

3. まとめ

まとめ

AWSでは、様々なお客様の要求に対応可能な、安全な生成AIサービスを提供したいと考えています。

生成AIを取り巻くリスク環境は絶えず変化しているため、AWSはAmazon Bedrock Guardrailsなどの機能をアップデートを提供し続けています。

今後の生成AI活用の方向性として、AIエージェント、Agentic AIが活用されつつある状況が見られます。技術リスクとして、Agentic AIでは、主なリスクとして、「アライメント問題」が追加すべき事項になると考えます。

「アライメント問題」 - AIの目標や行動が人間の意図や価値観と一致しない、不正確な命令の解釈、アクセス権限の濫用のリスク

AWSでは、Amazon Bedrock AgentCoreというAgentic AIサービスの提供を開始し、MCPサーバの実現や、既存のMCPサーバとの連携における、完全なセッション分離、Amazon VPC 接続、AWS PrivateLink サポート、包括的なコントロールなど、エンタープライズグレードのセキュリティ機能をユーザに提供開始しています。

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Thank you!

Yoichi Takizawa

Director,
Japan Public Sector Tech Business unit,
Amazon Web Services Japan G.K.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.