Cloud Run ではじめる LLM/MCP サーバー運用とセキュリティ対策

新井 雅也
Syspective Inc.

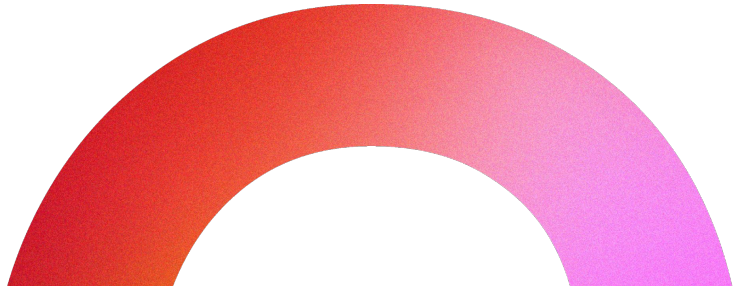
Google Cloud Next Tokyo

新井 雅也

株式会社 Synspective
プリンシパル
GDE (Google Developer Experts)

本日お話すること

LLM や MCP サーバーを Cloud Run で運用する際の
アーキテクチャ例とセキュリティ対策のポイント



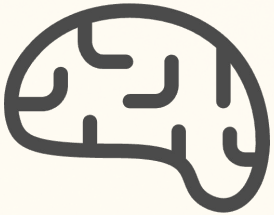
本日お話しないこと

- LLM や MCP サーバー、AI アプリケーション自体の実装に関連した内容
- AI アプリケーションに着目したセキュリティ対策
  - 一部関連する箇所のみ言及

みなさん、普段の業務や開発で
LLM や MCP サーバーを利用していますか？

まずは簡単に LLM と MCP のおさらいから

LLM（大規模言語モデル）とは

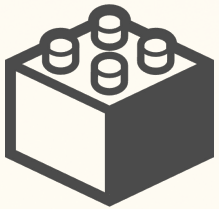


膨大なデータと深層学習により構築された言語モデル

- ラフに伝えるなら「プログラムがたくさんの言葉を学ぶことによって、人間のように文章を理解し作り出せるもの」

例）Gemini 2.5, GPT-4.5, Claude Sonnet 4

MCP とは



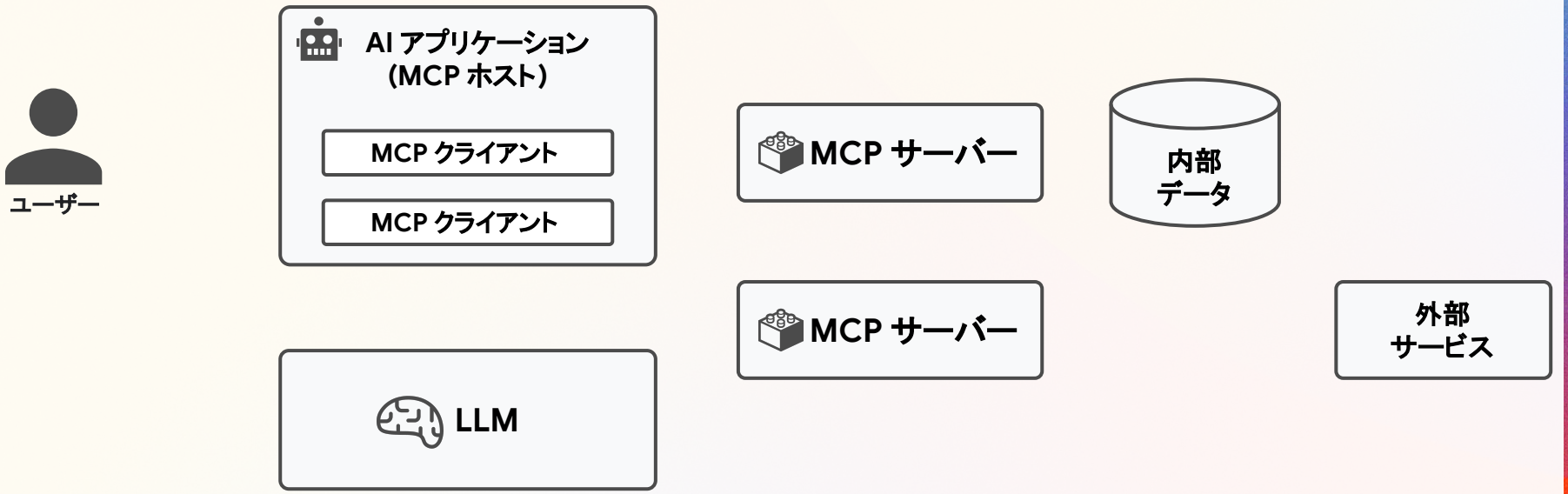
AI モデルが外部のツールを呼び出すための **標準化プロトコル**

- MCP ホスト、MCP クライアント、MCP サーバーで構成

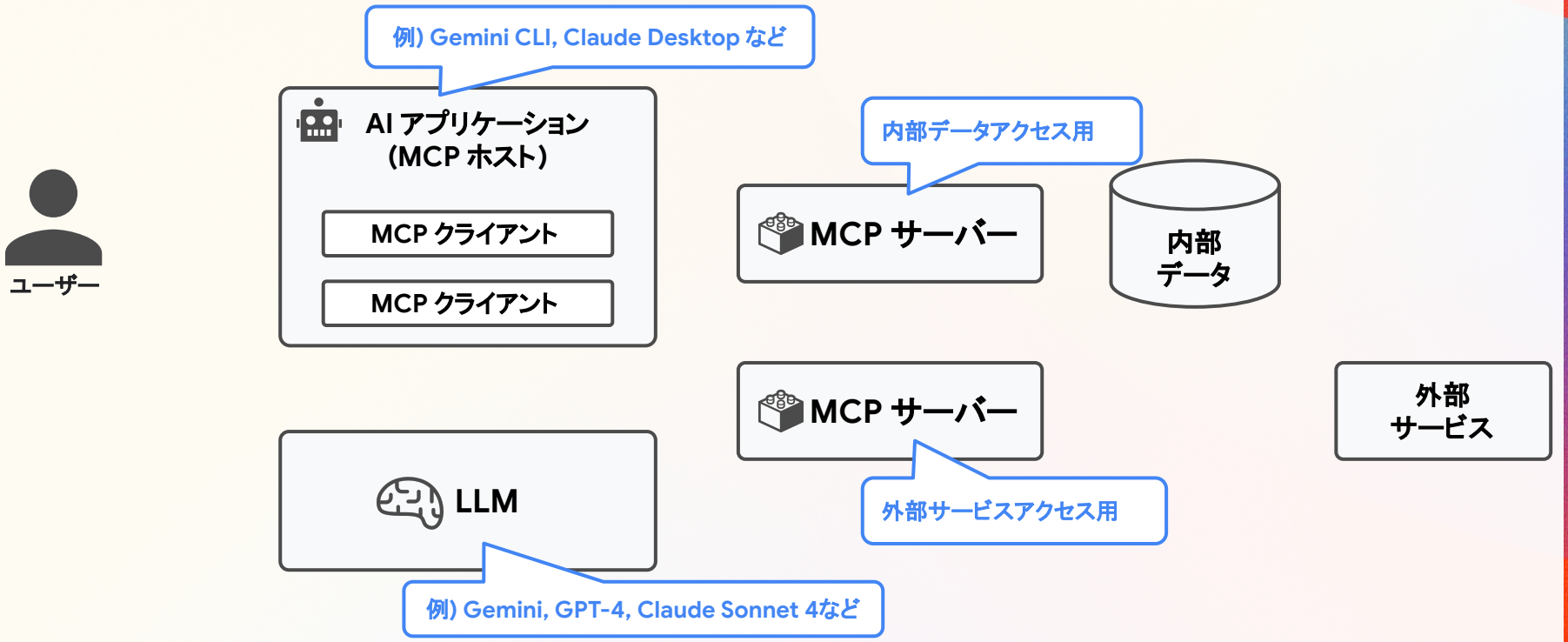
LLM が学習した以外の情報に対して、「 LLM の代わりに外部から情報を提供する」役割を担うのが **MCP サーバー**

LLM / MCP サーバーが利用された処理フロー

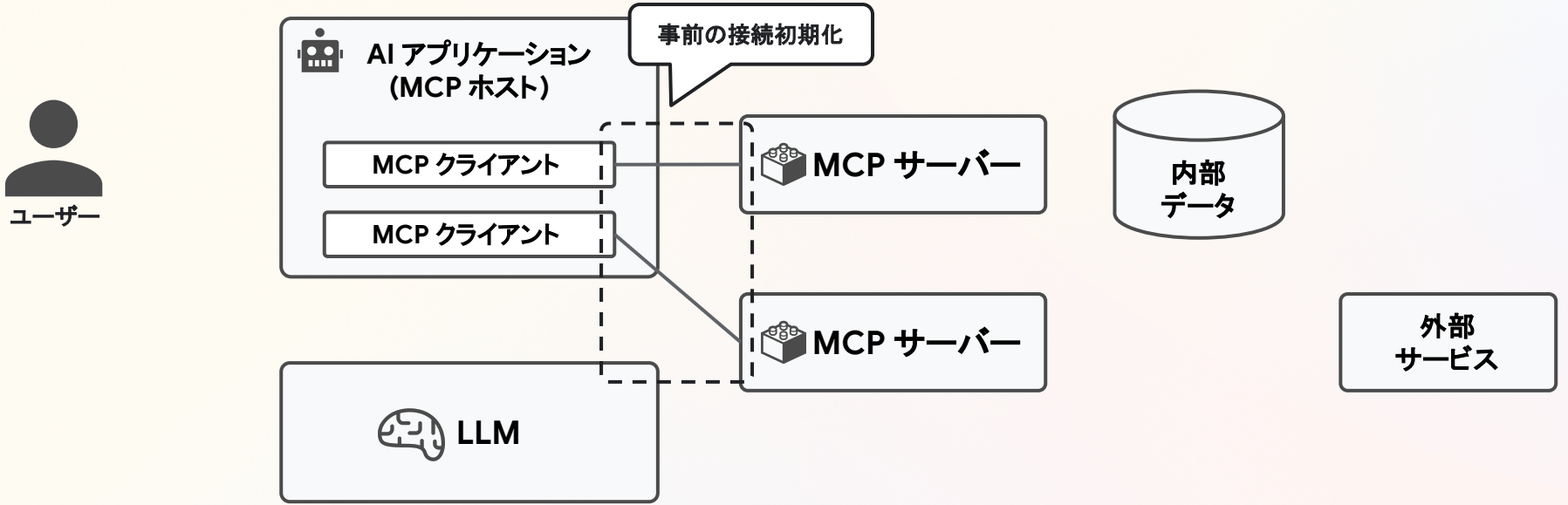
LLM / MCP サーバーが利用された処理フロー



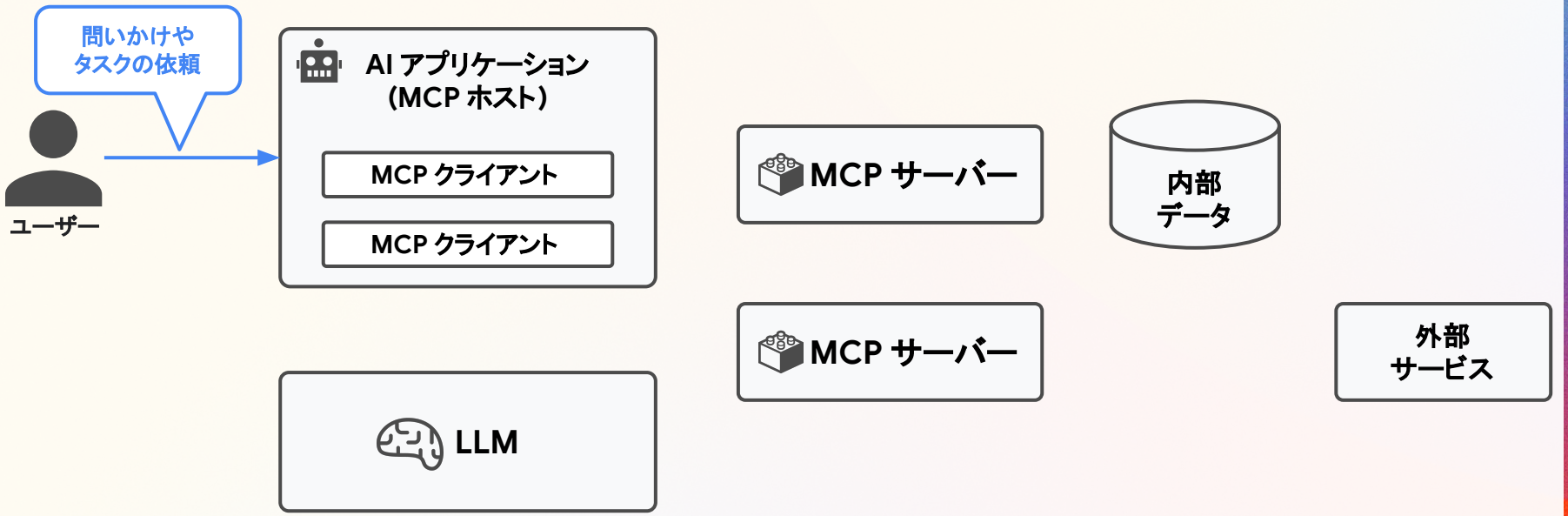
LLM / MCP サーバーが利用された処理フロー



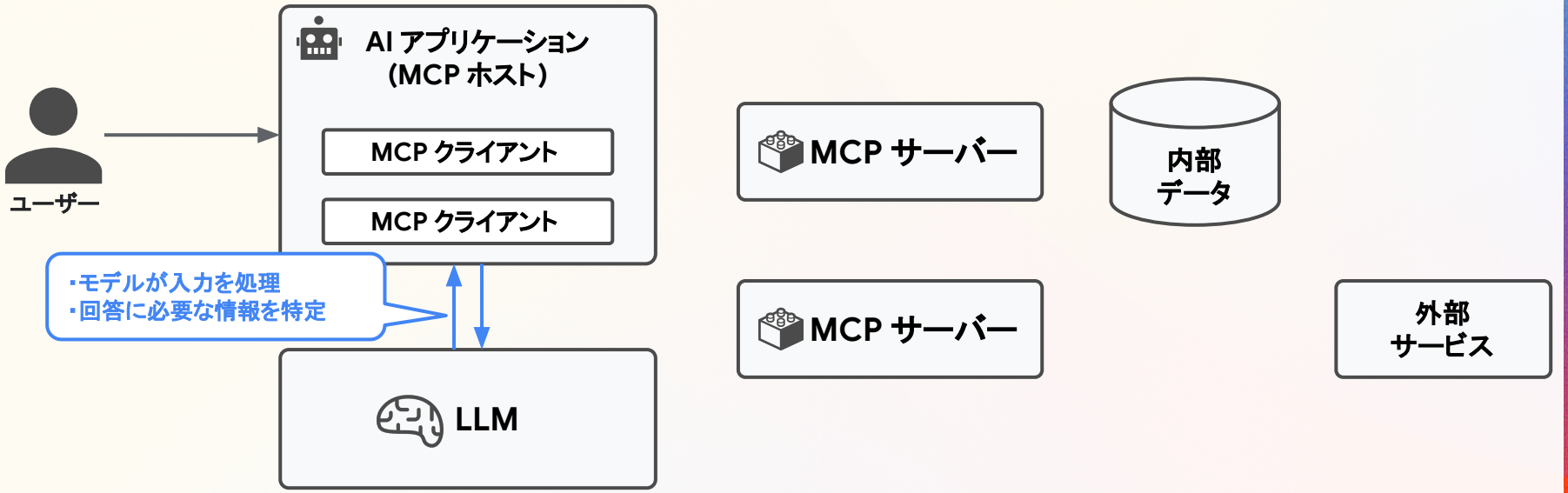
LLM / MCP サーバーが利用された処理フロー



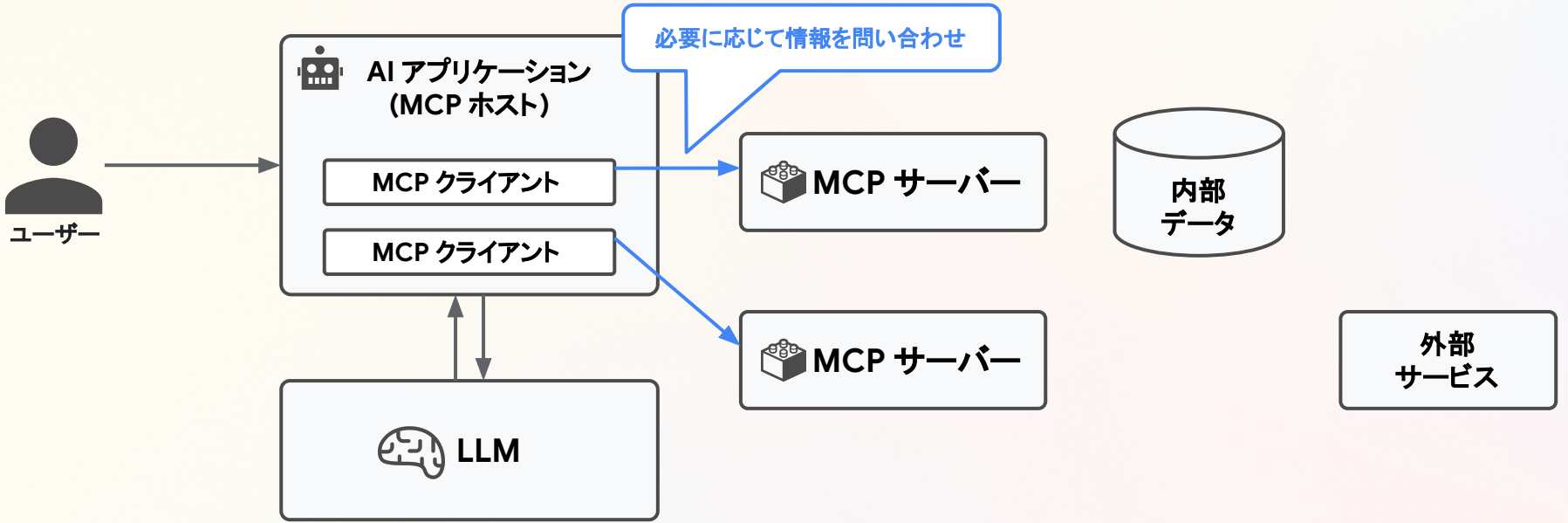
LLM / MCP サーバーが利用された処理フロー



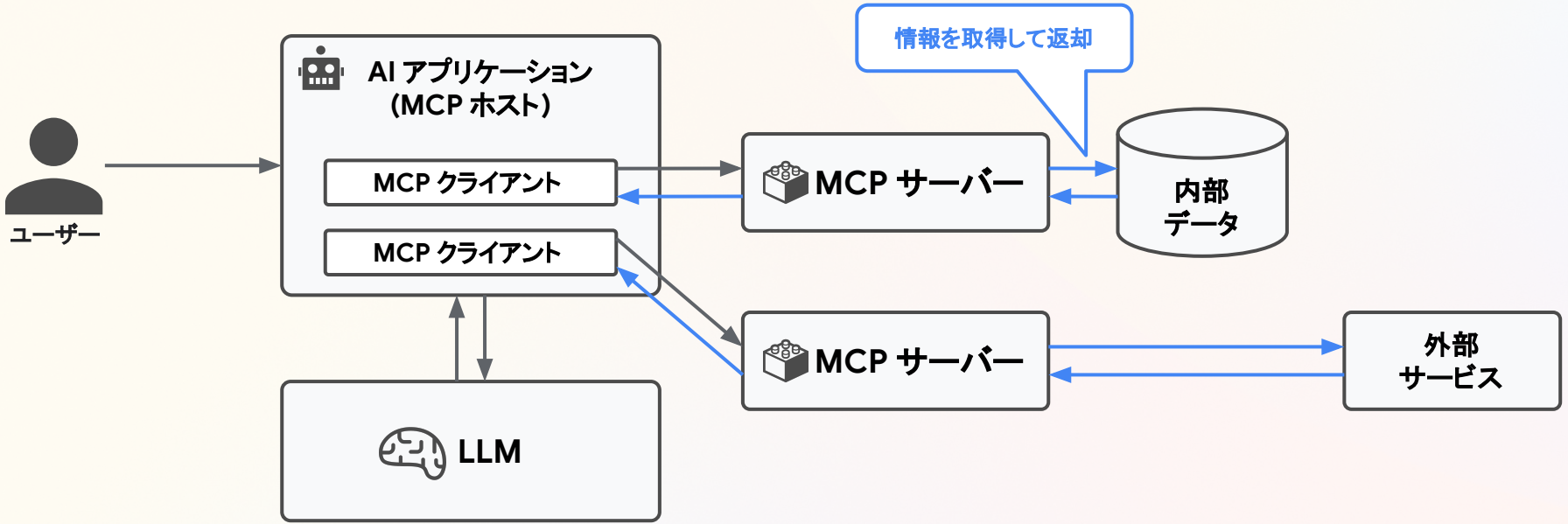
LLM / MCP サーバーが利用された処理フロー



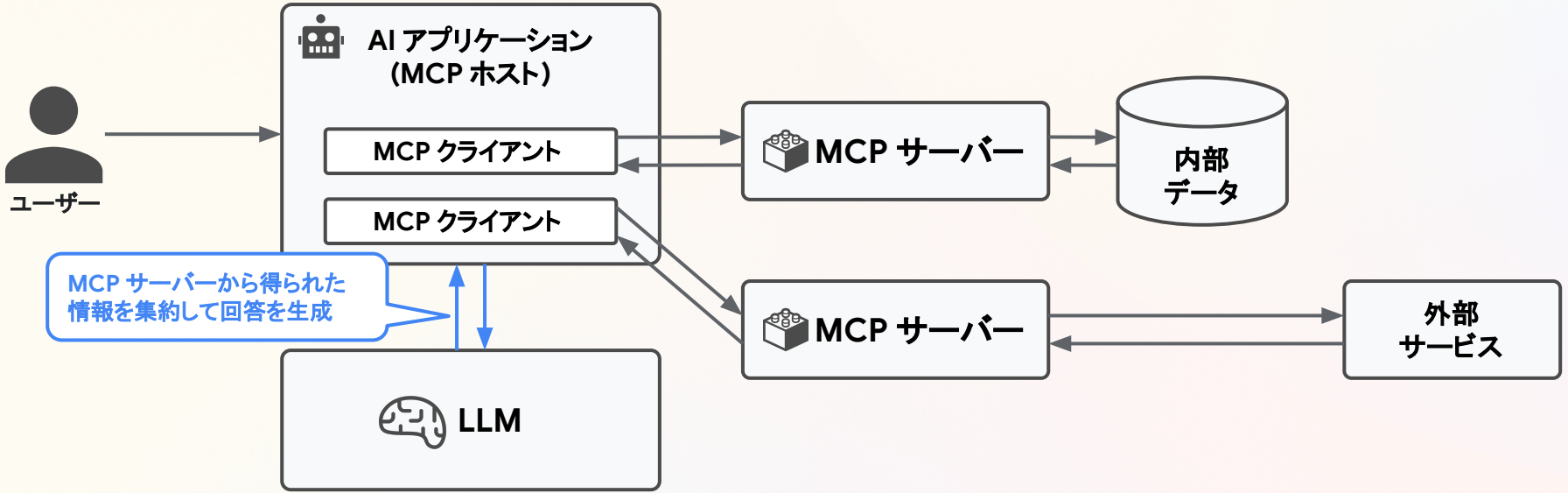
LLM / MCP サーバーが利用された処理フロー



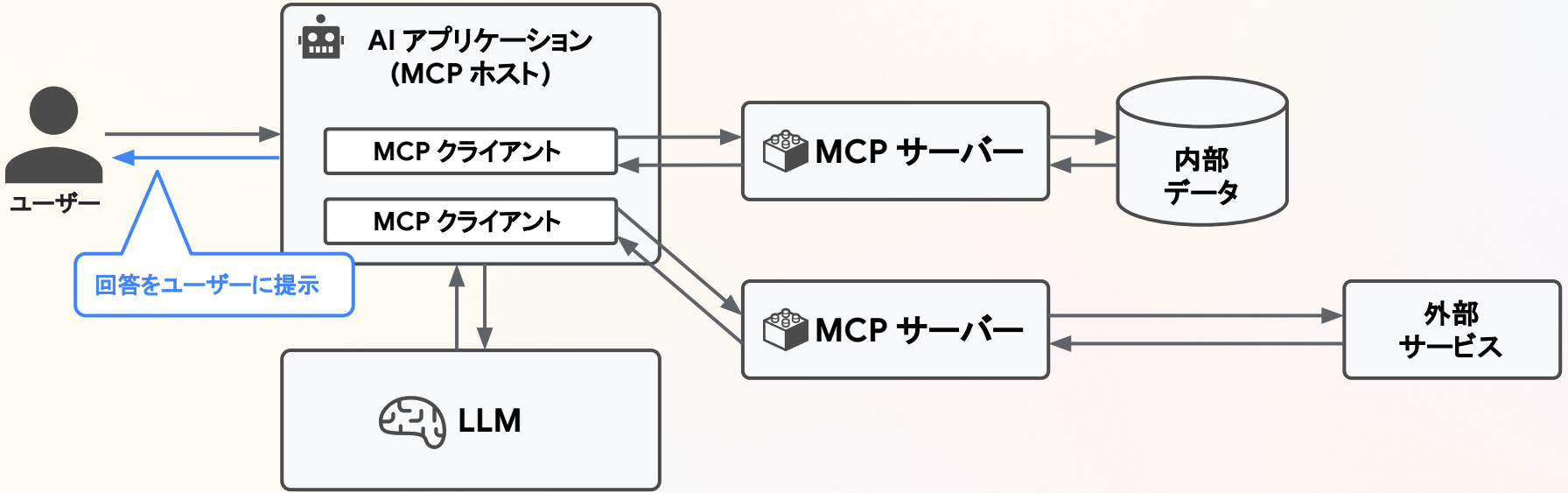
LLM / MCP サーバーが利用された処理フロー



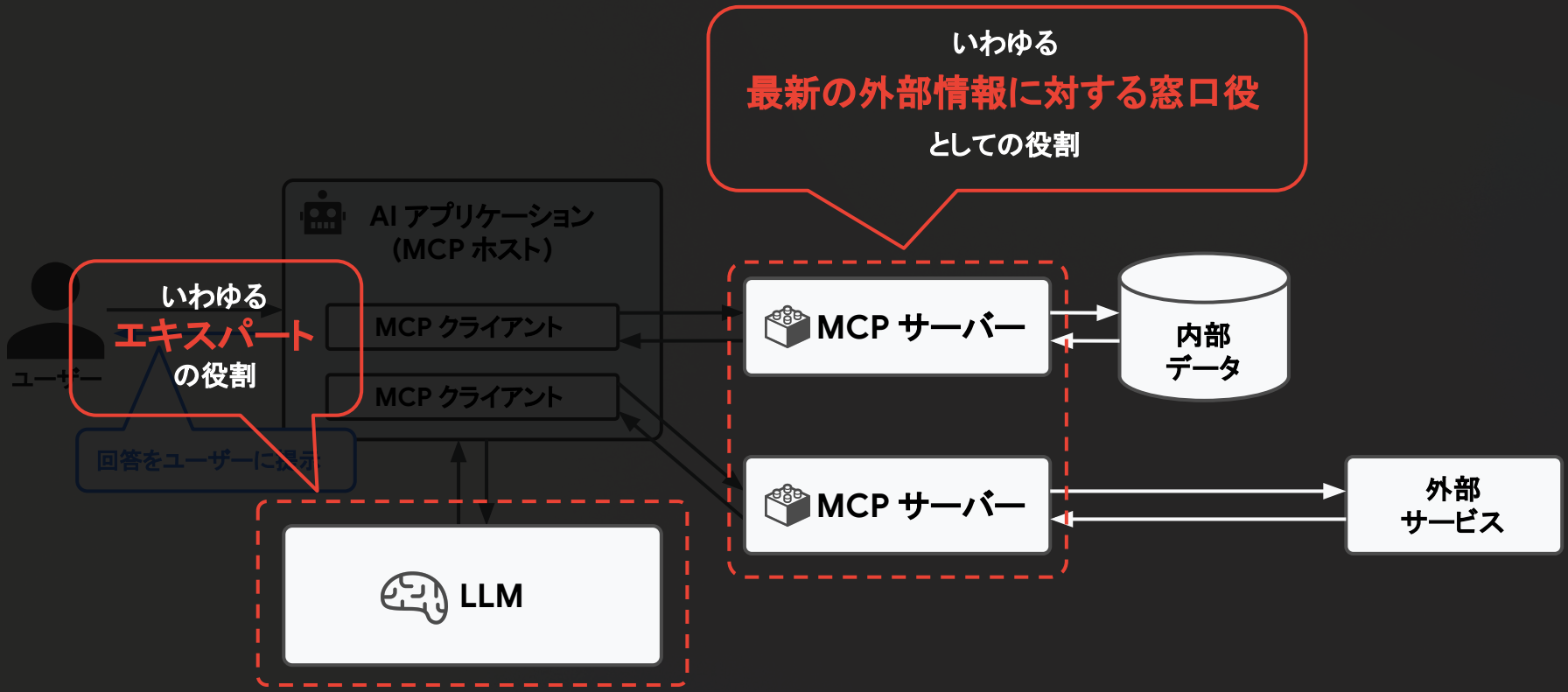
LLM / MCP サーバーが利用された処理フロー



LLM / MCP サーバーが利用された処理フロー



ここまでのまとめ


いわゆる
最新の外部情報に対する窓口役
としての役割
AI アプリケーション
(MCP ホスト)
いわゆる
エキスパート
の役割
ユーザー
MCP クライアント
MCP クライアント
回答をユーザーに提示
MCP サーバー
MCP サーバー
内部
データ
外部
サービス
LLM

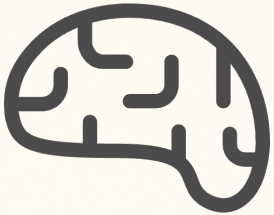
自前の

みなさん、普段の業務や開発で LLM や MCP サーバーを利用していますか？

自前の

なぜ、LLM や MCP サーバーを使う必要があるのでしょうか？

LLM を自前で用意するユースケース



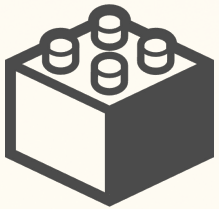
ドメインに特化した モデルの利用やカスタマイズ

- 金融知識、法的知識

ガバナンス・コンプライアンス要件

- 特定地域外へのデータ転送・保存が制限される規制産業等

MCP サーバーを自前で用意するユースケース



ビジネスユースケースに特化した情報の参照

既存のシステム・内部 API による情報参照

- 公開 MCP サーバーでは提供されていない情報の取得

自前の

なぜ、LLM や MCP サーバーを使う場合に Cloud Run を選択するのか？

一番のメリットは ゼロスケール

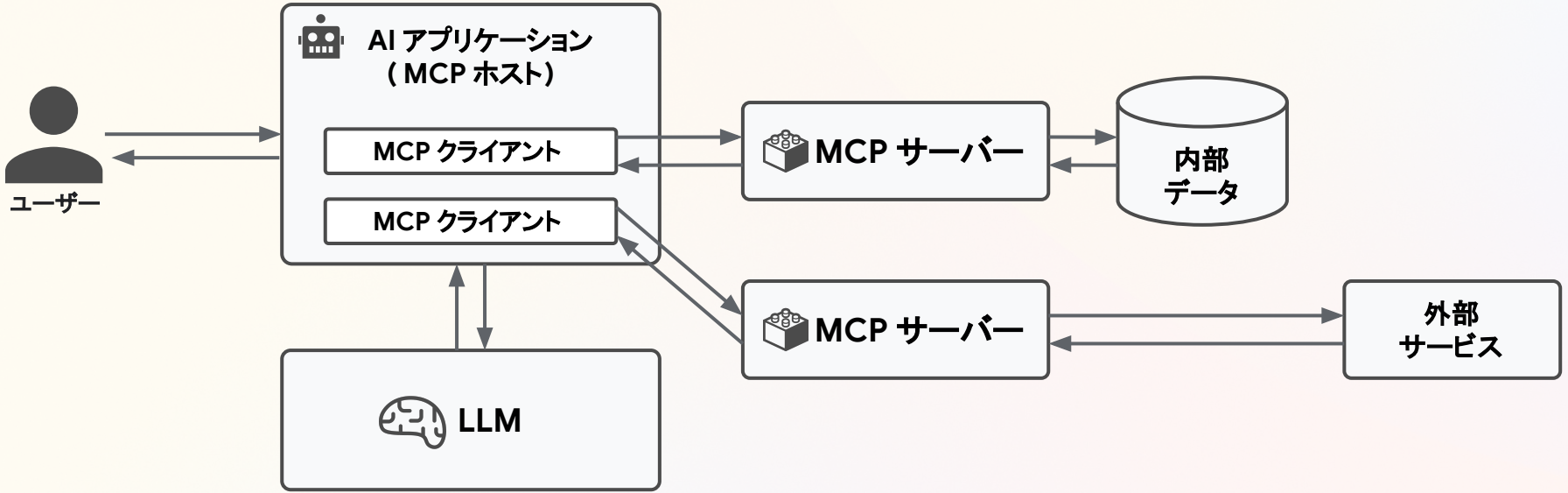
- 未利用時はインスタンス数を 0 にできる
  - アイドルコストが発生しない
  - 後述する GPU 利用時のコストを抑えられる
- 散発的な利用特性 と相性が良い
  - 課金は秒単位

LLM ワークロード向けに GPU が利用可能

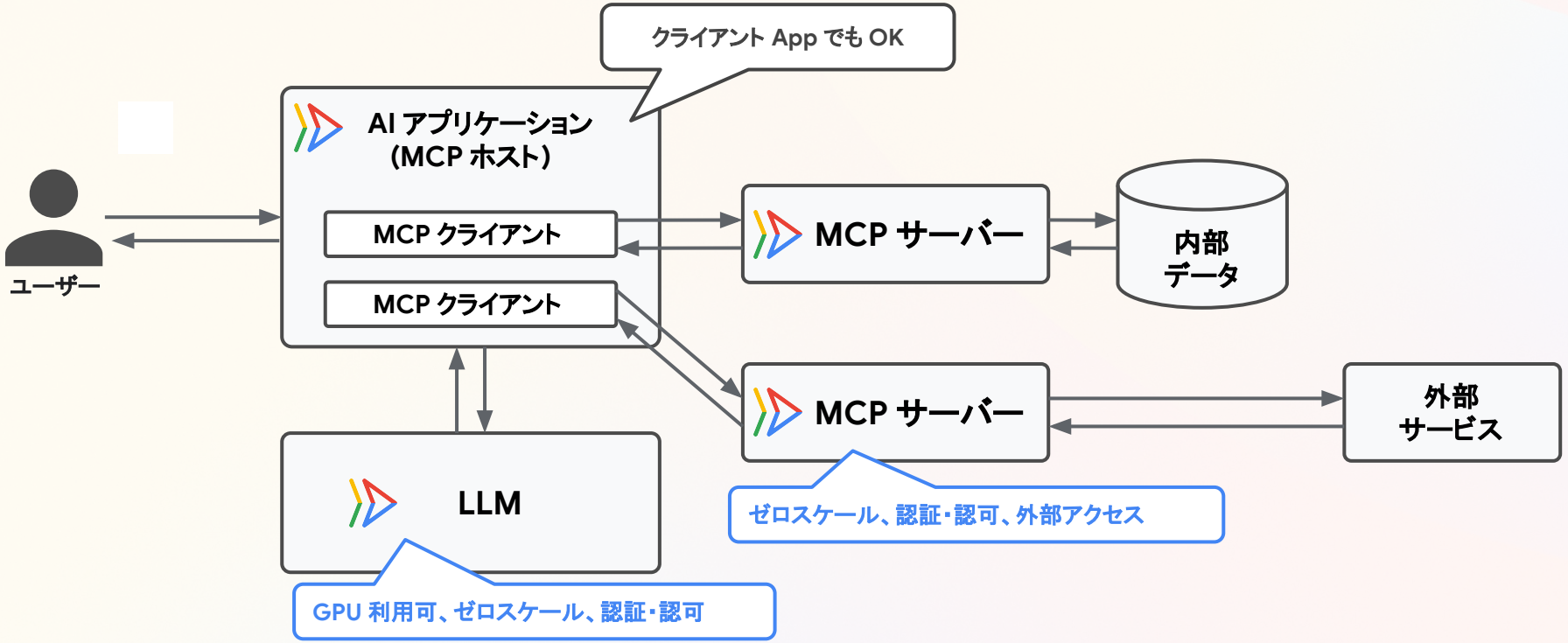
- GPU のドライバーインストール等は不要
- 数～数十秒 でデプロイ
- 初期利用時は クォーター増加が必要
- 残念ながら、本発表準備時点 （2025 年 7 月 23 日)では
  東京・大阪リージョンでは未対応

コンプライアンス要件の観点では
今後に期待 ...

Cloud Run による LLM / MCP サーバーのアーキテクチャ例

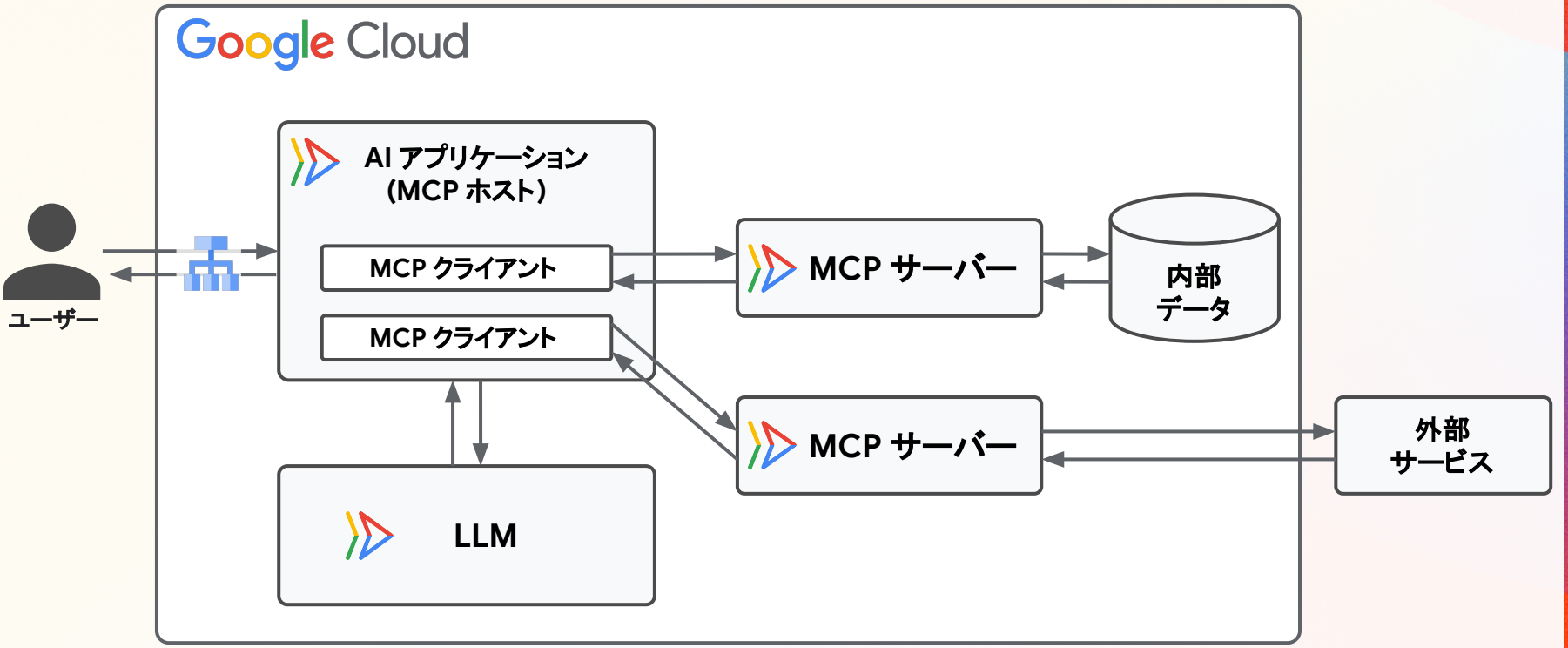


Cloud Run による LLM / MCP サーバーのアーキテクチャ例



LLM や MCP サーバーを Cloud Run で運用する際、必要なセキュリティ対策のアプローチ は？

想定するアーキテクチャ構造



LLM / MCP サーバーだろうと、
結局のところは Web アプリケーション

LLM / MCP サーバーだろうと、結局のところは Web アプリケーション

- 依然として、従来の Web アプリケーションに関するセキュリティ対策 は効果的
- 加えて、AI アプリケーション特有の観点 も考慮

Web App セキュリティ × 生成 AI App セキュリティ

Web App セキュリティ × 生成 AI App セキュリティ

OWASP Top 10

Web アプリケーションにとって
最も重要なセキュリティリスク を
まとめたもの

※https://owasp.org/Top10/

OWASP Top 10 for AI Applications 2025

生成 AI アプリケーションにおける
主要な 10 のセキュリティリスク を
まとめたもの

※https://genai.owasp.org/llm-top-10/

Web App セキュリティ × 生成 AI App セキュリティ

OWASP Top 10

不適切な アクセス制御	暗号化の失敗
インジェクション	セキュアでない 設計
セキュリティ設定の 不備	脆弱で古い コンポーネント
識別と認証の 失敗	ソフトウェアと データの整合性の失敗
セキュリティログと モニタリングの失敗	SSRF

OWASP Top 10
for AI Applications 2025

プロンプト インジェクション	機密情報の漏えい
サプライチェーン	データやモデルの 汚染
不適切な出力の ハンドリング	過剰な 代理行為
システムプロンプト の流出	ベクトル化と 埋め込みの脆弱性
不正確な情報	際限のない消費

LLM や MCP サーバーを Cloud Run で運用する際、Google Cloud 観点の具体的な セキュリティ対策 は？

Web App セキュリティ × 生成 AI App セキュリティ

OWASP Top 10

不適切な アクセス制御	暗号化の失敗
インジェクション	セキュアでない 設計
セキュリティ設定の 不備	脆弱で古い コンポーネント
識別と認証の 失敗	ソフトウェアと データの整合性の失敗
セキュリティログと モニタリングの失敗	SSRF

OWASP Top 10
for AI Applications 2025

プロンプト インジェクション	機密情報の漏えい
サプライチェーン	データやモデルの 汚染
不適切な出力の ハンドリング	過剰な 代理行為
システムプロンプト の流出	ベクトル化と 埋め込みの脆弱性
不正確な情報	際限のない消費

本発表にて割愛する項目

OWASP Top 10

開発プロセスや
アプリケーション側実装に依存する観点

- インジェクション
- セキュアでない設計
- 脆弱で古いコンポーネント
- 識別と認証の失敗
- ソフトウェアとデータの整合性の失敗
- SSRF

OWASP Top 10 for AI Applications 2025

LLM 出力や挙動の妥当性に関する観点

- システムプロンプトの流出
- ベクトル化と埋め込みの脆弱性
- 不正確な情報

Web App セキュリティ × 生成 AI App セキュリティ

OWASP Top 10

不適切な アクセス制御	暗号化の失敗
インジェクション	セキュアでない 設計
セキュリティ設定の 不備	脆弱で古い コンポーネント
識別と認証の 失敗	ソフトウェアと データの整合性の失敗
セキュリティログと モニタリングの失敗	SSRF

OWASP Top 10
for AI Applications 2025

プロンプト インジェクション	機密情報の漏えい
サプライチェーン	データやモデルの 汚染
不適切な出力の ハンドリング	過剰な 代理行為
システムプロンプト の流出	ベクトル化と 埋め込みの脆弱性
不正確な情報	際限のない消費

Cloud Run のIAM 権限に関連した対策

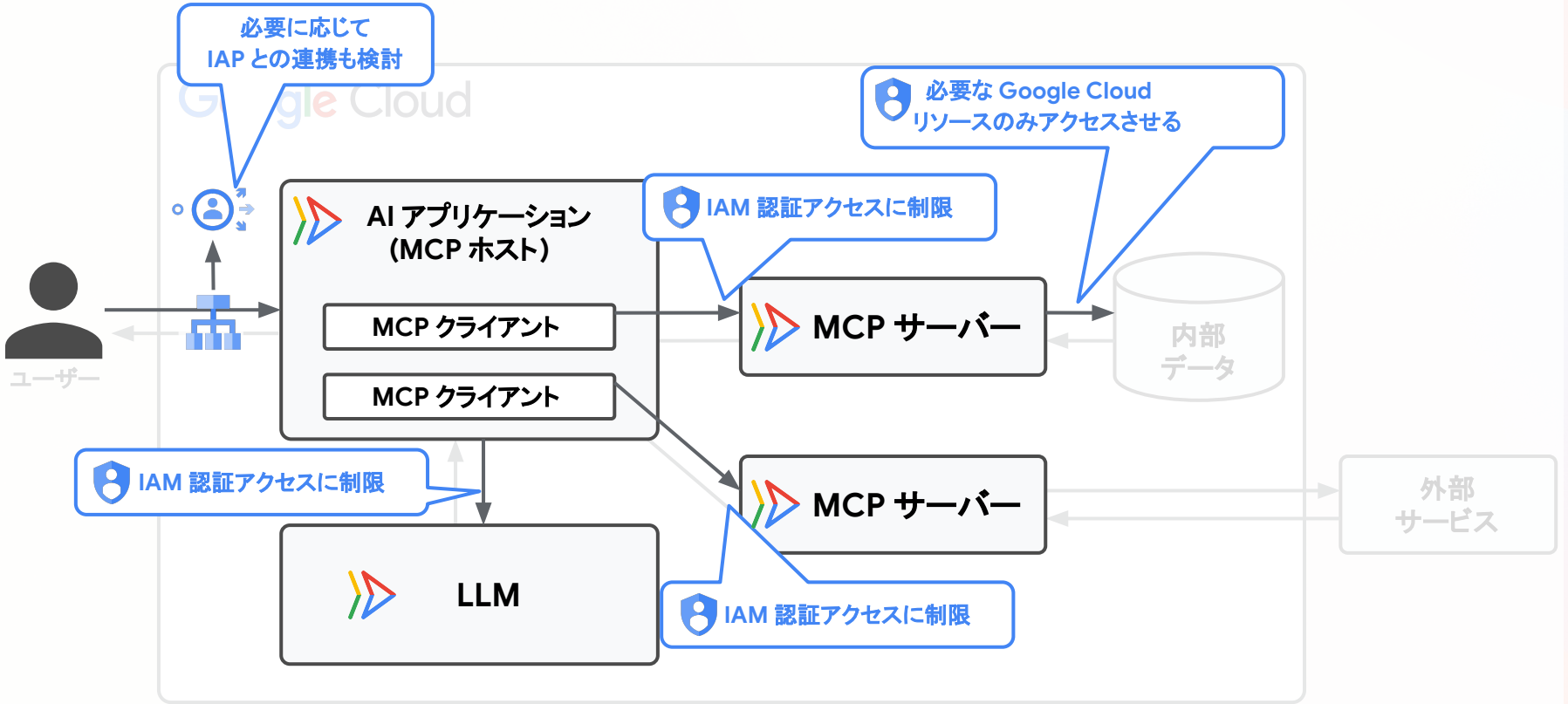
OWASP Top 10

- 不適切な
  アクセス制御
- セキュリティ設定の
  不備

OWASP Top 10
for AI Applications 2025

- プロンプト
  インジェクション
  - 意図しないLLM の
    挙動に対する考慮が必要
- 機密情報の漏えい
  - 意図しない機密データの
    出力への対策が必要

IAM による Cloud Run 実行とその他アクセスを制御



通信・データの暗号化に関連した対策

OWASP Top 10

暗号化の失敗

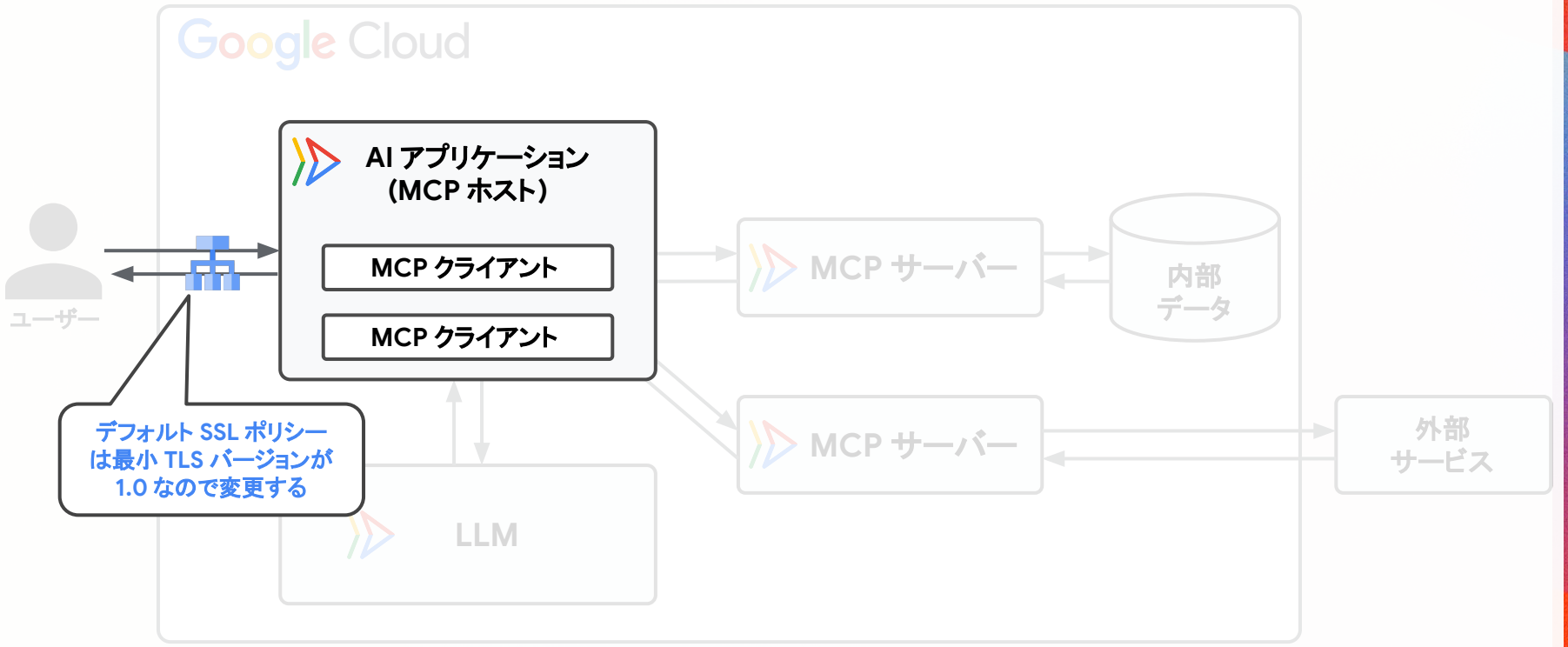
脆弱な暗号化アルゴリズムの利用に対する考慮

セキュリティ設定の不備

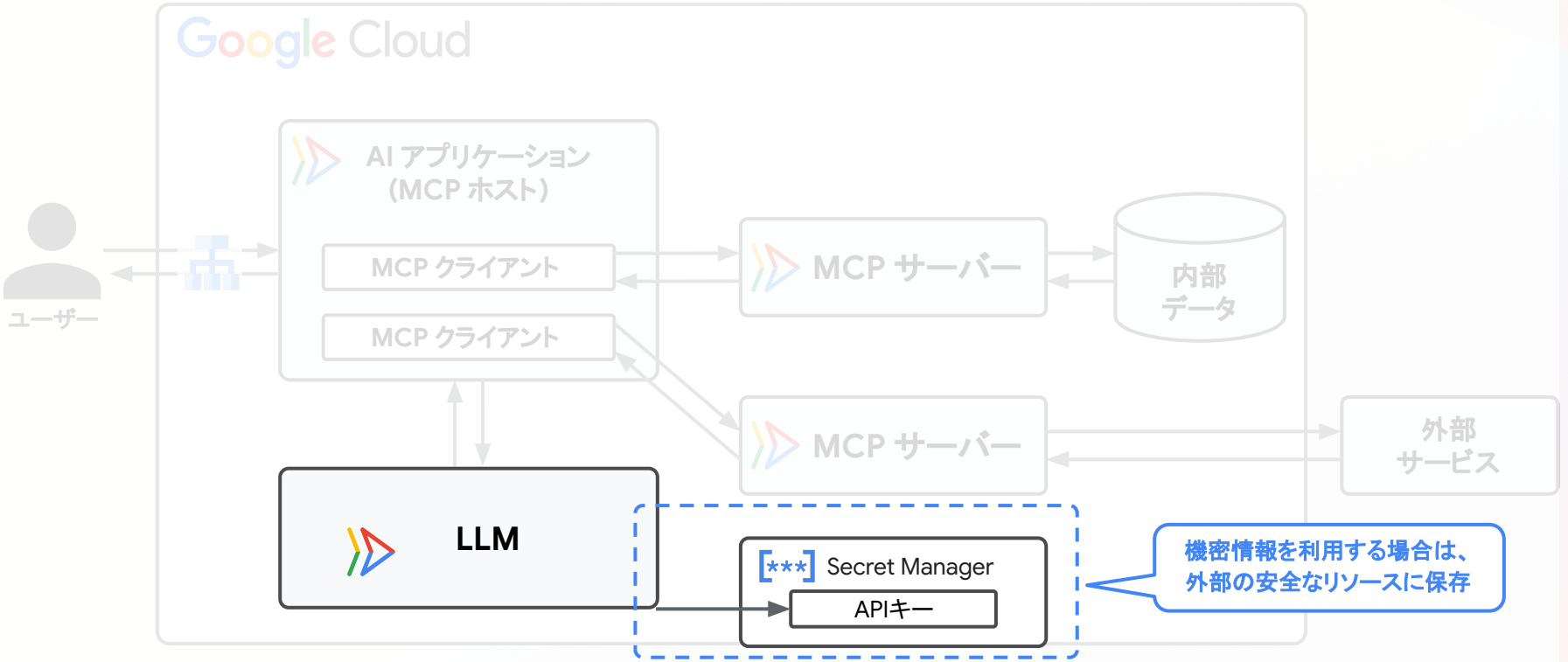
デフォルト値の利用や適切な機密情報の管理

OWASP Top 10 for AI Applications 2025

インターネットアクセスに External ALB を利用している場合は ポリシーの設定に注意



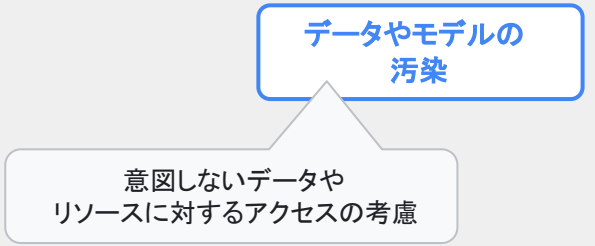
Secret Manager による機密情報の外部受け渡し



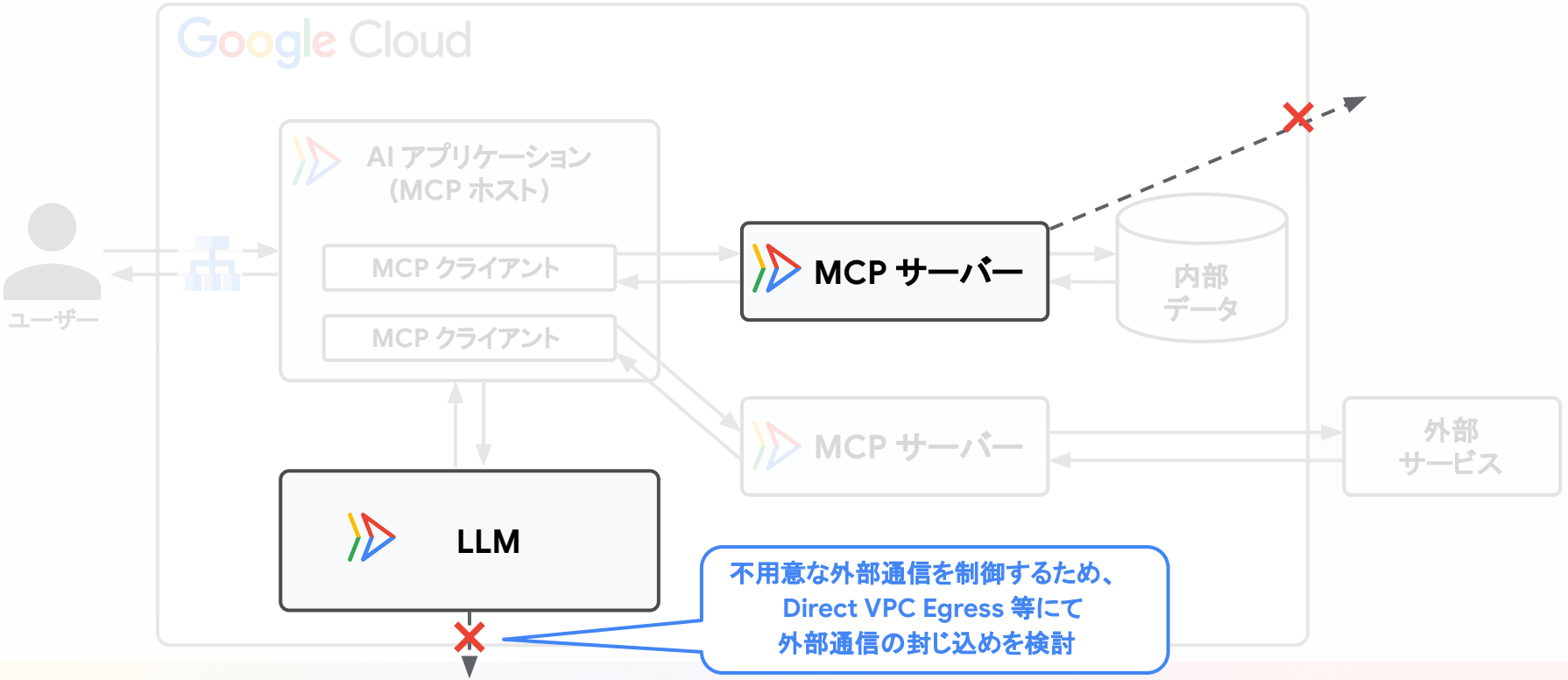
意図しないデータアクセス制御に関連した対策

OWASP Top 10

OWASP Top 10 for AI Applications 2025



外部通信が不要な場合は制御



異常な振る舞いに対するロギング観点の対策

OWASP Top 10

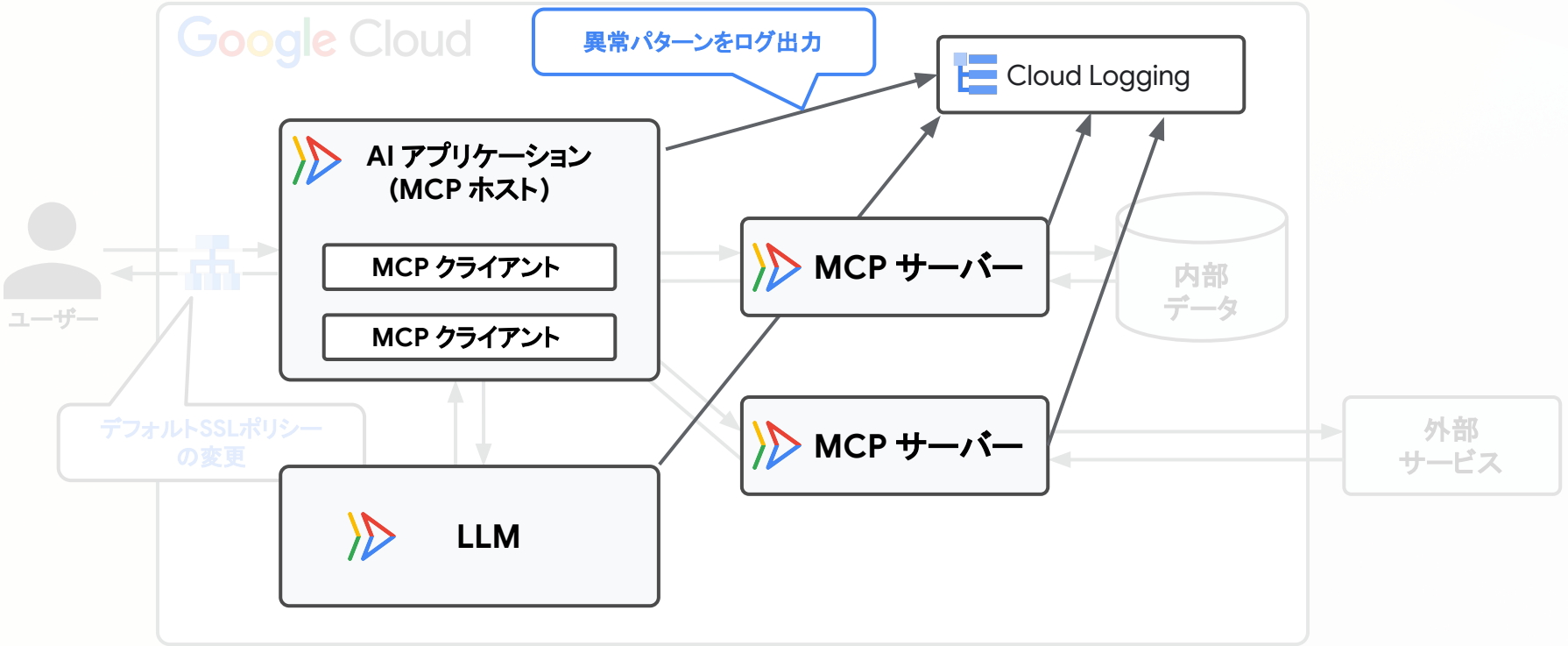
セキュリティログと
モニタリングの失敗

OWASP Top 10
for AI Applications 2025

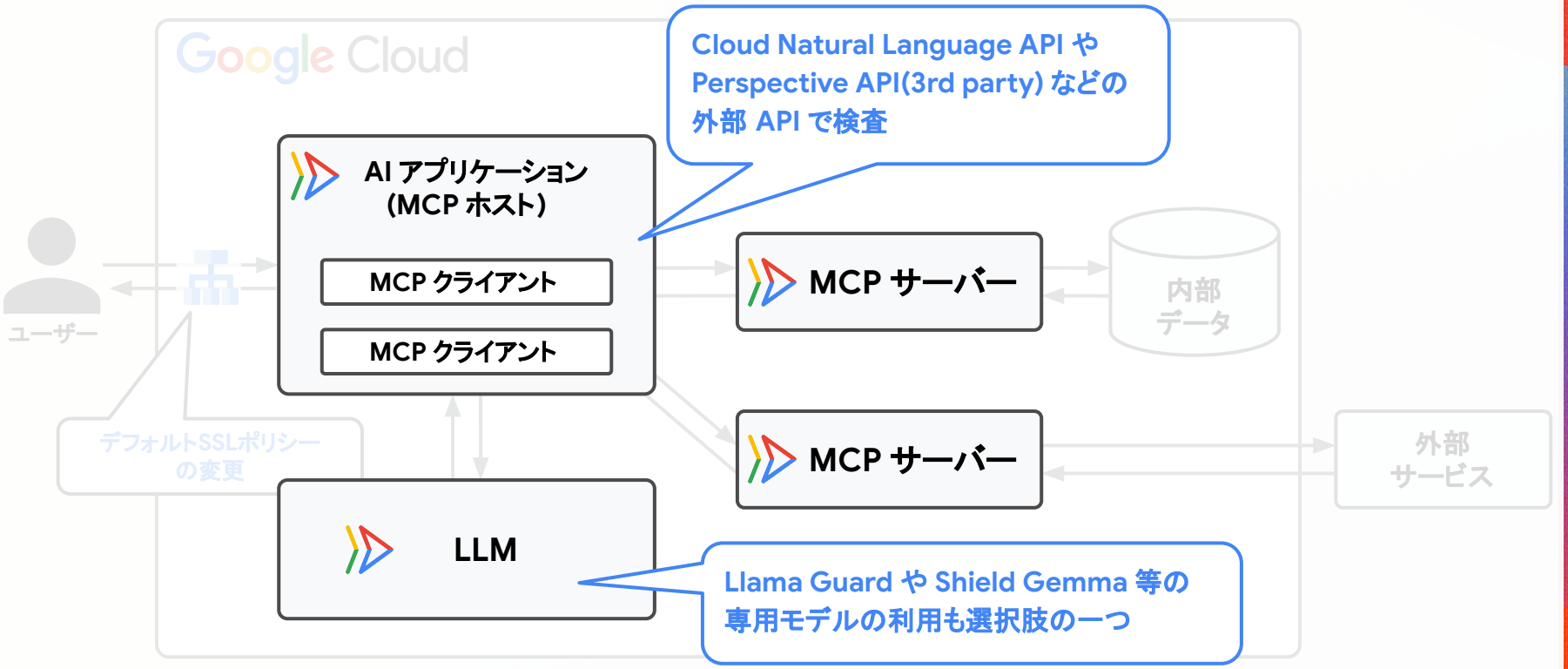
不適切な出力の
ハンドリング

異常なパターンを検出できるよう、ロギングや監視を検討

Cloud Logging に異常パターンが記録されるように留意



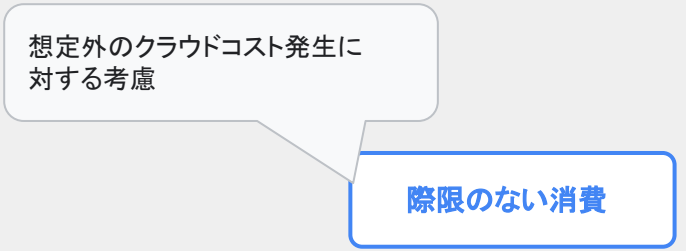
参考）不適切な入出力のブロックやフィルタリングは アプリケーション実装やモデル選定で対処



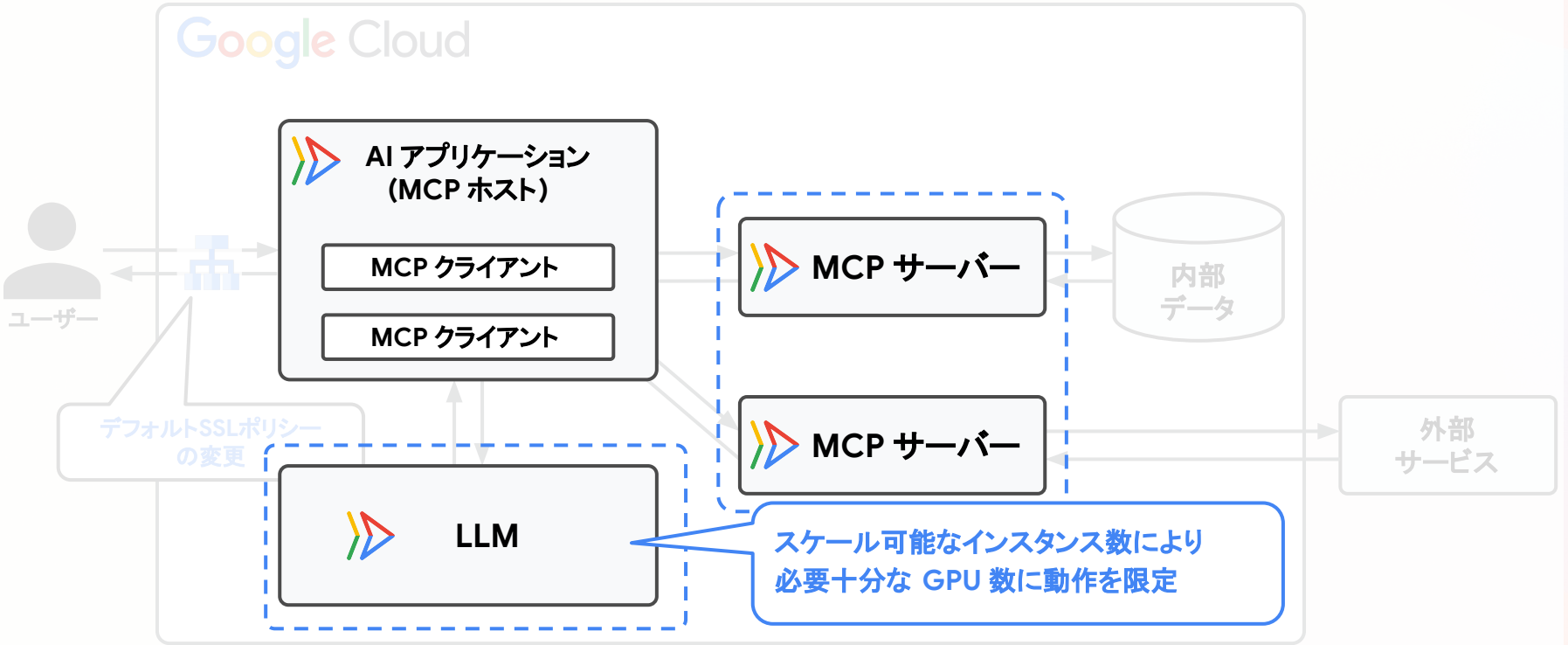
コスト消費を考慮したリソース制御に対する対策

OWASP Top 10

OWASP Top 10 for AI Applications 2025



需要に合わせた GPU のクォータとインスタンス数の制御



セキュリティ対策に関するまとめ

- IAM 認証による Cloud Run アクセス制御
- IAP を利用した認証
- External ALB におけるデフォルト SSL ポリシーの変更
- Secret Manager を利用した機密データの管理
- Direct VPC Egress 等によるパブリックアクセスの制御
- 異常パターンのロギング
- インスタンス数やクォータ制御 による想定外コスト発生の抑止

まとめ

本発表のまとめ

- LLM と MCP の概要
- 自前 LLM と MCP サーバー運用のユースケース
- LLM と MCP を Cloud Runで運用する利点とアーキテクチャ例
- OWASP Top 10 & OWASP Top 10 for AI Applications 2025
- Cloud Run 上のセキュリティ対策のポイント

ご清聴いただき、
ありがとうございました。