

# Google Cloud 生成 AI パートナー エ コシステムのご紹介

Google Cloud

パートナー事業本部

清水 岳之

# Google Cloud 生成 AI ソリューション支援パートナー

## トランスフォーメーション プログラム

戦略から導入まで支援



## 生成 AI ロケットスターター・パッケージ

生成 AI 導入・技術支援



## AI ソリューション・パッケージ

すぐに適用可能なソリューションをご提供

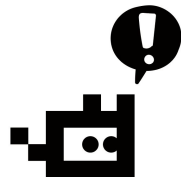


## ISV ソリューション

LLM 組込サービス開発中



# GennAI Buddy (2024 年春リリース予定)



## 知識の宝庫である社内データをもとに、 "自律的" に業務を行ってくれる相棒。

タスクの実行に必要な権限やデータを自律的に判断・収集し、事前に定義されていないタスクを実行できる。



例: 今週 Salesforce で活動予定になっているのに、アポの入っていない営業一覧をチェックリストにしてつくってください。

仕事を依頼



例: 作成した Google スプレッドシートの情報を依頼者に送信

完了結果の報告

自律的に社内システムの情報をもとに、推論をして依頼内容を遂行

例: Salesforce の情報と、Google カレンダーの情報を取得し Google スプレッドシートでチェックリストを作成

社内のあらゆるデータにLLM を介してアクセス。  
連携できるデータソースも多種多様。

社内マニュアルや規定集だけでなく、様々な動的なシステム(SFA,CRM,Slack など)を横断的に学習！静的な非公開情報だけでなく、動的な非公開情報までカバー！

### 社内システム

財務会計	管理会計	経費精算
労務	給与	勤怠
評価	採用	購買
法務	IdP	CS
SFA	CRM	販売管理
インフラ	コード管理	ナレッジ
ストレージ	ワークフロー	ヘルプデスク
グループウェア	コミュニケーション	
MDM/EMM	エンドポイントセキュリティ	

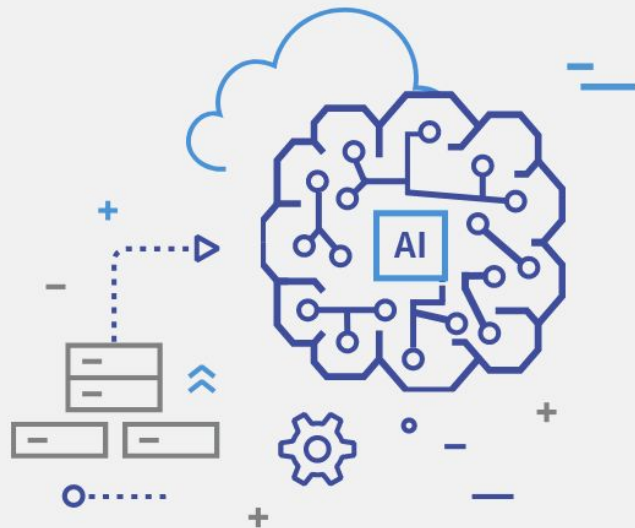
# Google Cloud 活用ソリューションのご紹介

## Google Cloud ではじめる Generative AI 活用支援ソリューション

- G-gen のエンジニアが Gen AI の活用を支援

### Google Cloud ではじめる Generative AI 活用支援ソリューション

- ① Google の **最新の生成 AI** を活用支援
- ① PoC から本番レベルの **実装まで可能**
- ① Google Cloud **有資格者** が担当



# シナリオ作成不要！ Chatbot へビルトイン型の生成 AI サービス！



## News Release



2023年11月15日  
SCSK株式会社

すぐに試せる！効果が分かる！  
「その問い合わせ、もうシナリオを作らなくても生成 AI がお答えします」  
チャットボット・ボイスボットへのビルトイン・生成 AI サービスを提供開始

SCSK株式会社(本社:東京都江東区、代表取締役 執行役員 社長:當麻 陸昭、以下 SCSK)は、チャットボット・ボイスボットに代表されるコンタクトセンターソリューションにおいて、「ビルトイン<sup>※1</sup>型・生成 AI サービス」を2023年12月1日より提供開始します。生成 AI の活用方法などで課題を抱えているお客様に対して、すぐに試して、効果が分かる環境を提供し、お客様業務の負担軽減に貢献いたします。

※1 お客様が既に利用されているサービスの一部として組み込みが可能

### 1. 背景

生成 AI の導入ニーズが高まるなか、お客様は生成 AI に対して「使い方」「活用方法」「効果」などが分からないといった共通の課題を抱えています。

SCSK は、AI を活用したチャットボットやボイスボットを提供しており、これまでヘルプデスクやコンタクトセンターにおける人手不足解消や業務自動化に貢献してまいりました。特に、ボイスボットの分野においては、Google Cloud の会話型 AI ソリューション「Contact Center AI (CCAI)」を当社のコンタクトセンターソリューションに組み込み、2023年度より提供を開始しました。

## 日本経済新聞

2023  
ご支援  
た。



### SCSK、チャットボット・ボイスボットへのビルトイン・生成 AI サービスを提供開始

企業：SCSK 産業：商社・サービス  
2023年11月15日 11:43



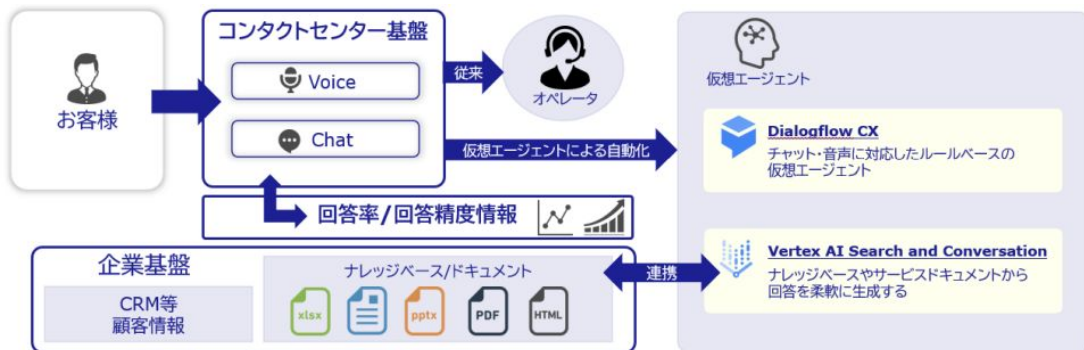
【プレスリリース】発表日:2023年11月15日

すぐに試せる！効果が分かる！



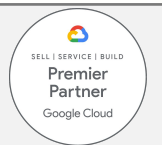
↑↑↑  
プレスリリース  
はこちら

- 最短 1 週間でご利用開始可能！  
お持ちのナレッジベース / ドキュメント読み込ませるのみ
- 新規環境用意が不要！  
APIなどを介して接続が可能
- 効果測定が可能  
回答率や回答精度を測定する BI を予めビルトイン



## お問い合わせはこちらまで

SCSK株式会社  
クラウドサービス事業本部サービス開発推進部  
→cbdc-all@scsk.jp



# マルチ生成 AI ボットから見た 各社のモデル比較

株式会社スリーシェイク

代表取締役社長

吉田 拓真

井上 秀一

# 自己紹介

株式会社スリーシェイク

代表取締役 吉田 拓真

2011年 DeNA 入社後、グループ会社のインフラを担当

AWS 日本進出時からのユーザー

2013年にソシャゲスタートアップに参画、システム統括しつつ、ゲームプロデューサー、事業戦略室室長など、ベンチャー創業期の技術・事業・経営を全般的にリード

2015年1月に株式会社スリーシェイクを設立

1987年生まれ



# 自己紹介

株式会社スリーシェイク

長期インターン生 (エンジニア)

井上 秀一

## インターン内容:

技術検証や記事執筆がメイン

LLM 関連の検証・アプリケーション開発にも従事

## 所属大学:

千葉工業大学 大学院 情報科学研究科

情報科学専攻 信川研究室 修士 2 年

エッジ系の機械学習の研究に従事

## 趣味 / 好きな技術:

サーバ開発、運用

Docker, Kubernetes, SRE





# 各生成 AI の動きと活用動向

# 各プラットフォームの生成 AI 開発状況

## Azure OpenAI

FY23~24

- コアモデルの Update
- GPT4.5 Turbo
  - Token : 128K
  - 関数呼び出し
  - 2~3 倍コスト安
- ファインチューニング対応
- カスタムモデル対応
- Gobi(マルチモーダル対応)

**GPT モデル自体の高度化 / コスト安に注力**

**生成 AI モデルにひたすら注力 (なんでもできる AIG を目指す?)**

## Google Cloud



FY23~24

- 特化型生成 AI モデルの発表
  - Sec-PaLM
  - Med-PaLM 2
  - Duet AI (Codey)
- 生成 AI 周辺の充実
  - Vertex AI Data Connectors
  - Vertex AI Pipelines
  - Vertex AI Search and Conversation
  - Vertex AI Model Garden

**PaLM 2 + 周辺系と特化型強化**

**生成 AI モデルも頑張りつつパフォーマンス (実用性) を重視?**

## AWS



FY23~24

- 生成 AI 周辺の充実
  - BedRock
    - モデル選択が自由 (Llama2 や Claude など)
    - ファイン チューニング
  - Retrieval Augmented Generation (RAG) with Knowledge Base
    - S3 ベースの RAG 実装
- 独自モデルの発表
  - Amazon Titan Multimodal Embeddings G1 (マルチモーダル)

**各生成 AI モデルと AWS の親和性強化**

**生成 AI システム実装の容易さを重視?**

# スリーシェイク社内での活用事例

## 業務効率化

### マーケティング / PR

- プレスリリースの草案作成
- インタビュー記事の校正
- バナー広告の文言草案作成

### セールス

- トークスクリプトの生成

### カスタマーサポート

- 問い合わせに対する 1 次回答の生成
- 技術問い合わせに対する 1 次切り分け時アドバイス

### コーポレート

- 与信分析
- 契約書の 1 次 Review

## エンジニアリング改善

### Development

- コード解析 ⇒ コメント自動作成 (可読性向上)
- 1 次 Review ⇒ Reviewer の負担削減
- PRD (仕様書) の作成

### SRE

- Terraform / Helm Chart コードの生成
- 作成したコードの Review
- Mermaid を利用した構成図の自動生成
- エラーコードの解析

### QA

- テストシナリオの作成
- ドキュメントの作成

### Research

- 技術論文 / 情報のサマリー作成 / Wiki 化

# スリーシェイク生成 AI 導入支援パッケージ



## 支援内容

### PaLM 2 をチャットベースで動作させるための環境構築支援

- ・ PaLM 2 を社内のあらゆる従業員が利用できるようにするための環境を整備した、必要な場合には管理機能の実装もご支援

(社内の全てのメンバーが PaLM 2 をスムーズに活用し、業務やタスクの効率を向上可能)

### Langchain ライブラリの活用による様々な AI モデルの統合

- ・ Langchain ライブラリを活用して様々な AI モデルを統合

(複数の AI 技術を組み合わせ、幅広い用途に対応した効果的な AI ソリューションの構築を実現。より高度で複雑なタスク処理が可能)

### データ基盤の整理

- ・ AI モデルの学習や適用に最適な形でのデータ処理基盤を構築。  
(RAG やファインチューニング含めた) 生成 AI モデルのフル活用を実現

## 技術的な比較

# h1-slack-bot の紹介

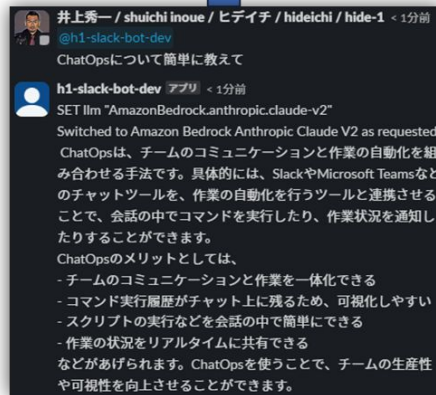
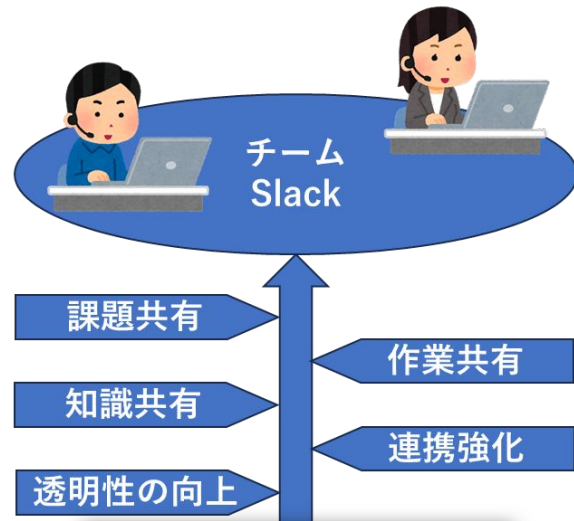
- h1-slack-bot とは
  - Slack 上で LLM と対話、比較を実現
    - OpenAI, GPT-4 Turbo, gpt-4-1106-preview
    - Google Cloud, PaLM 2 API, text-bison
    - Amazon Bedrock, Anthropic Claude 2
  - 3-shake の Slack で現在稼働中
- Slack と LLM、何が嬉しい？

## ChatOps の実現

運用・開発作業・情報検索 × チャットツール (Slack)


問題解決の迅速化、作業の一貫性確保、透明性の向上

**業務の生産性・効率を向上**




# h1-slack-bot 採用済: 各生成 AI の機能的な比較

## 高コスト・低速レスポンス 高品質レスポンス

 井上秀一 / shuichi inoue / ヒデイチ / hideichi / hide-1 22分前  
@h1-slack-bot-dev  
OpenAI GPT-4 Turbo gpt-4-1106-previewに質問です。  
ChatOpsとはなんですか？

3件の返信

 h1-slack-bot-dev アプリ 21分前  
SET IIm "OpenAI.ChatGPT"  
Switched to OpenAI GPT-4 Turbo model on your request.  
ChatOpsは、チャットツールを中心に据えて、ソフトウェア開発や運用管理の作業を行う文化や手法です。このアプローチでは、チャットボットや自動化されたスクリプトをチャット環境に統合して、開発チームがコミュニケーションを行いながら、デプロイメント、サーバーの再起動、アプリケーションの監視などのタスクを実行できるようにします。  
ChatOpsの主な利点は次のとおりです：  
- \*\*コラボレーションの向上\*\*：チームメンバーがリアルタイムでコミュニケーションを取り、作業を進めることができます。  
- \*\*透明性の向上\*\*：作業の履歴がチャットに記録されるため、誰が何をしたのかが明確になります。  
- \*\*効率の向上\*\*：繰り返し行われる作業を自動化することで、手作業によるエラーや作業時間を削減します。  
- \*\*知識の共有\*\*：チャットによる作業のログが知識のベースとなり、チーム内での情報共有が容易になります。  
ChatOpsはDevOpsの文化の一部と見なされることもあり、チームがより迅速に、より効率的に作業を行えるようにすることを目的としています。  
SlackのようなチャットツールにHubotやその他の自動化ツールを統合することで、ChatOpsは実現されます。

	OpenAI GPT-4 Turbo gpt-4-1106-preview	Google Cloud PaLM2 API text-bison	Amazon Bedrock Anthropic Claude 2
概要	高コスト低速レスポンス ハイクオリティなchat用 チューニングモデル	低コスト 1対1の会話にチューニング された汎用モデル	高コスト レスポンス測度やや早め コンテキスト上限高
コスト	入力 \$0.01 / 1K tokens 出力 \$0.03 / 1K tokens	入力 \$0.0010 / 1K charctors 出力 \$0.0010 / 1K charctors ※ 1 token = 4 charctors 入力 \$0.004 / 1K token 出力 \$0.004 / 1K token	入力 \$0.01102 / 1K tokens 出力 \$0.03268 / 1K tokens
制限	128k context	入力 8192 tokens 出力 1024 tokens	100k
トレーニング データ	2023年4月まで	2023年2月まで	-
	<a href="https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo">https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo</a>	<a href="https://cloud.google.com/vertex-ai/docs/generative-ai/learn/models?hl=ja">https://cloud.google.com/vertex-ai/docs/generative-ai/learn/models?hl=ja</a>	<a href="https://aws.amazon.com/jp/bedrock/pricing/">https://aws.amazon.com/jp/bedrock/pricing/</a>

# h1-slack-bot 採用済: 各生成 AI の機能的な比較

シンプルな応答  
やや早いレスポンス

最も低コスト  
他二種よりも約 60% 安価

文字数課金制  
コスト予測が簡単

	OpenAI GPT-4 Turbo gpt-4-1106-preview	Google Cloud PaLM2 API text-bison	Amazon Bedrock Anthropic Claude 2
概要	高コスト低速レスポンス ハイクオリティなchat用 チューニングモデル	低コスト 1対1の会話にチューニング された汎用モデル	高コスト レスポンス測度やや早め コンテキスト上限高
コスト	入力 \$0.01 / 1K tokens 出力 \$0.03 / 1K tokens	入力 \$0.0010 / 1K charctors 出力 \$0.0010 / 1K charctors ※ 1 token = 4 charctors 入力 \$0.004 / 1K token 出力 \$0.004 / 1K token	入力 \$0.01102 / 1K tokens 出力 \$0.03268 / 1K tokens
制限	128k context	入力 8192 tokens 出力 1024 tokens	100k
トレーニング データ	2023年4月まで	2023年2月まで	-
	<a href="https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo">https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo</a>	<a href="https://cloud.google.com/vertex-ai/docs/generative-ai/learn/models?hl=ja">https://cloud.google.com/vertex-ai/docs/generative-ai/learn/models?hl=ja</a>	<a href="https://aws.amazon.com/jp/bedrock/pricing/">https://aws.amazon.com/jp/bedrock/pricing/</a>





## Tokenについての補足:

テキストを、スペース、句読点、特殊文字を  
特定のルールに従って、より小さなセグメントに分離

## 以下はtokenizerによるtokenの算出:

GPT-3.5 & GPT-4 GPT-3 (Legacy)

こんにちは！  
私はインターン生です。

Clear Show example


Tokens Characters  
12 19

こんにちは！  
私はインターン生です。


Google Cloud PaLM2 API text-bison	Amazon Bedrock Anthropic Claude 2
低コスト 1対1の会話にチューニング された汎用モデル	高コスト レスポンス測度やや早め コンテキスト上限高
入力 \$0.0010 / 1K charctors 出力 \$0.0010 / 1K charctors ※ 1 token = 4 charctors 入力 \$0.004 / 1K token 出力 \$0.004 / 1K token	入力 \$0.01102 / 1K tokens 出力 \$0.03268 / 1K tokens
入力 8192 tokens 出力 1024 tokens	100k
2023年2月まで	-
<a href="https://cloud.google.com/vertex-ai/docs/generative-ai/learn/models?hl=ja">https://cloud.google.com/vertex-ai/docs/generative-ai/learn/models?hl=ja</a>	<a href="https://aws.amazon.com/jp/bedrock/pricing/">https://aws.amazon.com/jp/bedrock/pricing/</a>

# h1-slack-bot 採用済 : 各生成 AI の機能的な比較

高品質なレスポンス  
 やや早いレスポンス  
 高コスト

 井上秀一 / shuichi inoue / ヒデイチ / hideichi / hide-1 24分前  
 @h1-slack-bot-dev  
 AmazonBedrock.anthropic.claude-v2に質問です。  
 ChatOpsとはなんですか？

3件の返信

 h1-slack-bot-dev アプリ 24分前  
 SET llm "AmazonBedrock.anthropic.claude-v2"  
 AmazonBedrock.anthropic.claude-v2に切り替えました。  
 ChatOpsは、チームのコラボレーションを促進するためにチャットツールと自動化ツールを統合する運用手法です。  
 ChatOpsでは、チャットツール上から自動化ツールを直接操作できます。例えば、Slackからシェルスクリプトを実行したり、Kubernetesのデプロイをしたりできます。  
 ChatOpsの利点は以下の通りです:  
 - チームのコミュニケーションと作業の自動化を1つのインターフェースで行える  
 - 作業状況の可視化が容易  
 - 手動作業と自動化作業の境界がなくなる  
 - 迅速なフィードバックループを構築できる  
 ChatOpsはDevOps文化を推進する上で重要な手法の1つです。チームの生産性とコラボレーションを高めるのに有効なアプローチだと言えます。

	OpenAI GPT-4 Turbo gpt-4-1106-preview	Google Cloud PaLM2 API text-bison	Amazon Bedrock Anthropic Claude 2
概要	高コスト低速レスポンス ハイクオリティなchat用 チューニングモデル	低コスト 1対1の会話にチューニング された汎用モデル	高コスト レスポンス測定やや早め コンテキスト上限高
コスト	入力 \$0.01 / 1K tokens 出力 \$0.03 / 1K tokens	入力 \$0.0010 / 1K charctors 出力 \$0.0010 / 1K charctors ※ 1 token = 4 charctors 入力 \$0.004 / 1K token 出力 \$0.004 / 1K token	入力 \$0.01102 / 1K tokens 出力 \$0.03268 / 1K tokens
制限	128k context	入力 8192 tokens 出力 1024 tokens	100k
トレーニング データ	2023年4月まで	2023年2月まで	-
	<a href="https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo">https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo</a>	<a href="https://cloud.google.com/vertex-ai/docs/generative-ai/learn/models?hl=ja">https://cloud.google.com/vertex-ai/docs/generative-ai/learn/models?hl=ja</a>	<a href="https://aws.amazon.com/jp/bedrock/pricing/">https://aws.amazon.com/jp/bedrock/pricing/</a>

# h1-slack-bot の搭載機能 : 概要

●スレッド毎に異なるLLMを使用可能

●LLMの切り替えもサポート

“PaLMに切り替えて”と発するだけで切り替え可能

●動的要約機能

Token上限を回避するには？

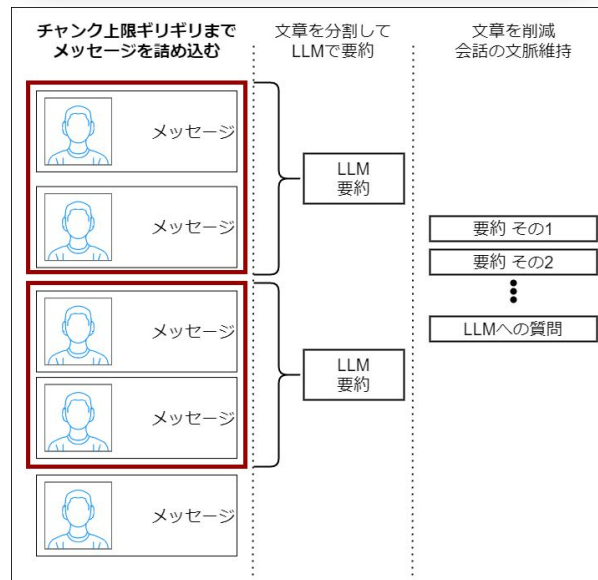
× ただの再帰的要約

○ ユーザが知りたい情報に集中して再帰的要約

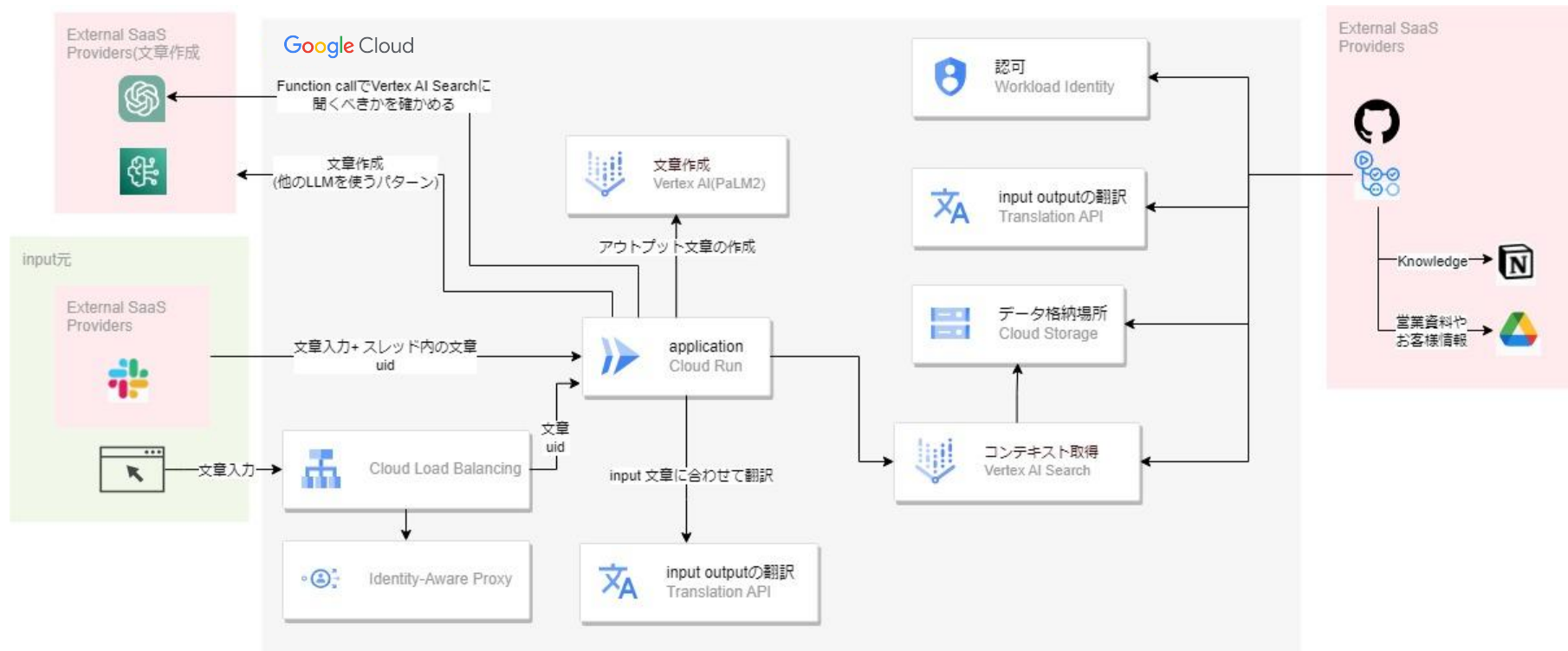
●RAG (Retrieval-Augmented Generation)

• Vertex AI Search and Conversationを利用

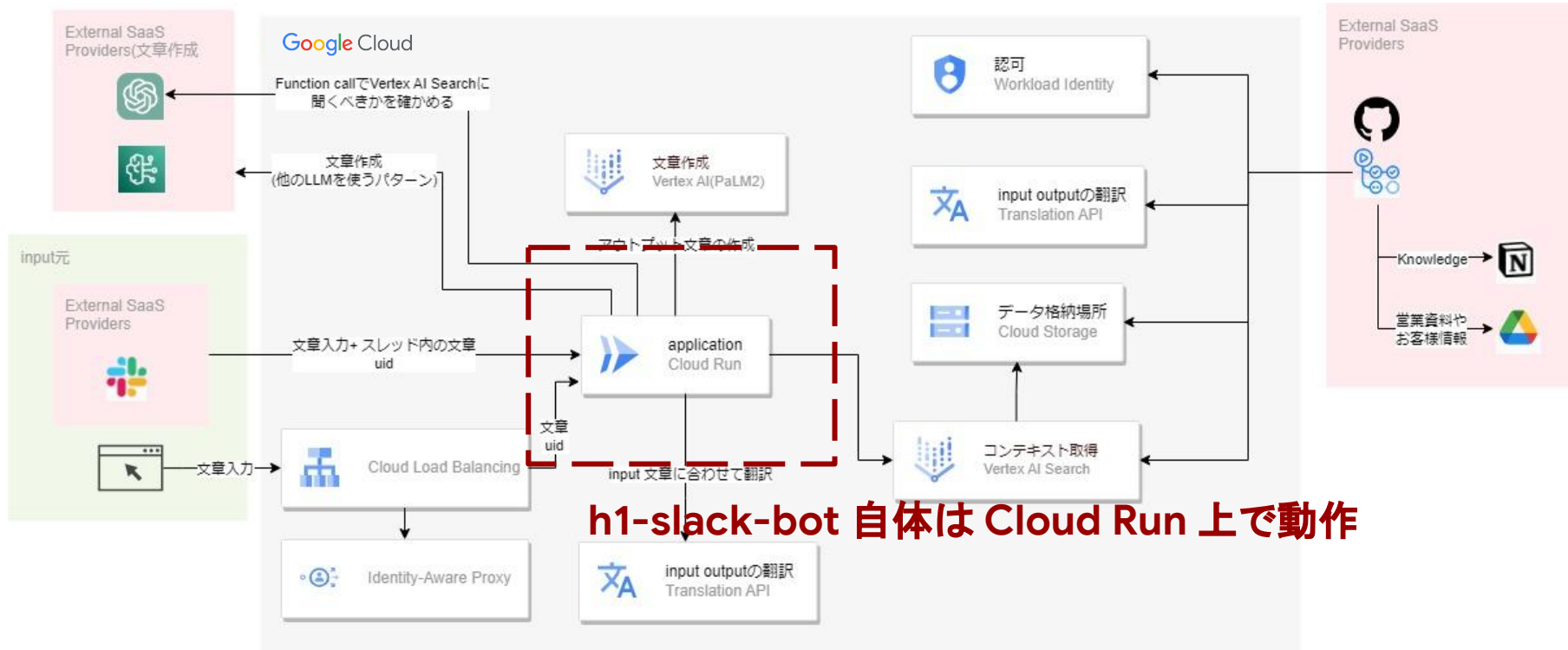
• 生成AIの誤回答を大幅軽減



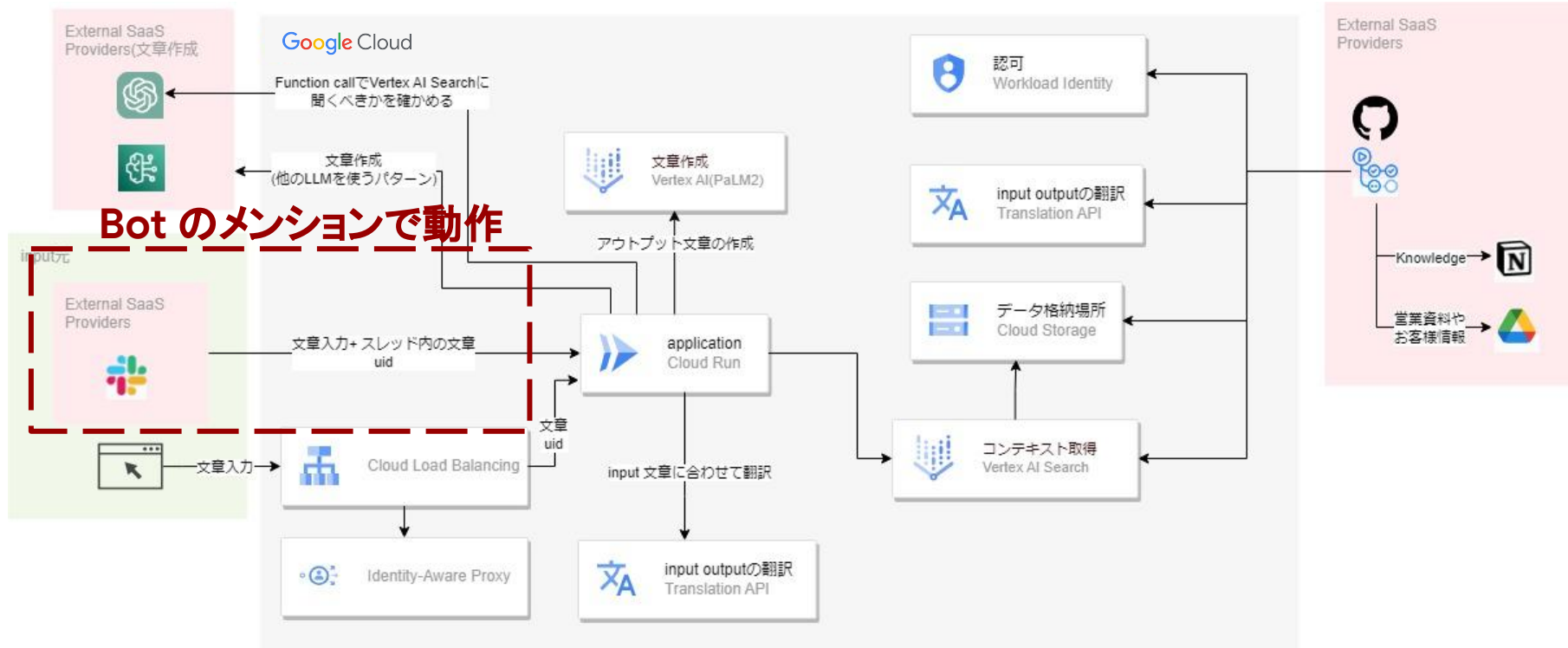
# h1-slack-bot のアーキテクチャ紹介



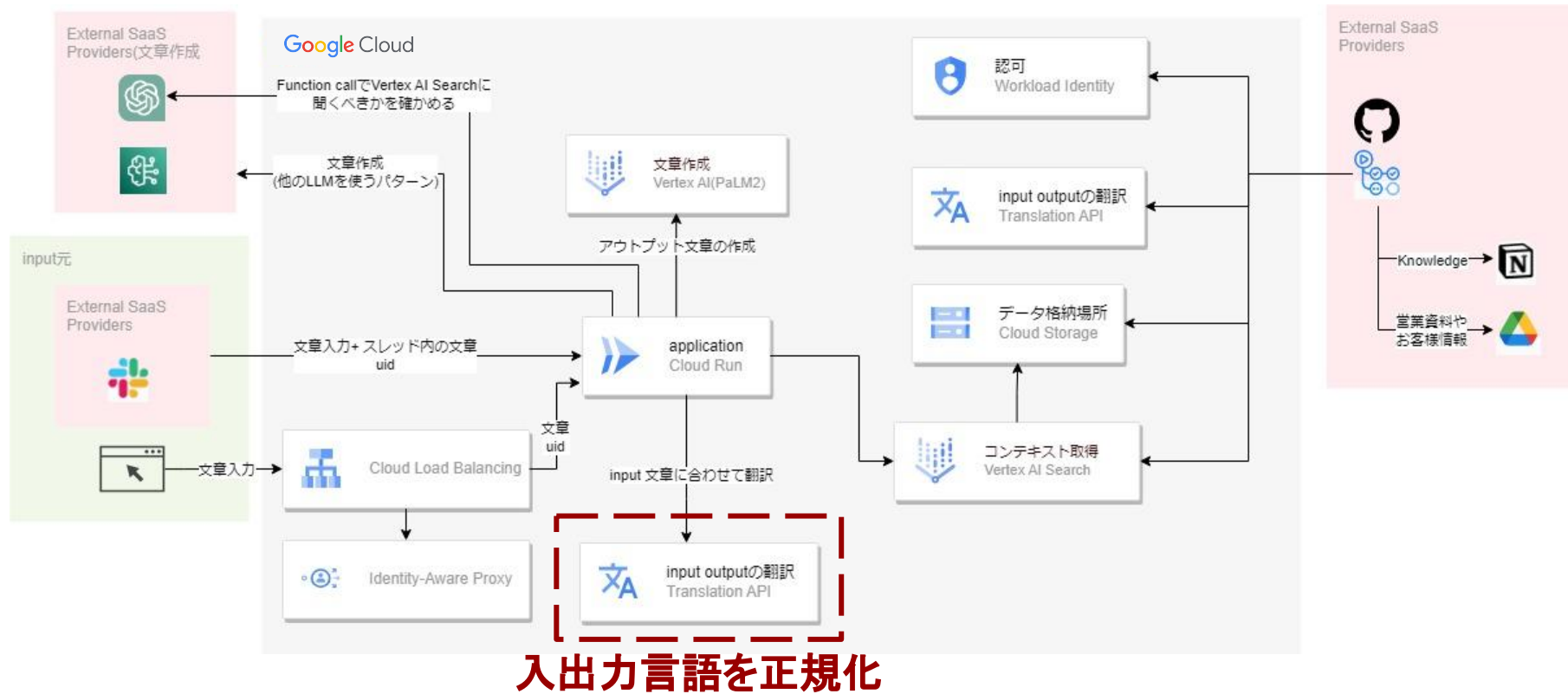
# h1-slack-bot のアーキテクチャ紹介



# h1-slack-bot のアーキテクチャ紹介

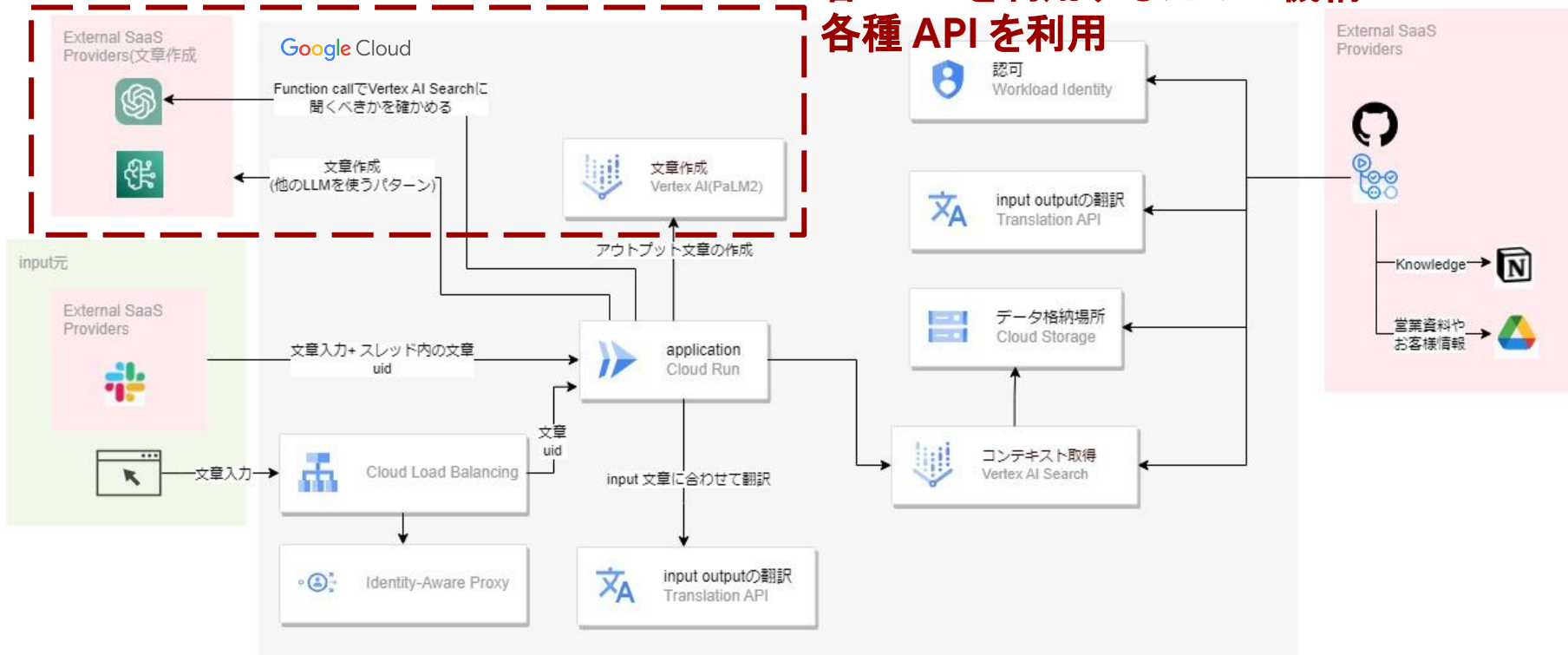


# h1-slack-bot のアーキテクチャ紹介



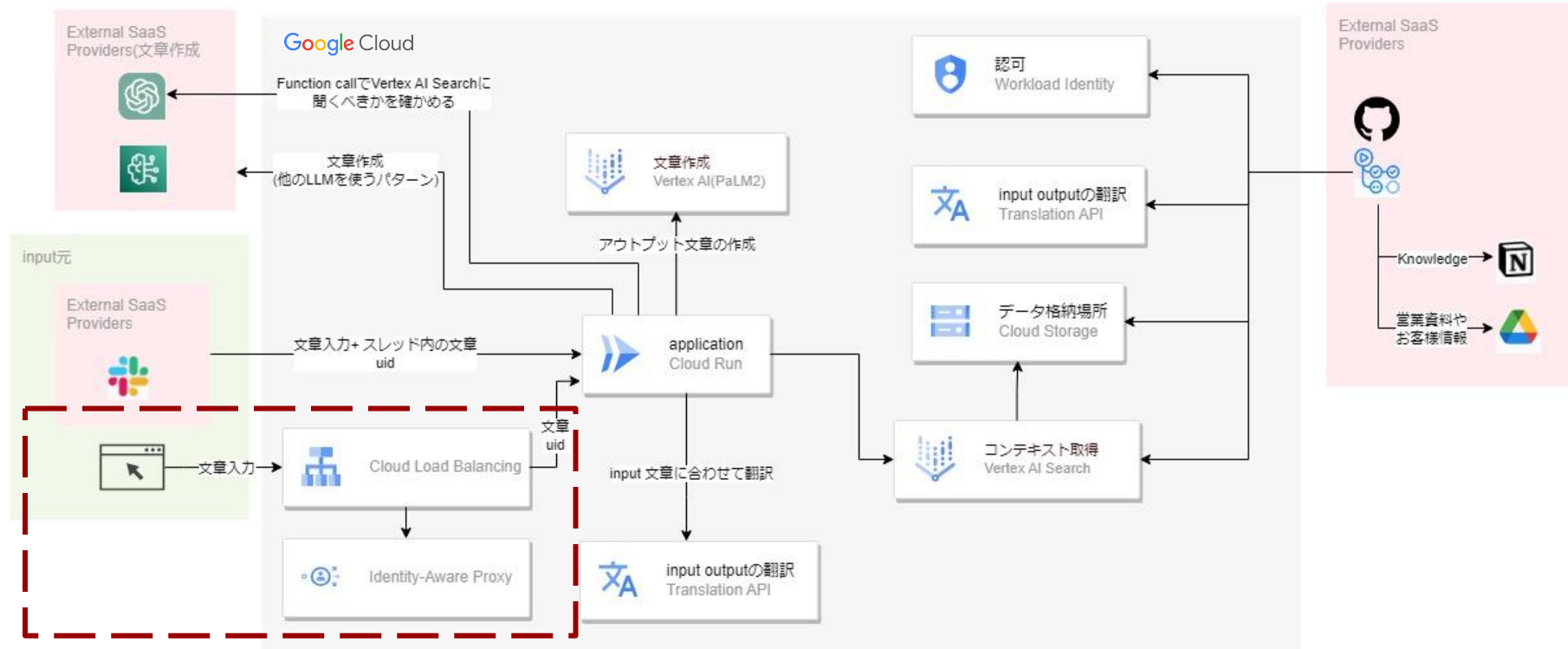
# h1-slack-bot のアーキテクチャ紹介

## 各 LLM を利用するための機構 各種 API を利用





# h1-slack-bot のアーキテクチャ紹介



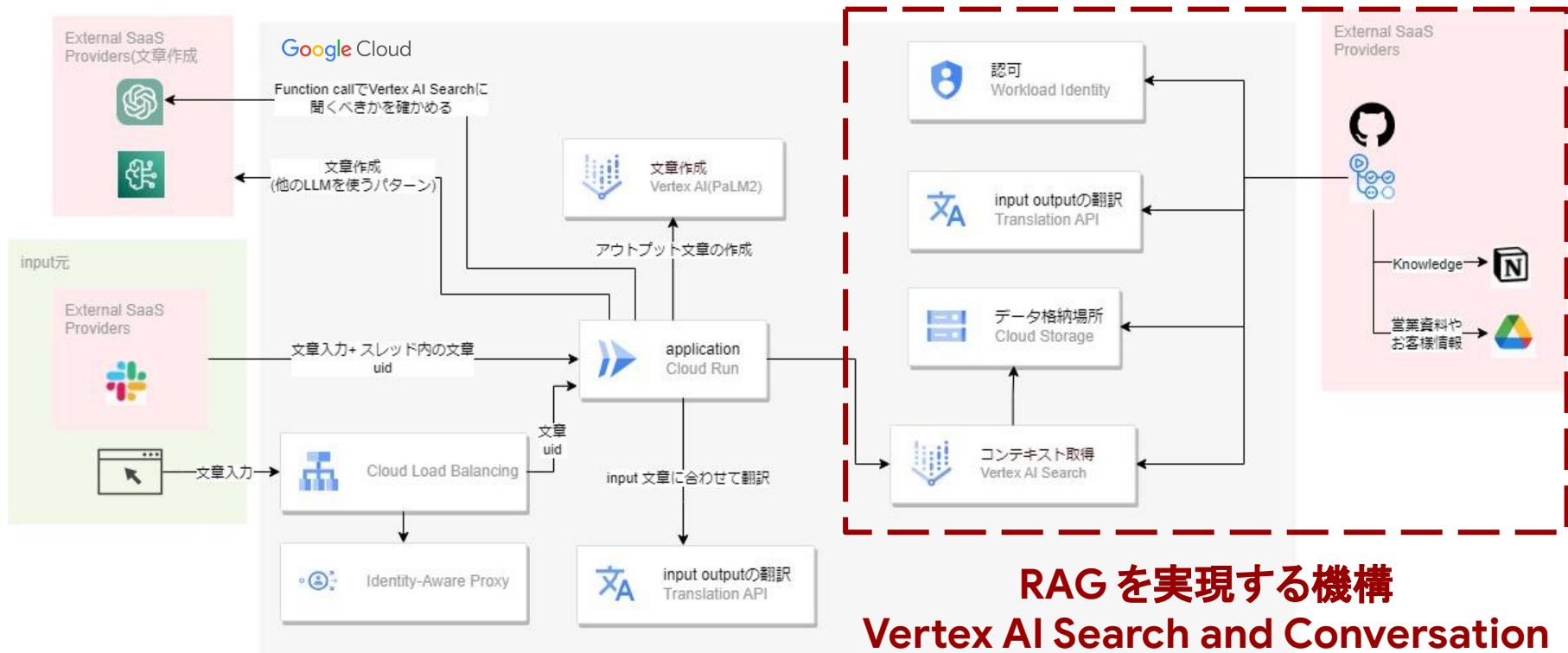
専用のチャット画面 (鋭意製作中)

# h1-slack-bot のアーキテクチャ紹介



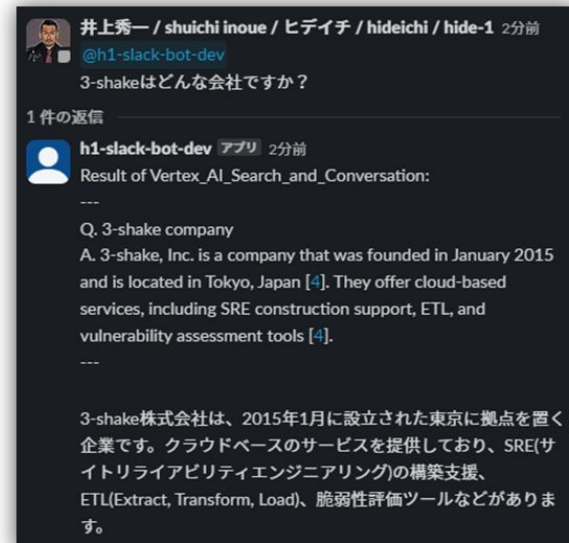
専用のチャット画面 (鋭意製作中)

# h1-slack-bot のアーキテクチャ紹介



# h1-slack-bot の搭載機能 : Vertex AI Search and Conversationの紹介

- Vertex AI Search and Conversation とは？
  - 生成 AI を使用した検索と会話構築
  - RAG (Retrieval-Augmented Generation) のキー
  - 社内情報やナレッジを利用可能
- H1-slack-bot における Vertex AI Search and Conversation:





**Thank you.**