

Agent Development Kit による商品検索 AI エージェント開発

Google
Cloud
Next

Tokyo

Proprietary



佐藤一憲

Google Cloud

Cloud AI

デベロッパー アドボケイト



@kazunori279



Dave Elliott

**Google Cloud
Developer Advocacy &
Engineering Manager, AI**



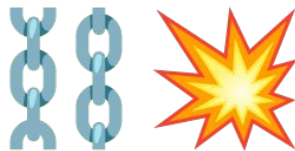
/in/davidlelliott



01. Vertex AI の エージェント開発スタック

AI エージェントの開発

その強さにとまなう難しさ



開発の見通しの悪さ

より速く開発するには？

エージェント間連携 の難しさ

依存関係や機能移譲を効果的に管理でき、エージェント間のワークフローをデバッグしやすくするには？

運用管理の難しさ

開発したエージェントのエコシステムを信頼性に優れた方法でデプロイ、スケール、監視、管理するには？

AI エージェント開発に必要なもの

ツールキット

標準化された効率的な開発の実現

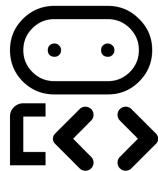
プロトコル

LLM、ツール、エージェント間を結ぶ

プラットフォーム

エージェントのデプロイと運用

Google のエージェント開発スタック



Agent Development Kit (ADK)

- オープンソースのツールキット
- 簡単に使える
- 柔軟性高く、すぐにデプロイできる



MCP

- Anthropic 開発のオープン プロトコル
- LLM にコンテキストを供給



Agent Engine

- Google Cloud のマネージド プラットフォーム
- エージェント システムを実運用向けにスケール



A2A

- Google 開発のオープン標準
- エージェント間を結ぶ

スタック #1: Agent Development Kit (ADK)



Agent Development Kit



Proprietary

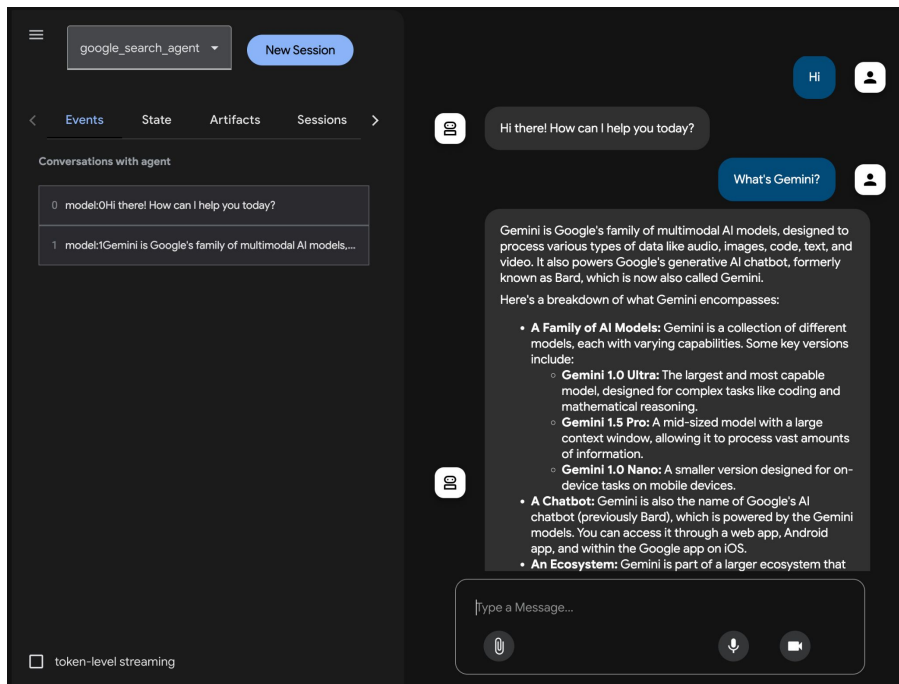
Google 開発のマルチエージェント フ
レームワーク

生成 AI 推論 によるアクション

Gemini に加え、LiteLLM によりサード
パーティ モデルをサポート

音声、動画、画像の双方向ライブ スト
リーミングをサポート

adk web tool



ADK のコアコンセプト



LlmAgent (Agent): エージェントの頭脳

- ✨ LLM を利用 (Gemini はじめ Claude などの他社モデル)
- ➡ 主要パラメータを設定 (エージェント名、モデル、プロンプト、ツールなど)

Tools (FunctionTool): エージェントが実行する機能

- エージェントにアクション実行の能力 (Python 関数) を提供
- 必要なスキーマ生成は ADK が担当

ADK のコアコンセプト



Runner: エージェントを実行 (*ADK run*, *ADK web*)

- エージェントを実行
- セッション管理(ユーザとの対話内容を保持)

Session とステート

-  Session = ユーザとの対話
-  ステート = セッションが保持する Dict オブジェクト

スタック #2: Model Context Protocol (MCP)



MCP とは

AI エージェントや AI アプリケーションと、外部ツールやデータソースをシームレスに結ぶオープン プロトコル

API

Web アプリケーション とバックエンドの連携方法を標準化

- サーバ
- データベース
- サービス

LSP

IDE(開発環境)と言語固有ツールの連携方法を標準化

- コードの閲覧
- コードの分析
- コードの理解

MCP

AI アプリケーション と外部システムの連携方法を標準化

- プロンプト
- ツール
- リソース

ADK と MCP



既存の MCP サーバを ADK で利用

- ADK エージェントが MCP クライアントとして動作



ADK ツールを MCP サーバ経由で公開

- MCP クライアントから ADK ツールを利用

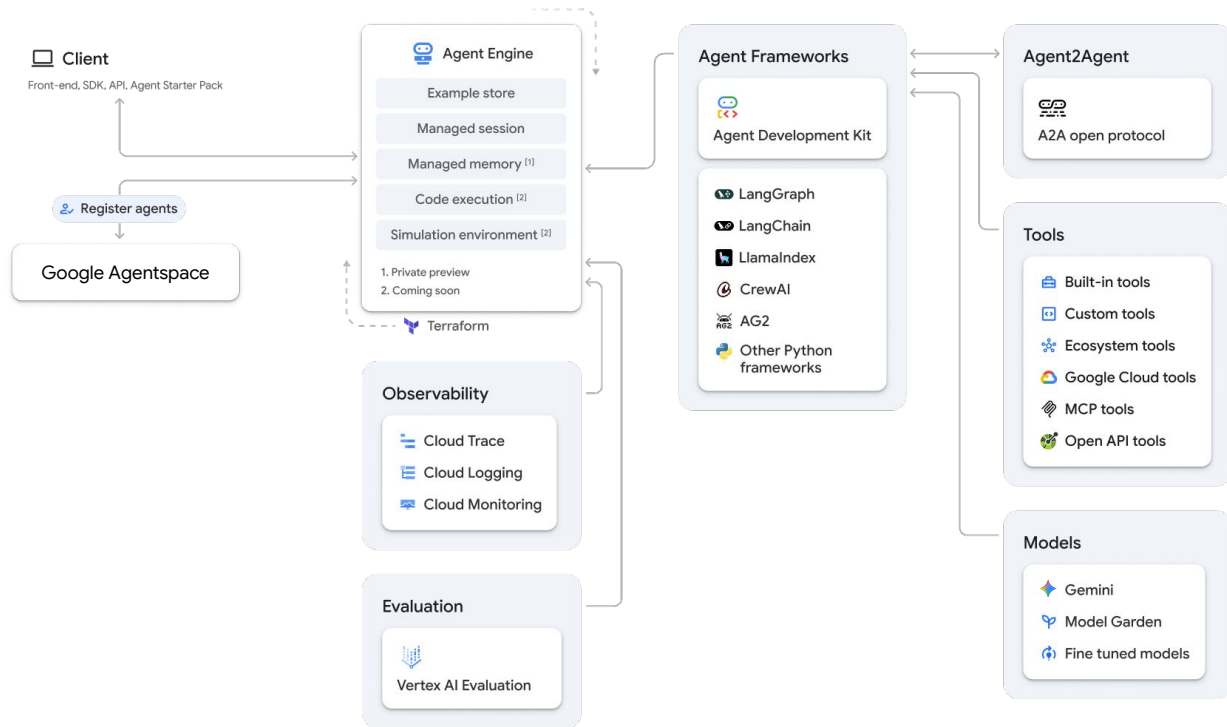
スタック #3: Vertex AI Agent Engine



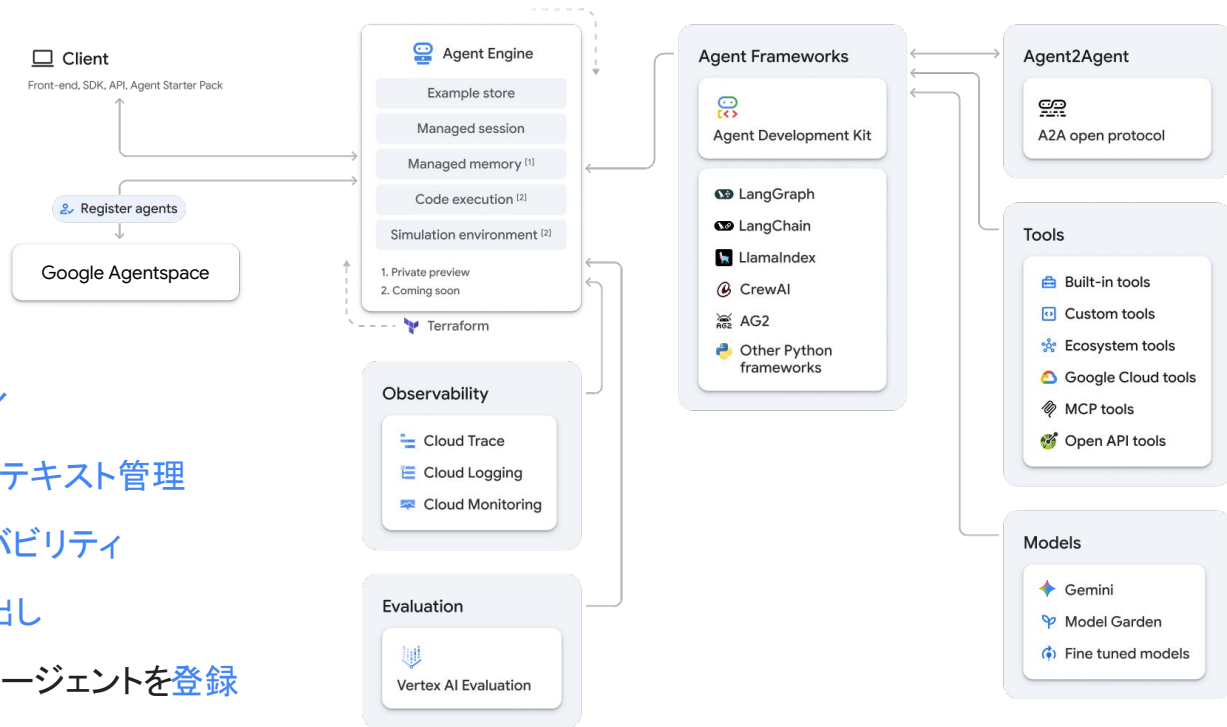
なぜ Agent Engine か

	Agent Engine なし	Agent Engine あり
コード開発	• FastAPI や Django 等でアプリケーション サーバを開発し、HTTP API でアプリ (LangChain や ADK) を提供 • Docker の知識が必要。コードからコンテナ イメージを作成する。 • Google Cloud 環境の設定	• Vertex AI SDK から <code>agent_engines.create()</code> を呼び出すだけ
インフラ運用	• ユーザのコードはユーザのプロジェクトにデプロイ • 全ての運用作業はユーザの責任に	• ユーザのコードをテナント プロジェクトにデプロイ • 自動スケーリング、リージョン拡張、コンテナ セキュリティ等を Vertex AI が管理
オブザーバビリティと評価	• ログ管理、監視ツール、トレース、評価等を手作業で設定	• オブザーバビリティ: Cloud Logging, Monitoring, および Tracing を利用 • Vertex AI Extensions をツールとして SDK から利用 • Vertex AI Gen AI Evaluation と統合

Vertex AI Agent Engine



Vertex AI Agent Engine



フレームワークでエージェントを**定義**

自動設定されたランタイムに**デプロイ**

セッション、メモリ、サンプル等の**コンテキスト管理**

エージェントのふるまいの**オブザーバビリティ**

アプリや UI からエージェントを**呼び出し**

Google AgentSpace にカスタムエージェントを**登録**

Stack #4: Agent to Agent (A2A)



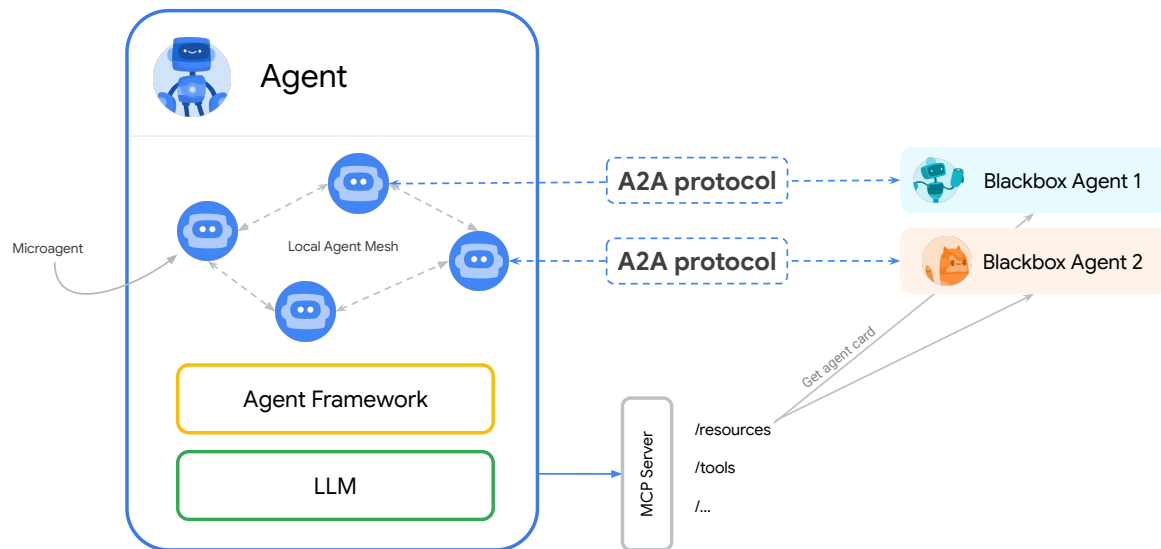
A2A (Agent2Agent)

エージェント間連携のためのオープンプロトコル

Agent Card でエージェントを発見

エンタープライズ対応のセキュアで統制の取れた環境

HTTP, SSE, JSON-RPC 等の標準技術で構成



A2A のコアコンセプト

Agent Skill (エージェント スキル)

- エージェントが備える能力や機能を記述
- エージェントがどのようなタスクに向いているかをクライアントに伝える

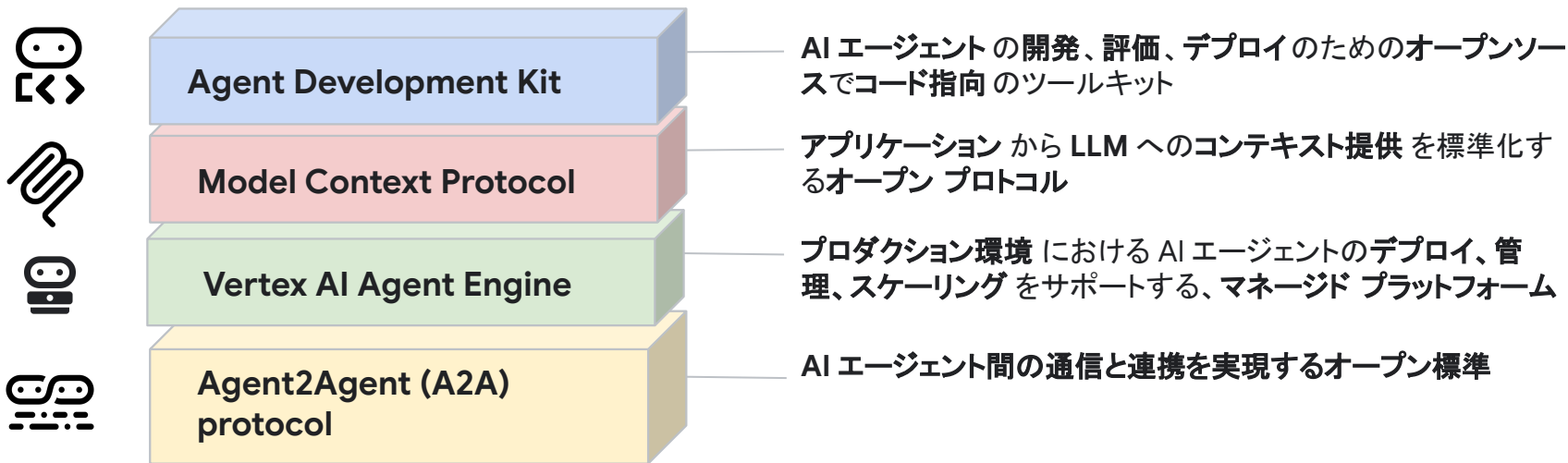
Agent Card (エージェント カード)

- JSON で記述されたエージェントの仕様
 - `.well-known/agent.json` に記述
- エージェントの「名刺」のようなもの

Vertex AI のエージェント開発スタック: まとめ



エージェント開発スタック



Agent Development Kit (ADK)



Github repo



Read the Docs



Sample Agents



goo.gle/adk-foundation
goo.gle/adk-using-tools
goo.gle/adk-mcp-a2a



/in/davidlelliott

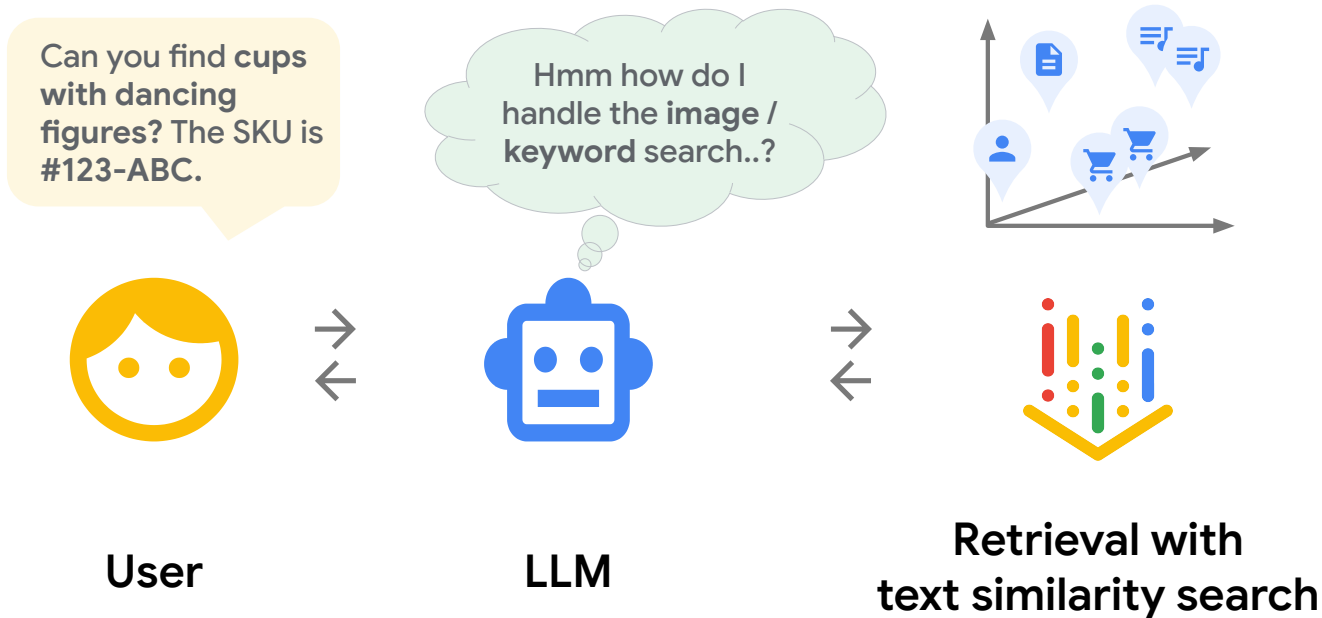


02. エージェント型 RAG のための ベクトル検索ベスト プラクティス

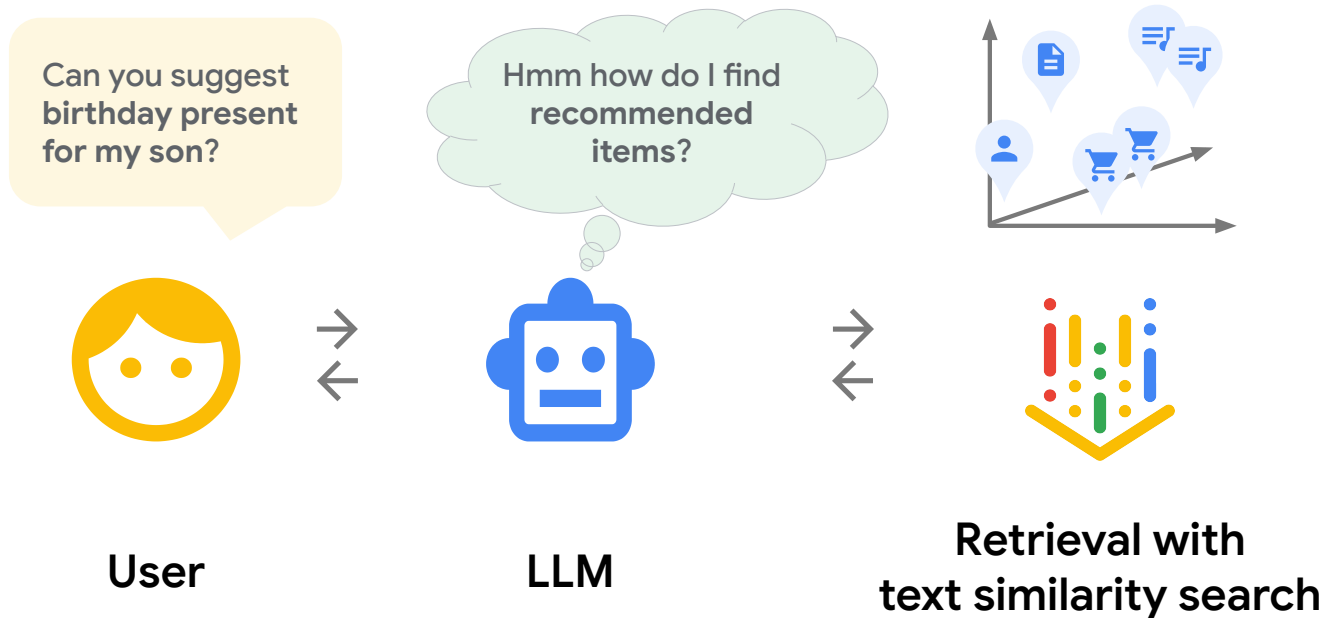
A typical RAG scenario for e-commerce



Challenge: Multimodal / Keyword search



Challenge: Recommendations



What makes search quality better



Multimodal search

Use images, audio and video semantics



Hybrid search

Use both semantic and keyword search



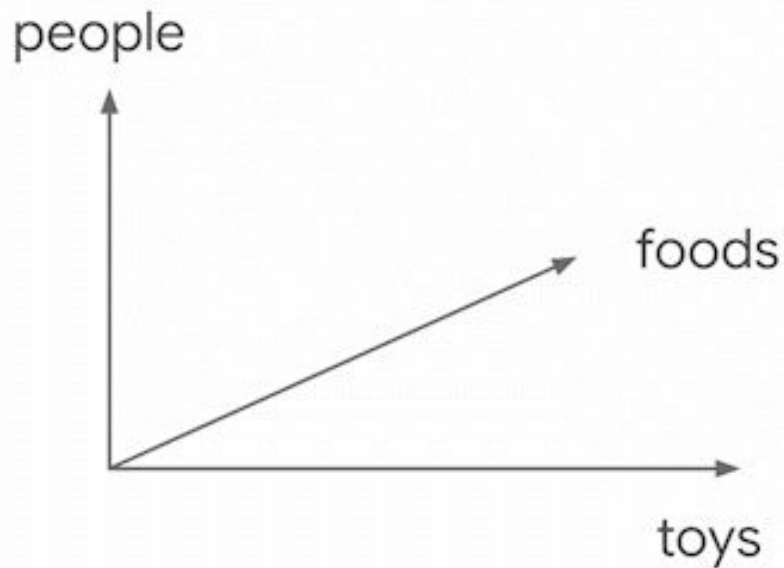
Task type embedding

To find by relevance, not by similarity

Multimodal Search

Using an embedding model that understands text and image

Proprietary



Hybrid Search

The weakness of the semantic search

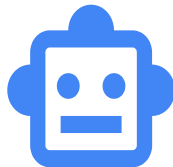
Can you find
item# 123-4567?
ASDFGHJ?



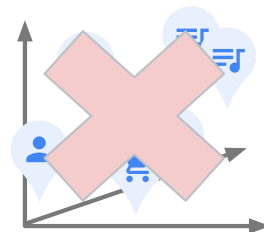
User



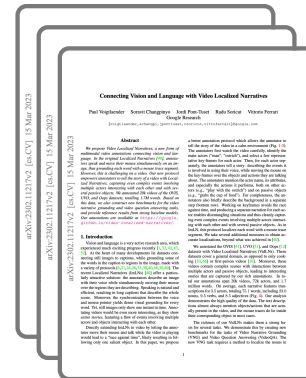
Hmm... what do they
mean???



LLM



Vector Search



Contents

Hybrid Search

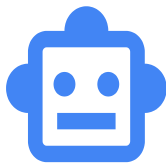
Combining the strength of semantic and keyword search

Can you find items matching "Kids"?

Here's the results with the Hybrid Search!



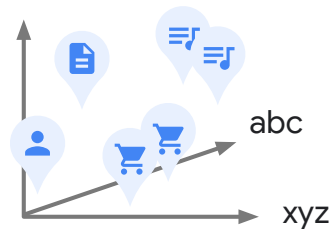
User



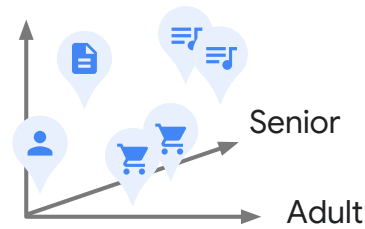
LLM



123



Youth



Google Blue Kids Sunglasses
 Google Red Kids Sunglasses
 YouTube Kids Coloring Pencils
 YouTube Kids Character Sticker Sheet

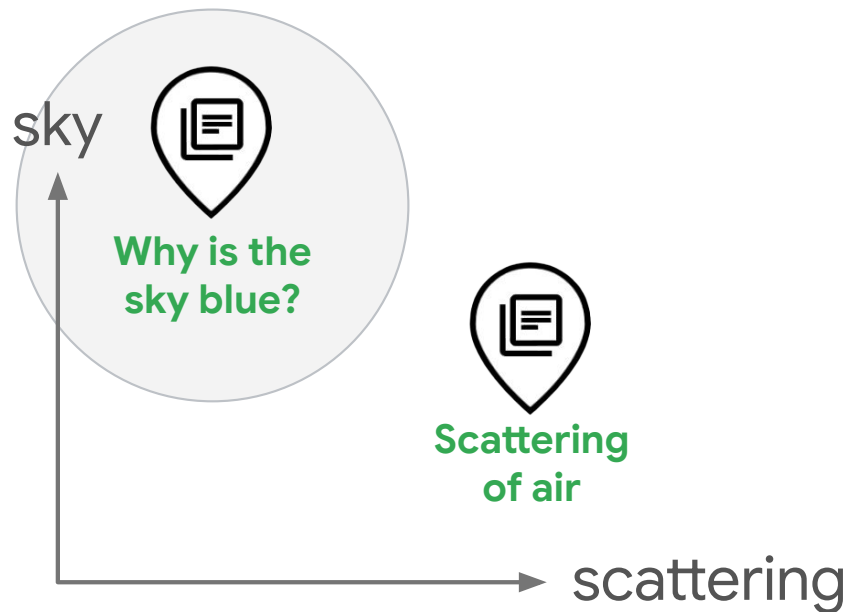
Google White Classic Youth Tee
 Google Doogler Youth Tee
 Google Indigo Youth Tee
 Google Black Classic Youth Tee
 Chrome Dino Glow-in-the-Dark Youth Tee
 Google Bike Youth Tee

Demo

Multimodal search and Hybrid Search

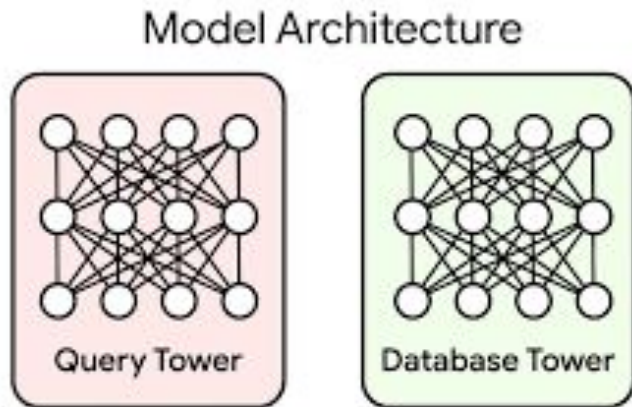
Task type embedding

Questions and their answers are not semantically similar



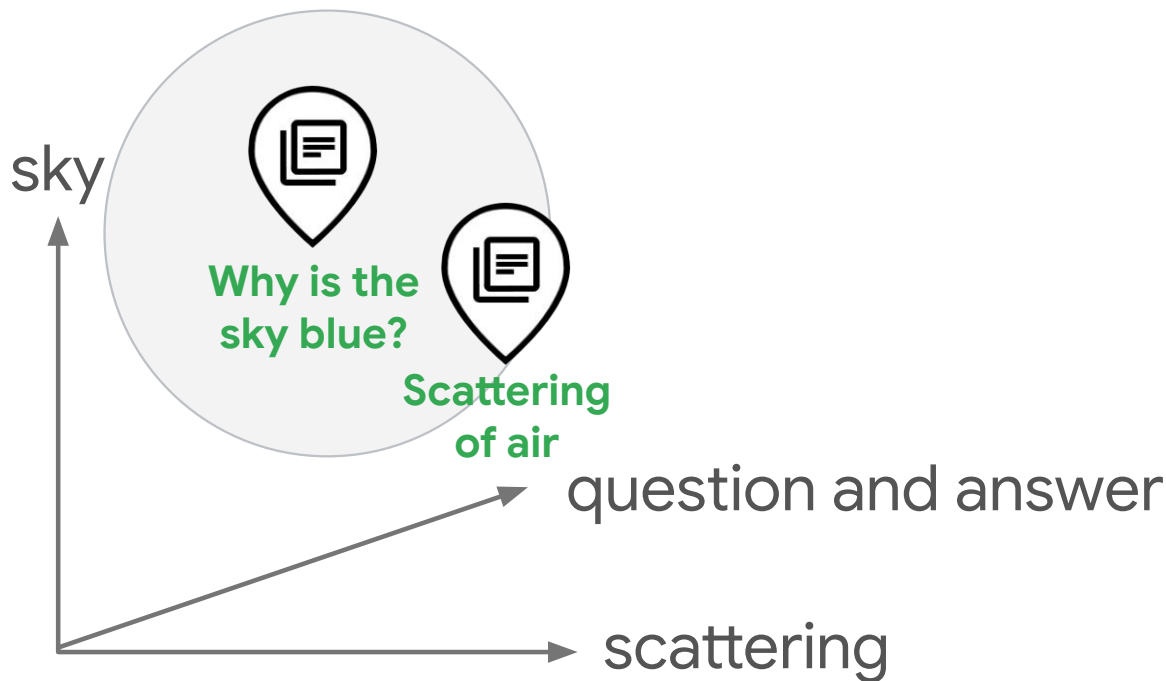
Task type embedding

Learns the relationship between questions and answers



Task type embedding

Maps the question and answer closer in the embedding space

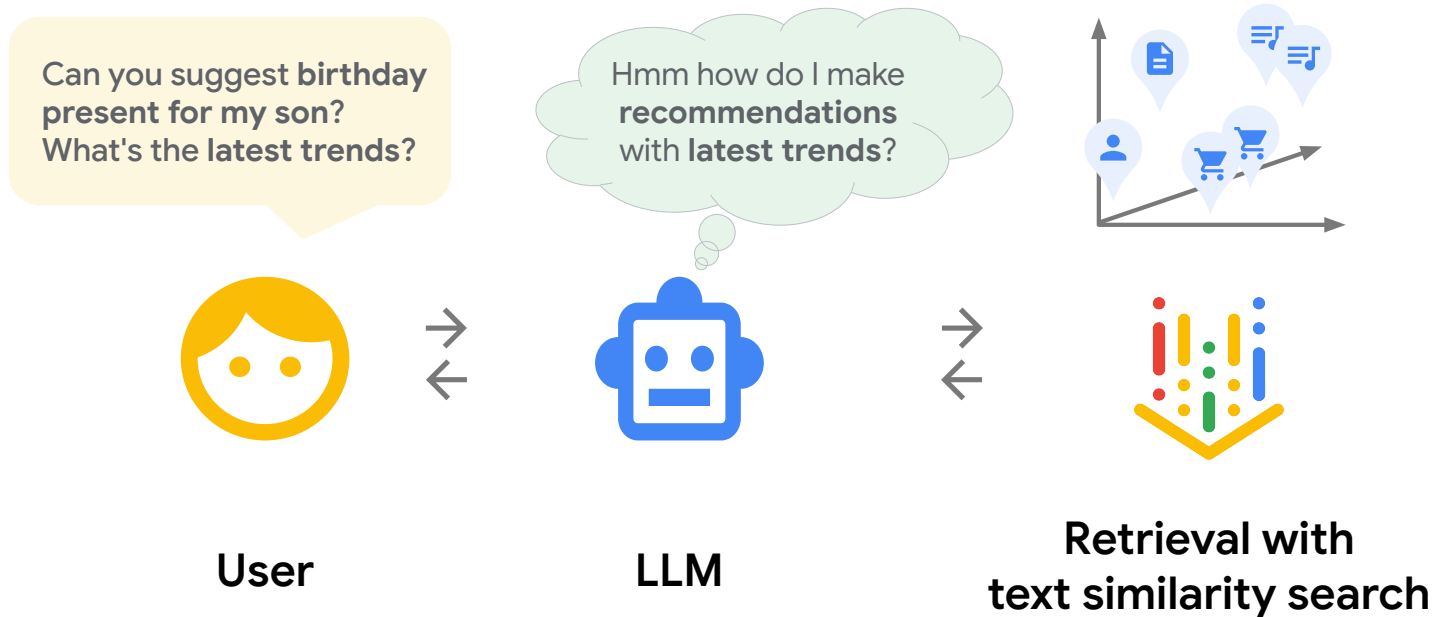


Demo

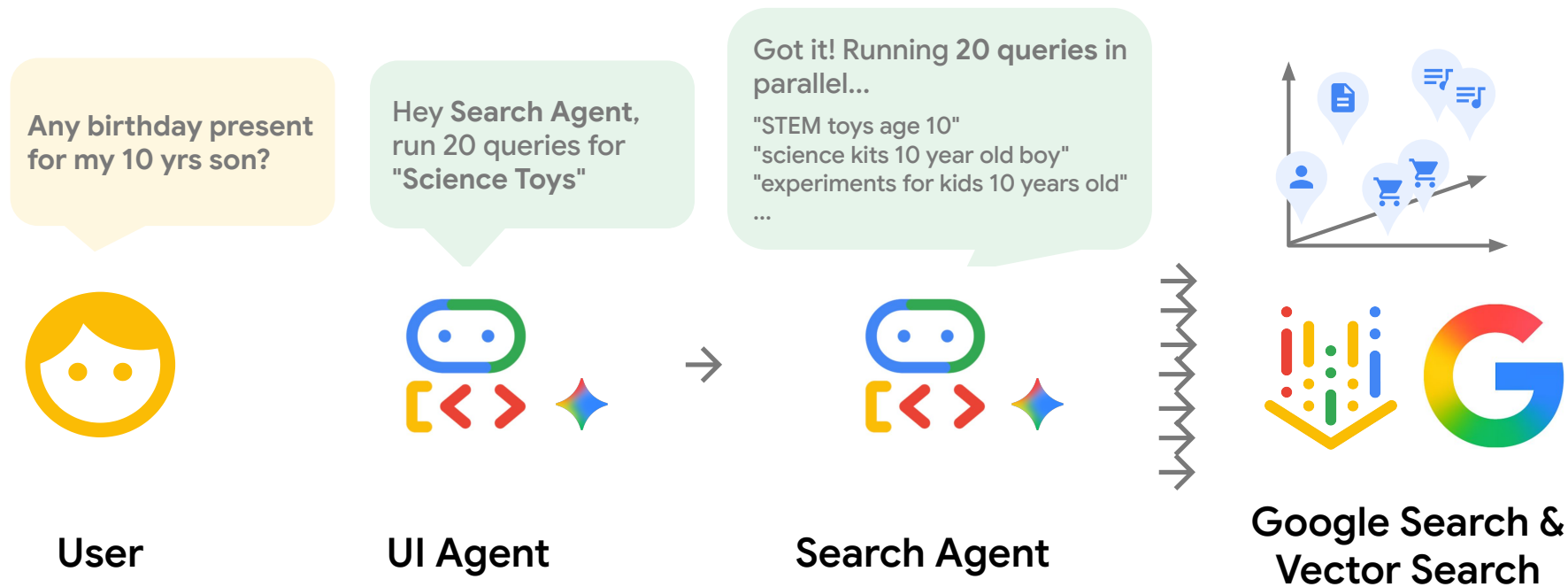
Task type embeddings

03. Vertex AI Vector Search によるエージェント型 RAG

Challenge: *Smart* Recommendations



Solution: AI Agents + Vector Search



Demo

Shopper's Concierge

Shopper's Concierge

Research with Google Search to find 5 item categories
(~500 ms)

Any birthday present
for my 10 yrs son?



User



Hey Google Search,
what do people buy for
birthday presents?



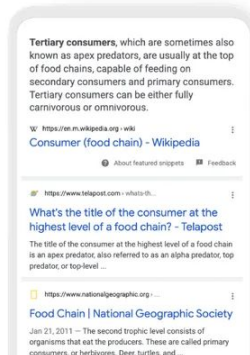
UI Agent



Sport goods,
Outdoor toys,
Science Toys

...

what's the title of the consumer at the highest level of a food chain



Google Search grounding

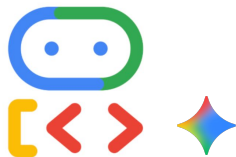
Shopper's Concierge

Generate and run 20 queries for each category
(From 3 million items to a few hundreds, in ~500 ms)



User

Hey Search Agent,
run 20 queries for
"Science Toys"



UI Agent



Got it! Running 20 queries in
parallel...

"STEM toys age 10"
"science kits 10 year old boy"
"experiments for kids 10 years old"
...



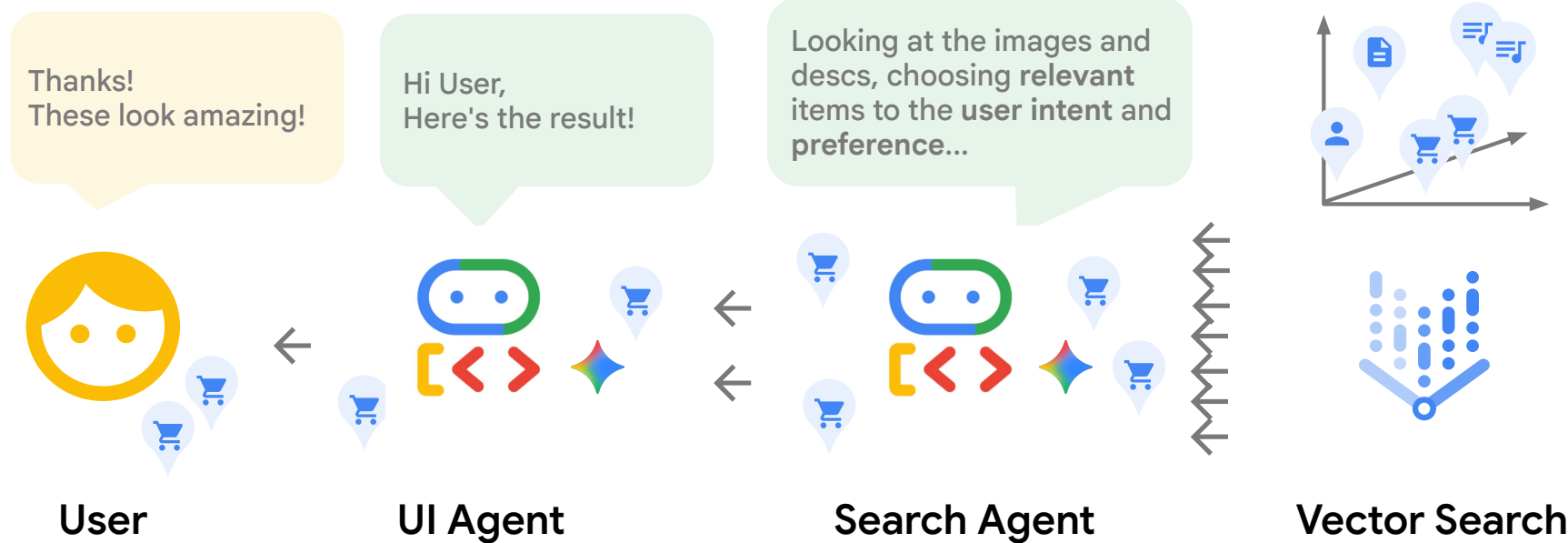
Search Agent



Vector Search

Shopper's Concierge

Personalized multimodal item curation for the user
(From a few hundreds to tens, in ~4 sec)



Multimodal item curation with Gemini 2.0 Flash

User intent: "decorating a living room", searching for "Wall Art & Mirrors"



Shopper's Concierge

Multimodal, multi-agent RAG
with Generative Recommendations



Multimodal,
Multi-agent search
with ADK

Real-time multimodal
communication



Generative
Recommendations

Google Search grounding
Query generation
Multimodal item curation



Vector Search

Multimodal search
Hybrid search
Task-type embeddings
Ranking API

Get Started



Vector Search
docs



ADK recsys
sample



@kazunori279

