Building Agentic RAG with ADK + Vector Search

Google Cloud Next Tokyo

Kaz Sato

Google Cloud
Cloud AI
Developer Advocate

@kazunori279

01. Agent dev stack on Google Cloud

To build an agentic app, you'd need...

Toolkit

for standardized & efficient development

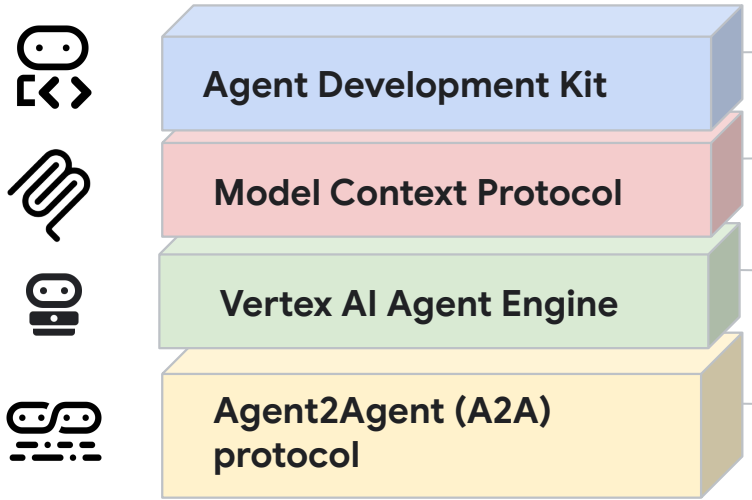
Protocols

for connecting LLMs, tools & agents

Agent platform

for managing deployment and operations

Agent dev stack on Google Cloud



Open-source, code-first toolkit for **building**, **evaluating**, and **deploying** AI **agents**.

Open protocol that standardizes how **applications provide context to LLMs.**

Managed platform to deploy, manage, and scale AI agents in production.

Open standard designed to enable **communication & collaboration between AI agents.**

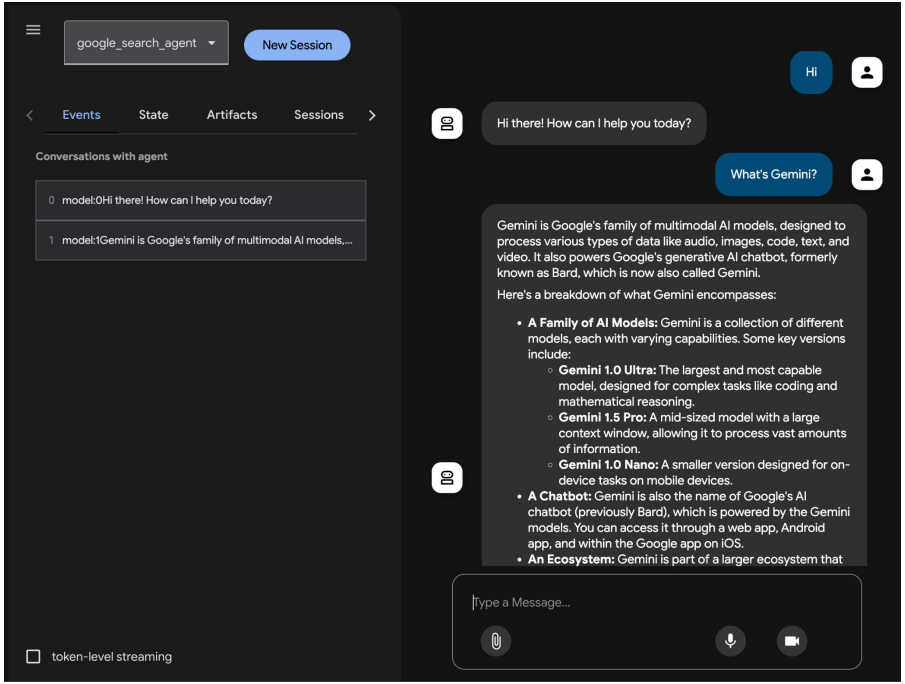
Agent Development Kit (ADK)

Multi-agent framework by Google

Actions driven by **gen AI reasoning**

Supports **Gemini** and **third party models** with LiteLLM

Live audio, video, screen bi-directional streaming



adk web tool

MCP: Model Context Protocol

MCP is an open protocol that enables seamless integration between AI apps & agents and your tools & data sources.

APIs

Standardize how **web applications** interact with the **backend**:

- Servers
- Databases
- Services

LSP

Standardizes how **IDEs** interact with **language-specific tools**:

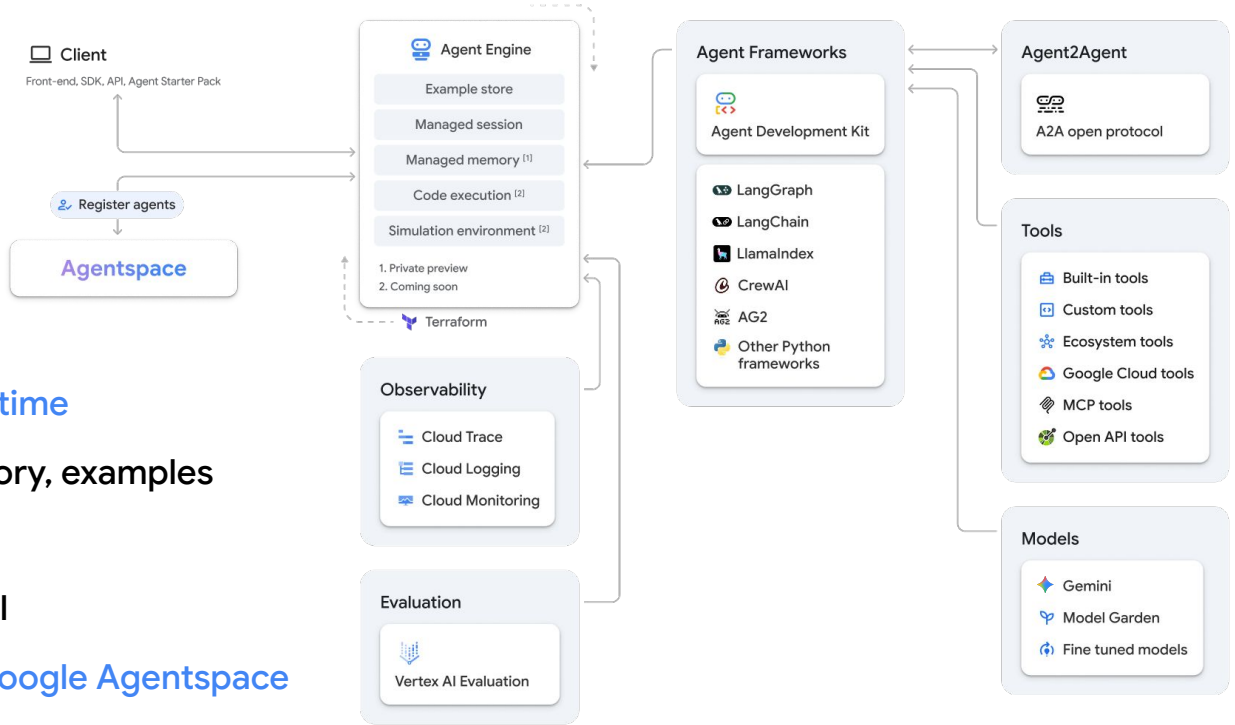
- Code navigation
- Code analysis
- Code intelligence

MCP

Standardizes how **AI applications** interact with **external systems**:

- Prompts
- Tools
- Resources

Vertex AI Agent Engine



Define an agent in a framework

Deploy to a fully provisioned runtime

Manage context: sessions, memory, examples

Observe agent behavior

Call your agent from your App/UI

Register your custom agent in Google Agentspace

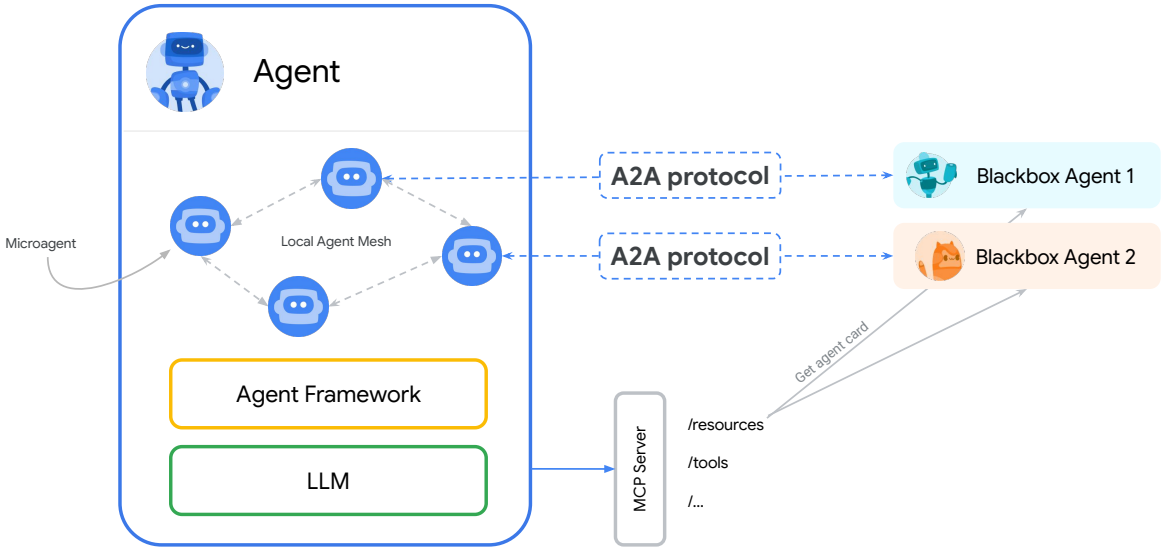
A2A (Agent2Agent)

Open protocol to handle agent collaboration

Enables agent discovery using Agent Card

Enterprise-ready for a secure and governed experience.

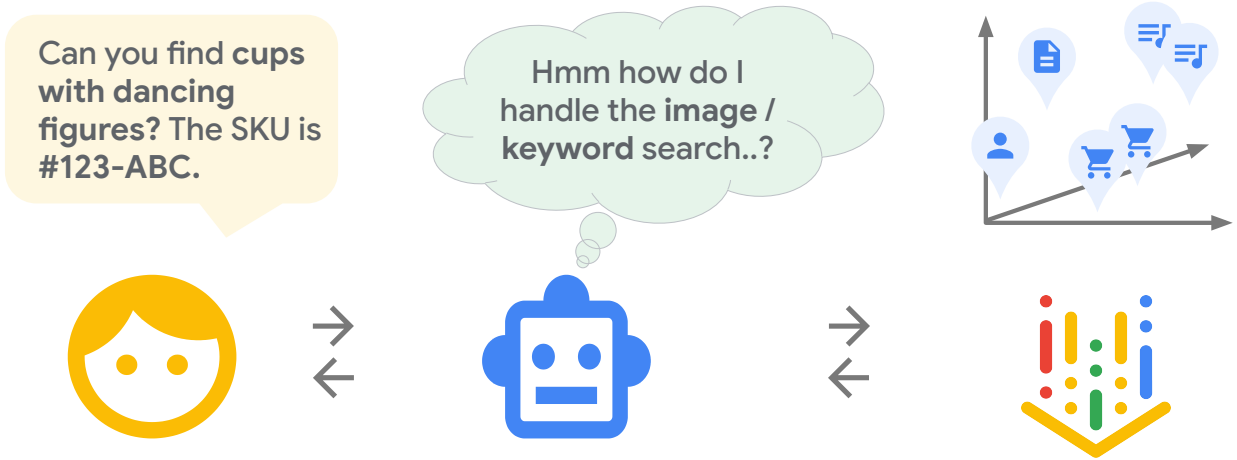
Built on top of existing, popular standards including HTTP, SSE, JSON-RPC



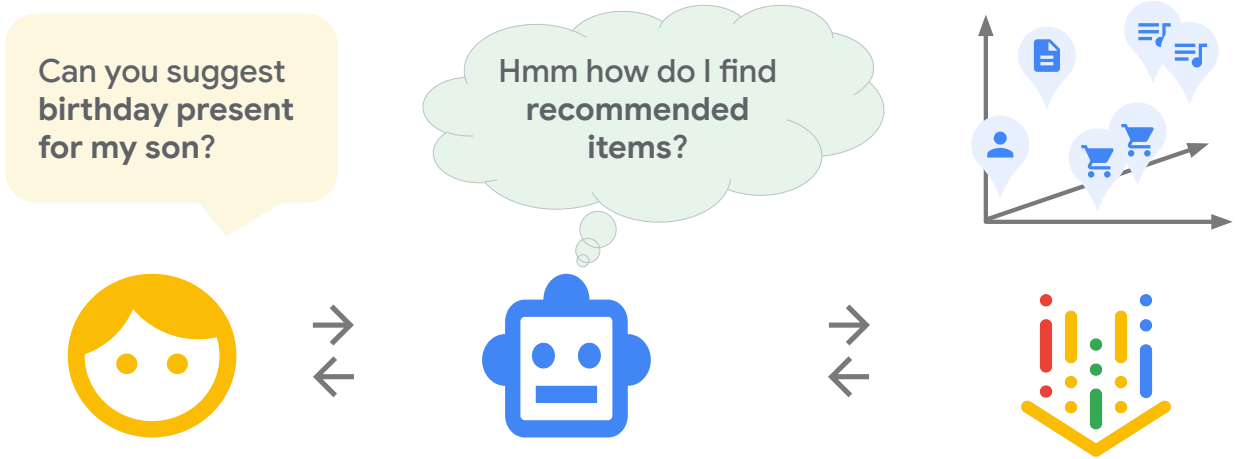
02. Vector Search advanced practices for Agentic RAG

A typical RAG scenario for e-commerce

Challenge: Multimodal / Keyword search



Challenge: Recommendations



User

LLM

Retrieval with text similarity search

What makes search quality better

Multimodal search

Use **images, audio and video** semantics

Hybrid search

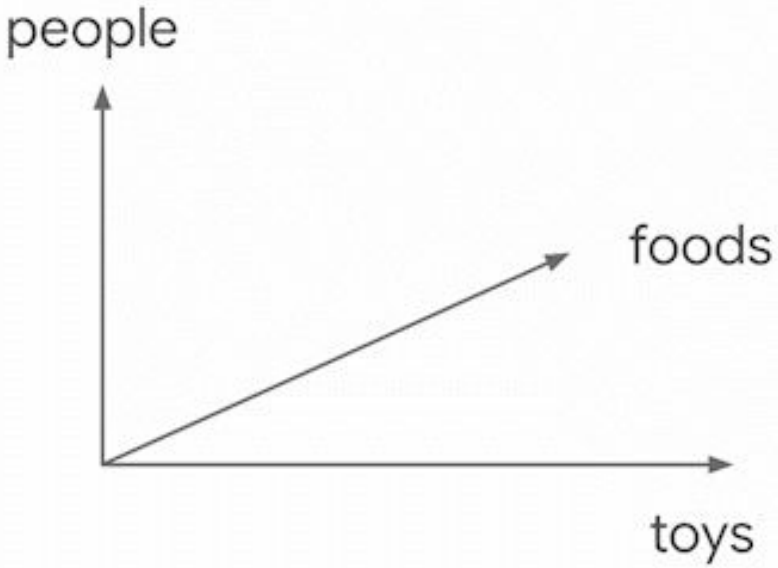
Use both **semantic** and **keyword search**

Task type embedding

To find by **relevance**, not by similarity

Multimodal Search

Using an embedding model that understands text and image

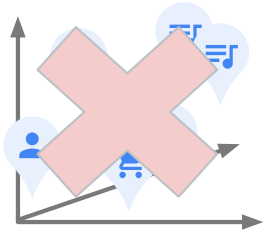


Hybrid Search

The weakness of the semantic search

Can you find **item# 123-4567? ASDFGHJ?**

Hmm... what do they mean???



User → **LLM** → **Vector Search** → **Contents**

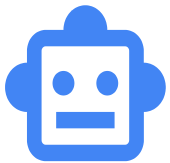
Hybrid Search

Combining the strength of semantic and keyword search

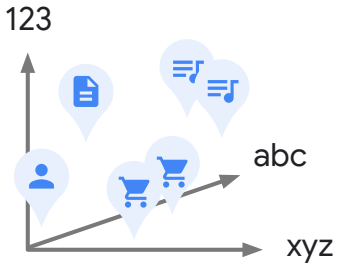
Can you find items matching **"Kids"**?

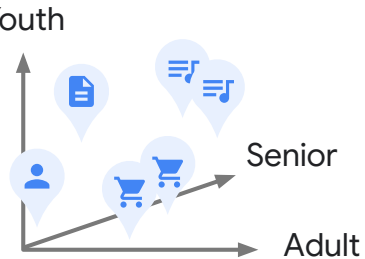
Here's the results with the **Hybrid Search!**

User



LLM






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

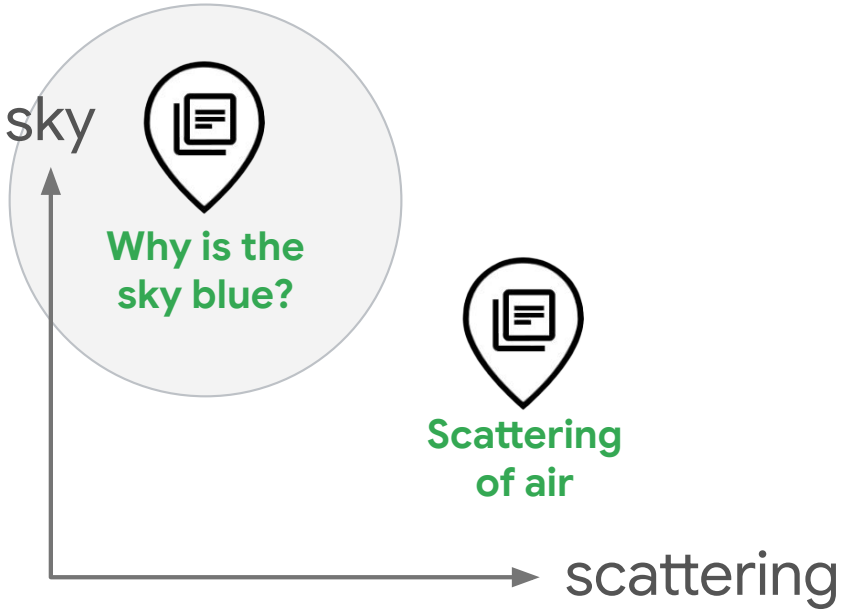
```
Google White Classic Youth Tee
Google Doogler Youth Tee
Google Indigo Youth Tee
Google Black Classic Youth Tee
Chrome Dino Glow-in-the-Dark Youth Tee
Google Bike Youth Tee
```

Demo

Multimodal search and Hybrid Search

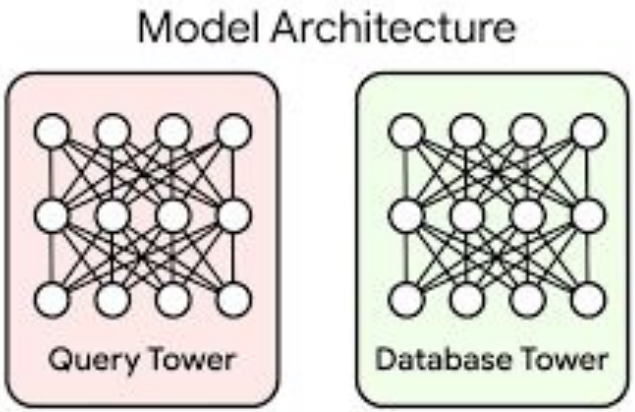
Task type embedding

Questions and their answers are not semantically similar



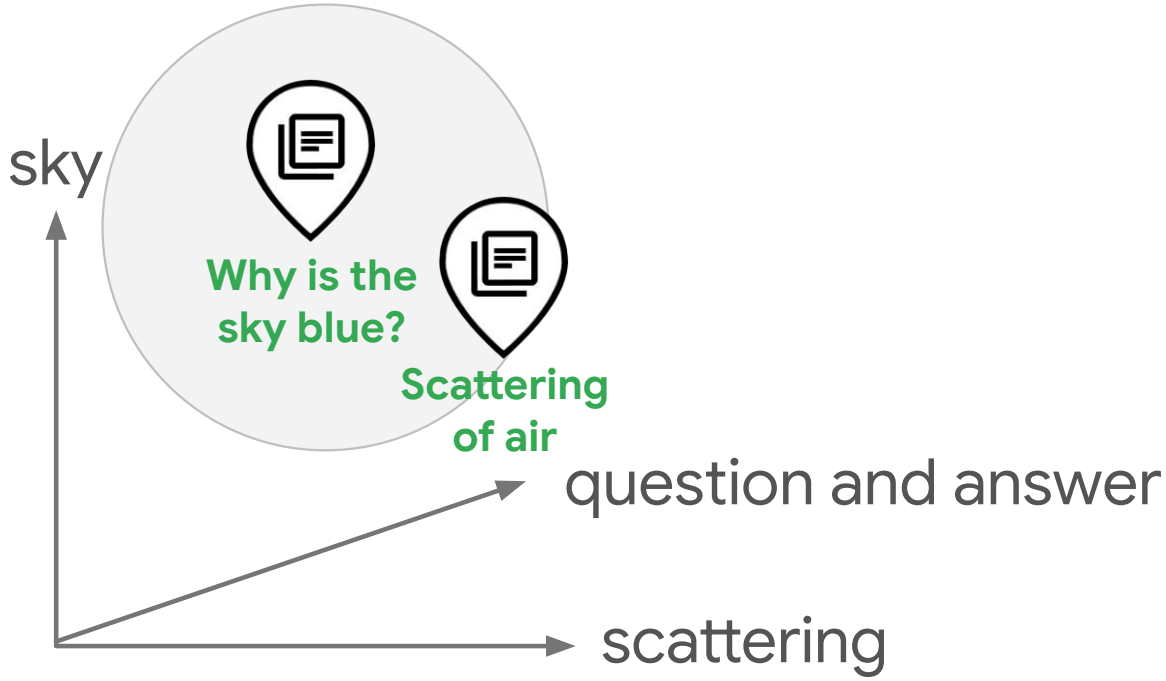
Task type embedding

Learns the relationship between questions and answers



Task type embedding

Maps the question and answer closer in the embedding space



Demo

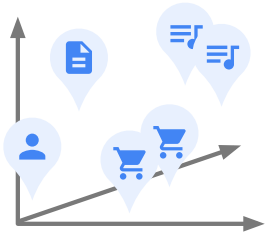
Task type embeddings

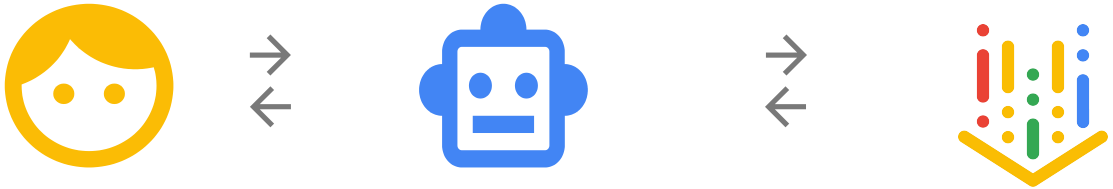
03. Building Agentic RAG with ADK + Vector Search

Challenge: *Smart* Recommendations

Can you suggest **birthday present for my son?** What's the **latest trends?**

Hmm how do I make **recommendations** with **latest trends?**





User **LLM** **Retrieval with text similarity search**

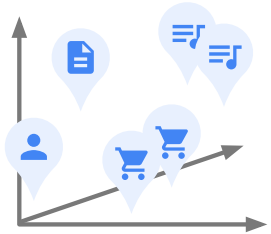
Solution: AI Agents + Vector Search

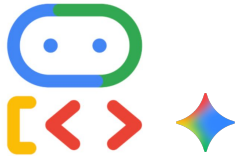
Any birthday present for my 10 yrs son?

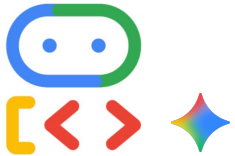
Hey Search Agent, run 20 queries for "Science Toys"

Got it! Running 20 queries in parallel...

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







User

UI Agent

Search Agent

Google Search & Vector Search

Demo

Shopper’s Concierge

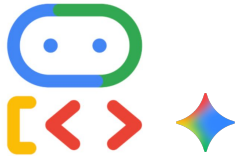
Shopper's Concierge

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

Any birthday present for my 10 yrs son?

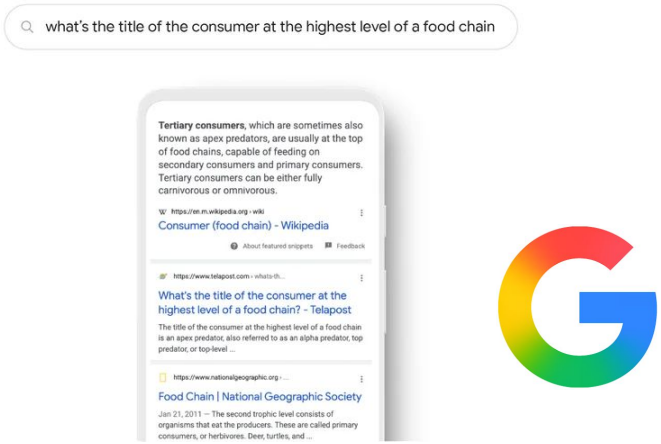
Hey Google Search, what do people buy for birthday presents?

User



UI Agent

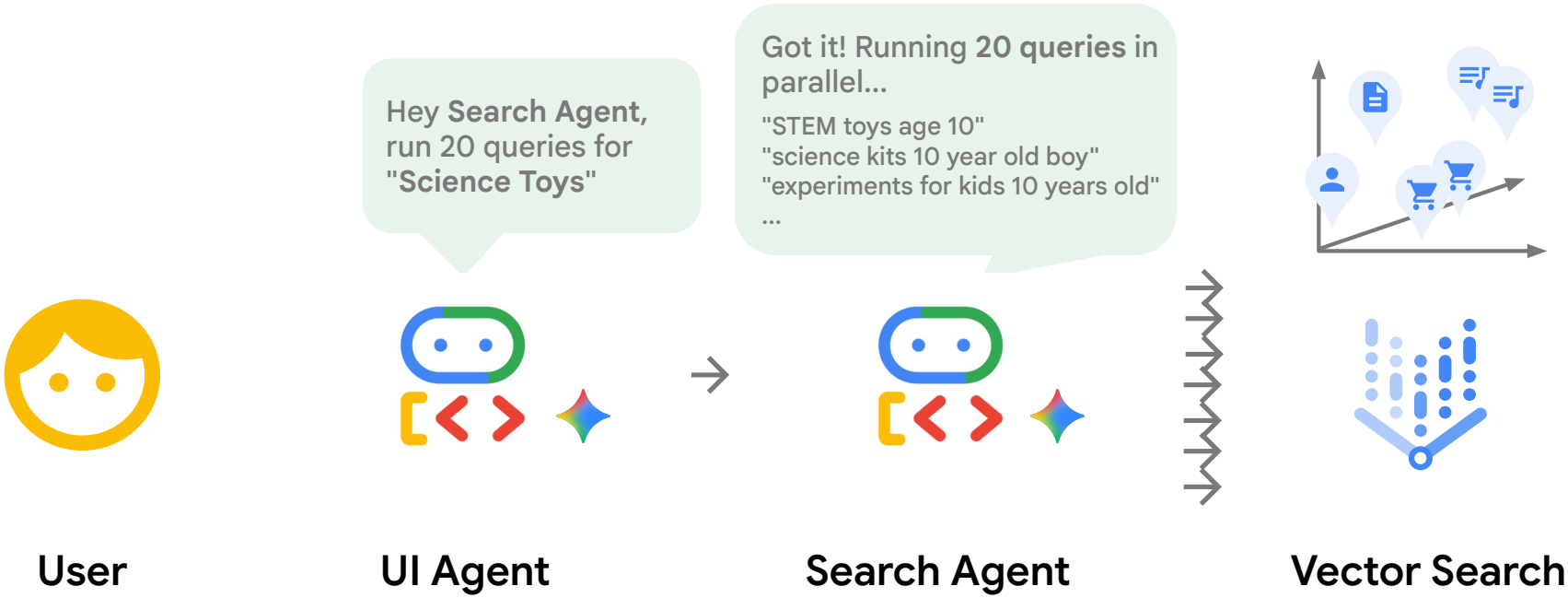
Sport goods,
Outdoor toys,
Science Toys
...



Google Search grounding

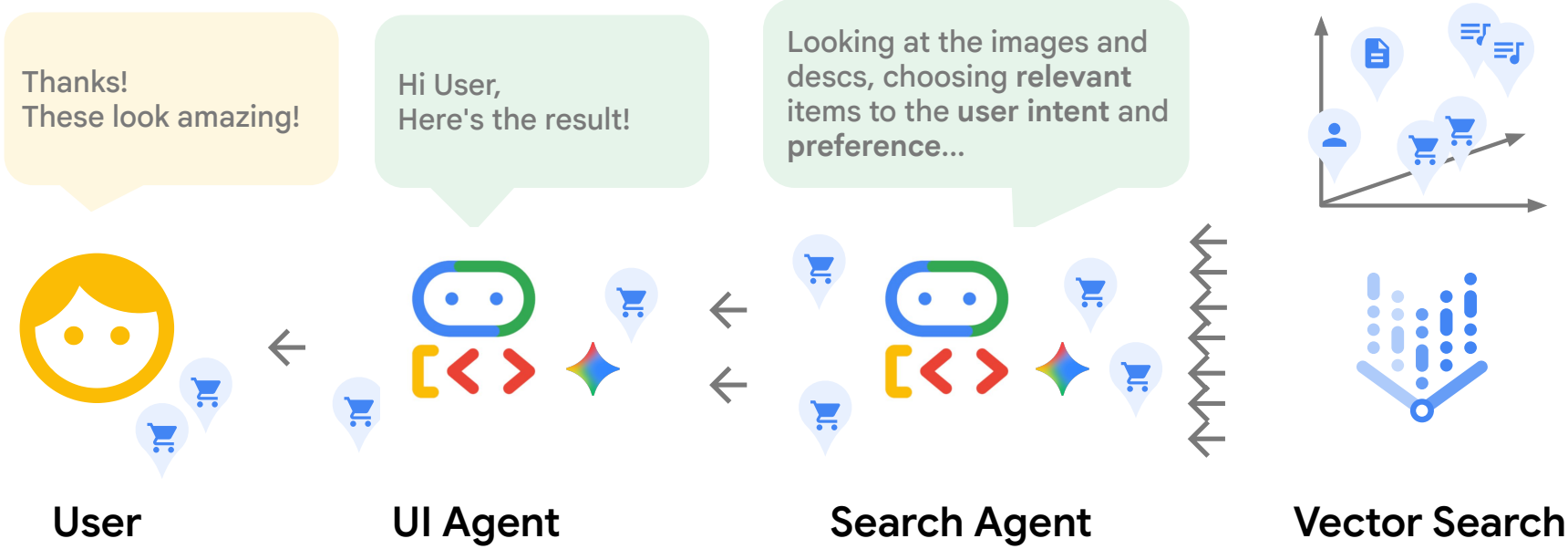
Shopper's Concierge

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



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"

04. Summary

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