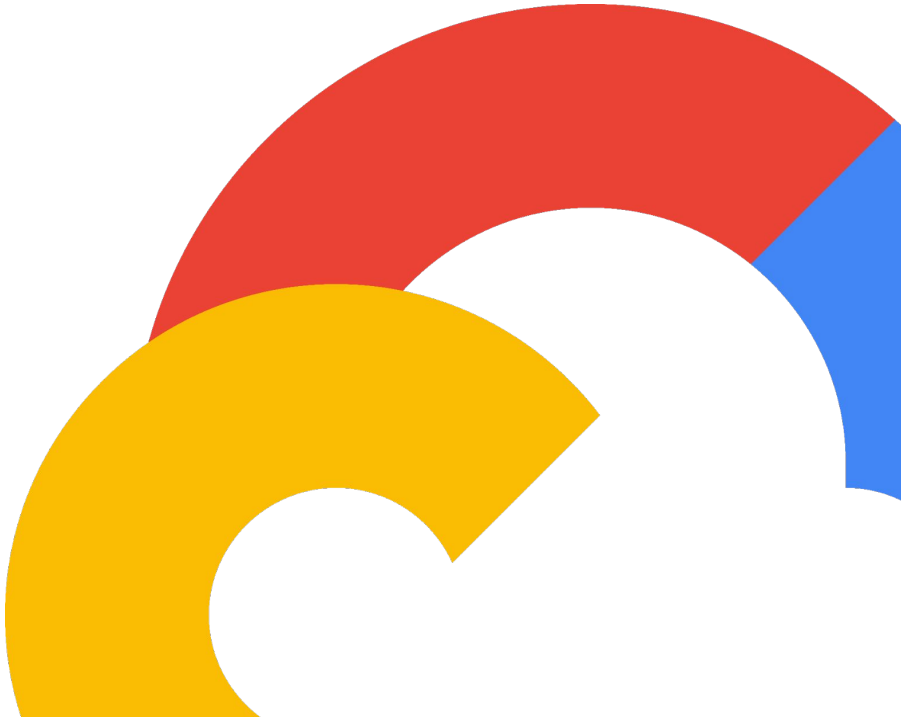
Intro to AI agents and how to build them

July 18th 2024

NERDERY

Google Cloud



Contents

01
Introductions

Speaker introduction

Justin spearheads Nerdery's data and AI team, expertly crafting and scaling solutions that drive impactful customer outcomes. He is passionate about building cross-functional teams — with a focus on becoming more data-driven.

Justin Richie

Nerdery

https://www.linkedin.com/in/justinrichie/

Speaker introduction

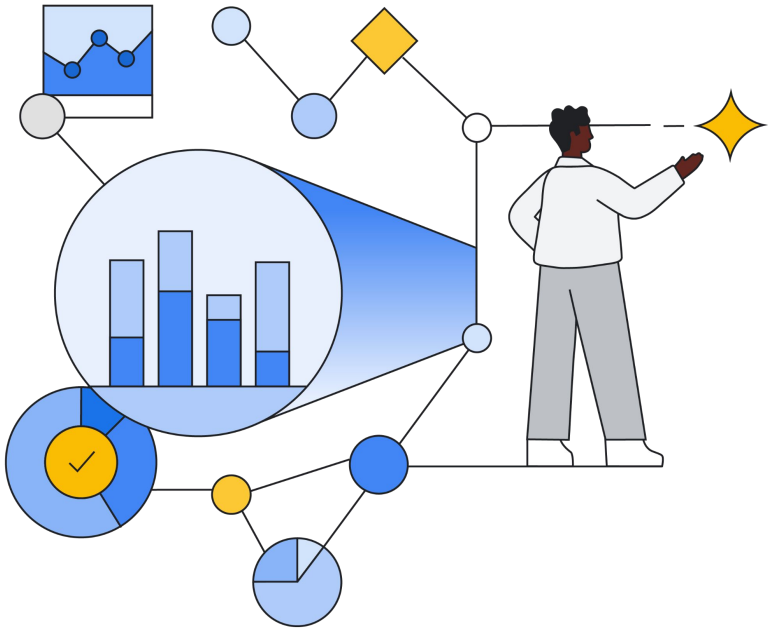
Korbin Graves

Google Cloud

Korbin Graves is a Customer Engineer at Google Cloud with experience working with State and local government agencies as well as higher education. He specializes in helping organizations use Google Cloud's data analytics and AI/ML solutions to solve their most pressing problems. Korbin is passionate about improving citizen services delivered by public sector organizations across all domains.

https://www.linkedin.com/in/korbingraves/

02
AI Agents 101



92%

of Fortune 500 firms have adopted generative AI

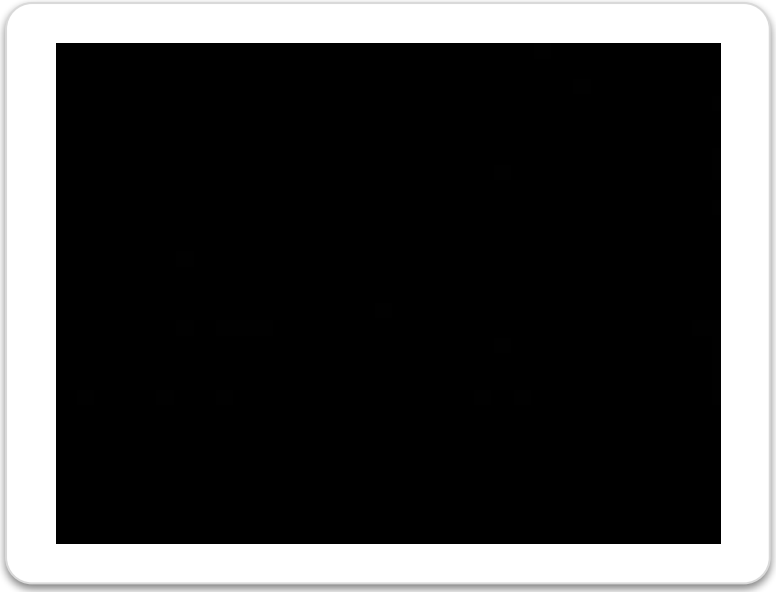
Source: Deloitte , AI Cheat Codes

An **AI agent** is an autonomous software entity designed to perceive its environment, make decisions, and take actions to achieve specific goals with limited or no direct human intervention.

What’s is the difference between LLM, chatbot and AI Agent?

An **AI agent** is the most general term. It refers to any program that can act intelligently in its environment. This can include chatbots, LLMs, and other types of AI systems.

AI agents are all about perceiving, learning, and acting to achieve specific goals. They might not necessarily use language for interaction.



The five components of building an AI Agent

01

Autonomous

It works by itself without needing constant human control.

02

Perceive

It can gather information from its surroundings to gather information.

03

Reason

It thinks about the information it gathers to make smart decisions and accomplish goals.

04

Act

It can do things to change or interact with its environment.

05

Learn

It gets better over time by learning from its experiences.

Types of Agents for this Webinar

Customer Agents

- Product Support
- Customer Orders + FAQs
- Discover + Experiences

Employee Agents

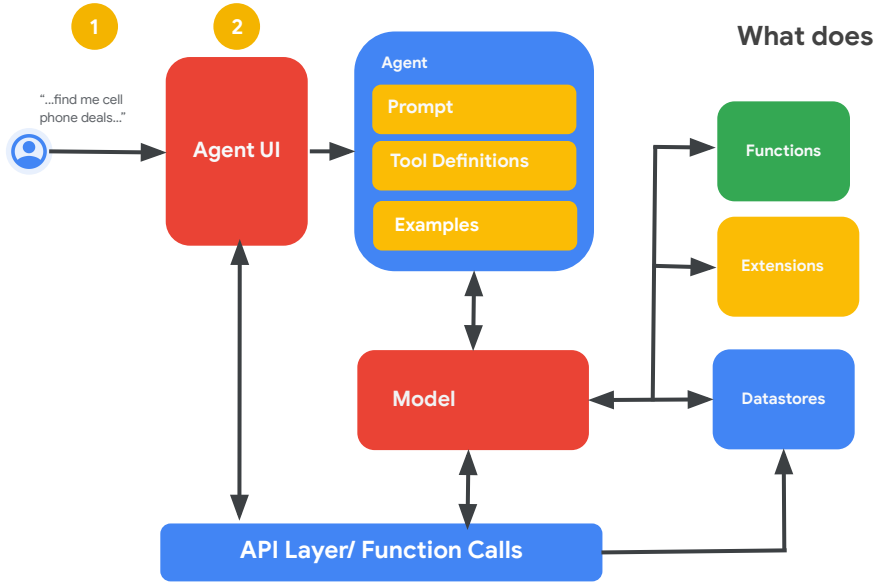
- Employee Onboarding
- Inventory Management
- Invoice + Fulfillment

Knowledge Agents

- Legal + Financial Services
- Mkt Research + Data Science
- Sales Agents

What is an AI Agent?

An AI agent is software that autonomously performs tasks or decisions, simulating human behavior and intelligence.



What does it do?

- Creates chatbots
- Automates tasks
- Integrates data sources
- Understands user intent
- Generates text responses

Advantages

- Build conversational AI
- Drag-and-drop interface
- Pre-built templates
- Multi-agent workflows
- Enterprise-grade security

Google has led the way to make AI Agents available to all organizations

2019

Dialogflow CX

Low-code platform
Intents, Flows & Webhooks
Visual Builder
Developer tools: build and deploy
Voice & IVR integrations
Chat client integrations
Sec-4 compliance

2022

Hybrid: Flows + Agents

Generative features on Flows:
- Generator fulfillment
- Generative fallback

Palm Text-bison LLM
20+ prebuilt templates

2023

Vertex AI Agents

No-code console
Gemini
Multi-lingual
RAG Support
Few shot example editors
Industry prebuilt templates
Multiple Region Support

2024

Multi-modal AI Agents

Gemini Pro
Multi-model: images, text and video support
Custom models and fine tuning available
RAG: More document types
Expanded Region Support

Key customer considerations Agent Builder addresses

Agent Builder provides a platform to build applications based on **customer challenges**

1. "How can we easily build **production-ready Agents**?"

2. "How do we ensure we use search and RAG to ground in our enterprise truth?"

3. "How can we balance the need for **control and customization** while building agents?"

4. "How do we **mitigate the risks of hallucinations?**" "How do I ensure safe and trusted brand reputation?

5. "**Security and privacy** are our main concerns when it comes to enterprise data; how are these concerns alleviated?"

Vertex AI Agent Builder Enterprise Ready Tools

Develop and deploy agents faster, grounded in your enterprise truth

Orchestrate

Create, launch, and manage your agents at scale

Build at any level: no code[1], low code, or full code options in **Vertex AI Agent Builder & Agent API**

Deploy and orchestrate custom agents with **LangChain on Vertex**
Public Preview

Ground & Augment

Increase generative AI output accuracy and relevancy

Ground with Google Search to access fresh, high quality information
Public Preview

Ground on your own enterprise data quickly with out-of-the-box RAG in **Vertex AI Search**
Generally Available

Build DIY RAG providers with **LlamaIndex on Vertex**
Public Preview

Use Tools

Connect LLMs to external tools; call APIs and Services

Create your own actions with **Function Calling** accessing custom or private APIs
Generally Available

Access pre-built reusable modules with **Extensions**[2]
Public Preview

[1] Using natural language (Public Preview)
[2] Two available: Code Interpreter and Vertex AI Search

03
AI Agent GCP Demo

Create App

LEARN

1 Type
2 Configuration
3 Data

Select app type

Select the type of application you want to create

Search

Get quality results out-of-box and easily customize the engine

SELECT

Chat

Answer complex questions out-of-the-box

SELECT

Recommendations

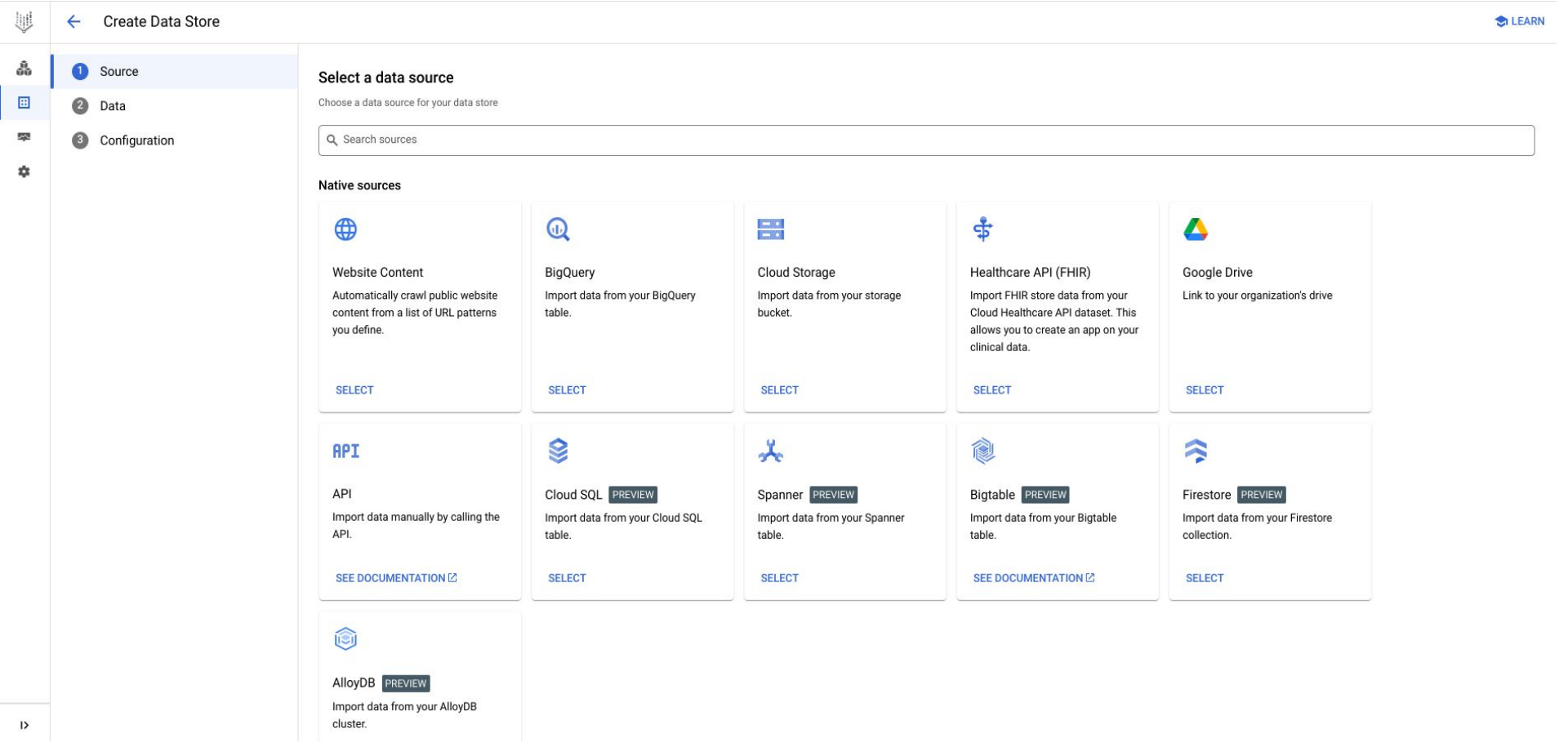
Create a content recommendation engine

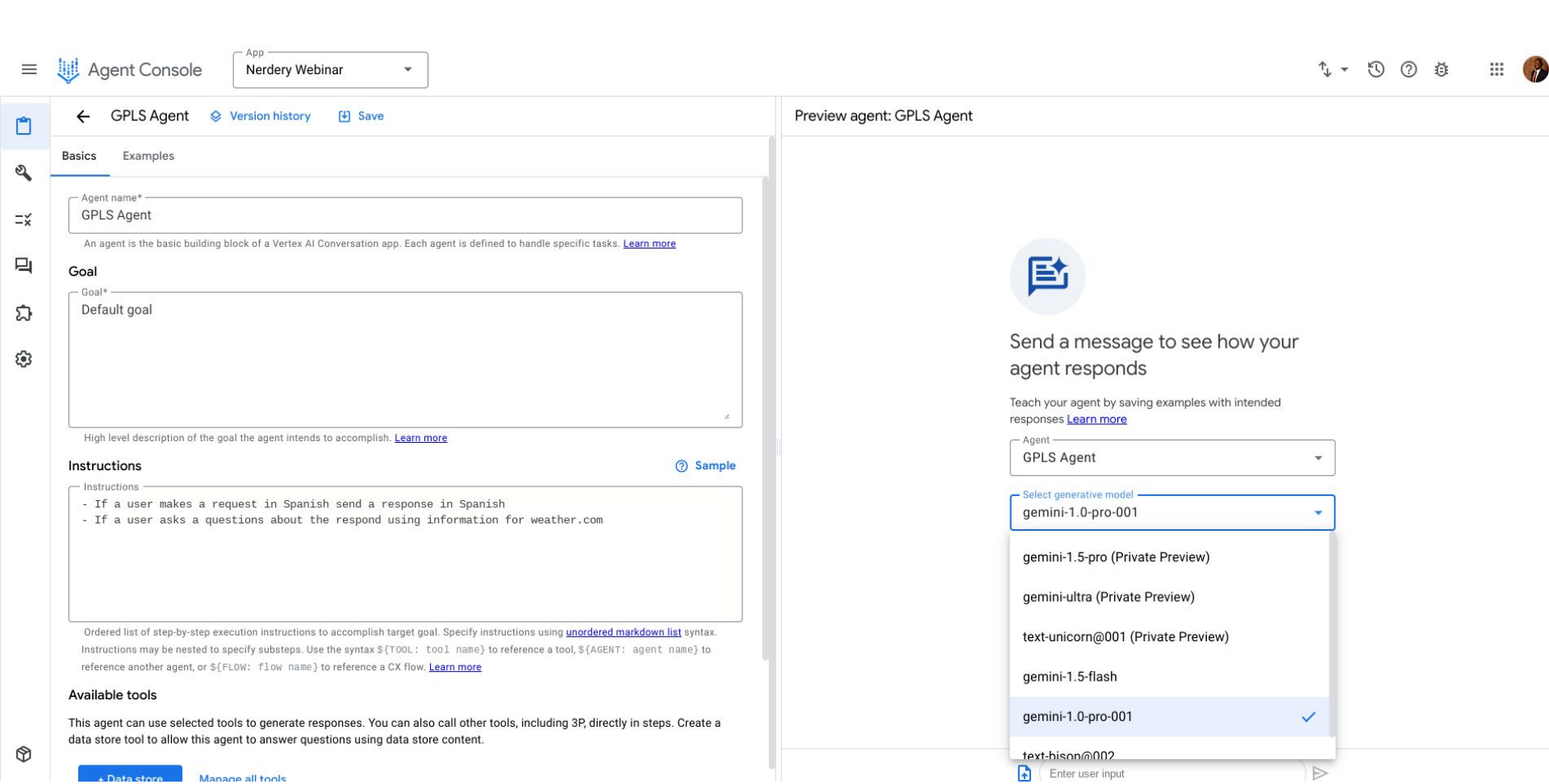
SELECT

Agent PREVIEW

Built using natural language, agents can answer questions from data, connect with business systems through tools, and more

SELECT


Create Data Store
LEARN
1 Source
2 Data
3 Configuration
Select a data source
Choose a data source for your data store
Search sources
Native sources
Website Content
Automatically crawl public website content from a list of URL patterns you define.
SELECT
BigQuery
Import data from your BigQuery table.
SELECT
Cloud Storage
Import data from your storage bucket.
SELECT
Healthcare API (FHIR)
Import FHIR store data from your Cloud Healthcare API dataset. This allows you to create an app on your clinical data.
SELECT
Google Drive
Link to your organization's drive
SELECT
API
API
Import data manually by calling the API.
SEE DOCUMENTATION
Cloud SQL PREVIEW
Import data from your Cloud SQL table.
SELECT
Spanner PREVIEW
Import data from your Spanner table.
SELECT
Bigtable PREVIEW
Import data from your Bigtable table.
SEE DOCUMENTATION
Firestore PREVIEW
Import data from your Firestore collection.
SELECT
AlloyDB PREVIEW
Import data from your AlloyDB cluster.


Agent Console
App
Nerdery Webinar
GPLS Agent
Version history
Save
Basics
Examples
Agent name*
GPLS Agent
An agent is the basic building block of a Vertex AI Conversation app. Each agent is defined to handle specific tasks. Learn more
Goal
Goal*
Default goal
High level description of the goal the agent intends to accomplish. Learn more
Instructions
Sample
Instructions
- If a user makes a request in Spanish send a response in Spanish
- If a user asks a questions about the respond using information for weather.com
Ordered list of step-by-step execution instructions to accomplish target goal. Specify instructions using unordered markdown list syntax. Instructions may be nested to specify substeps. Use the syntax ${TOOL: tool name} to reference a tool, ${AGENT: agent name} to reference another agent, or ${FLOW: flow name} to reference a CX flow. Learn more
Available tools
This agent can use selected tools to generate responses. You can also call other tools, including 3P, directly in steps. Create a data store tool to allow this agent to answer questions using data store content.
+ Data store
Manage all tools
Preview agent: GPLS Agent
Send a message to see how your agent responds
Teach your agent by saving examples with intended responses Learn more
Agent
GPLS Agent
Select generative model
gemini-1.0-pro-001
gemini-1.5-pro (Private Preview)
gemini-ultra (Private Preview)
text-unicorn@001 (Private Preview)
gemini-1.5-flash
gemini-1.0-pro-001
text-bison@002
Enter user input

04
Extending AI Agents

Why build your own custom AI Agent?

Agentic Workflows

Builder tools offer pre-built functionalities, but a custom agent lets you **tailor the LLM and retrieval tasks** to your specific needs, **increasing performance,** in addition to multiple agents.

Data & Security Control

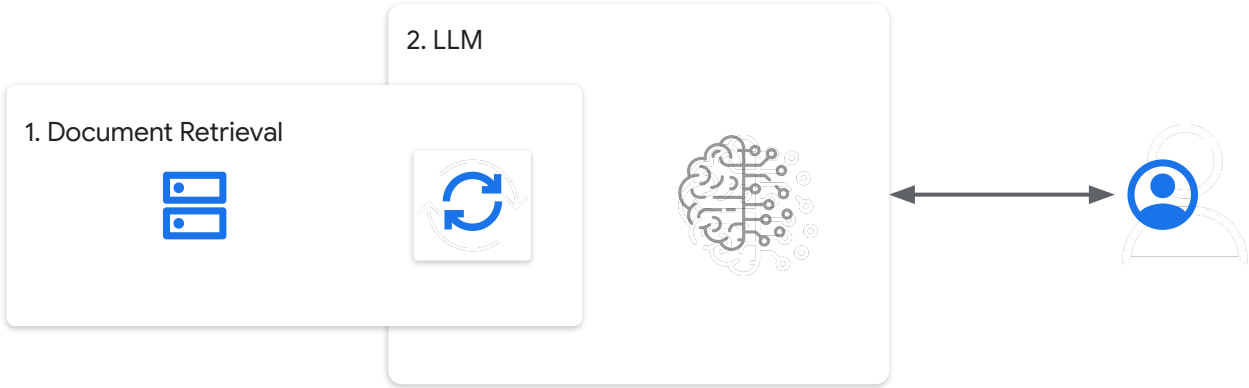
A custom approach gives you **complete control over data** used for training and **enhanced security** measures to protect sensitive information.

Custom Integrations

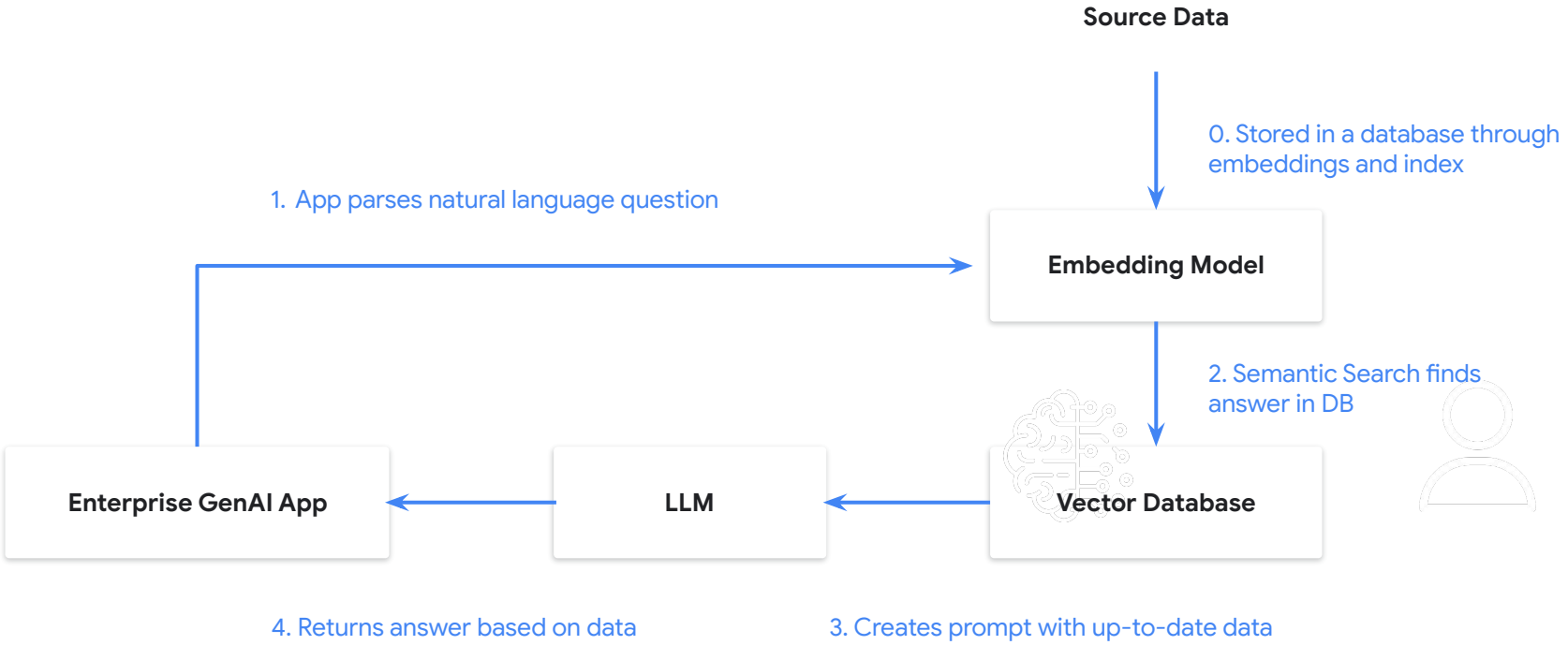
Builder tools may have limitations on handling **enterprise-level traffic**. Building custom allows scaling and flexibility to integrate with complex workflows and data pipelines. Function calls fit into this section.

RAG 101

RAG, or Retriever-Augmented Generation, combines information retrieval with generative models to enhance AI's ability to provide relevant responses



RAG architecture



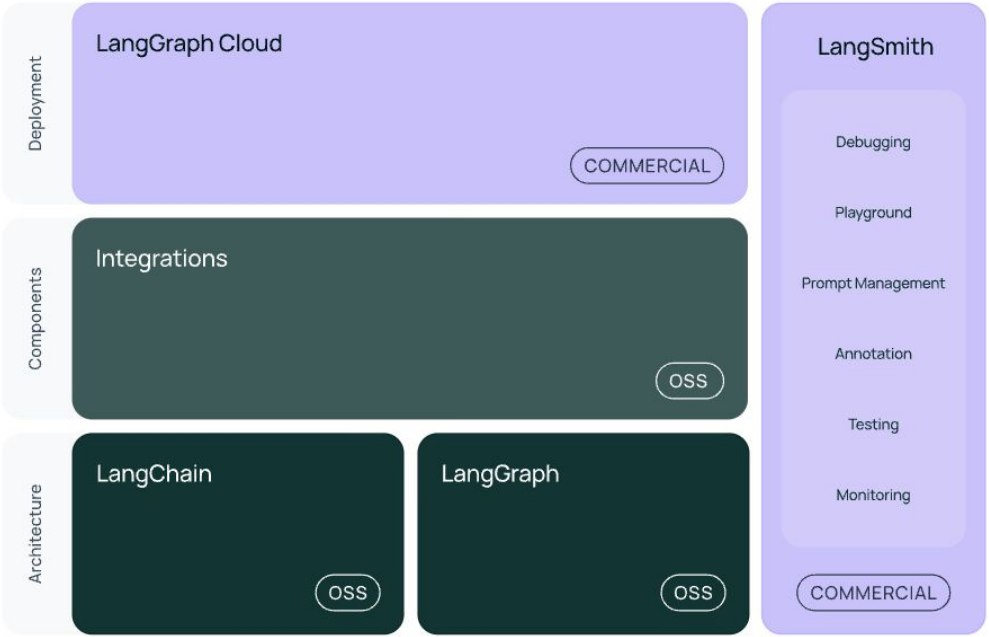
LangChain

LangChain is a framework for developing applications powered by large language models (LLMs).

Development: Build your applications using LangChain's open-source building blocks, components, and third-party integrations. Use LangGraph to build stateful agents with first-class streaming and human-in-the-loop support.

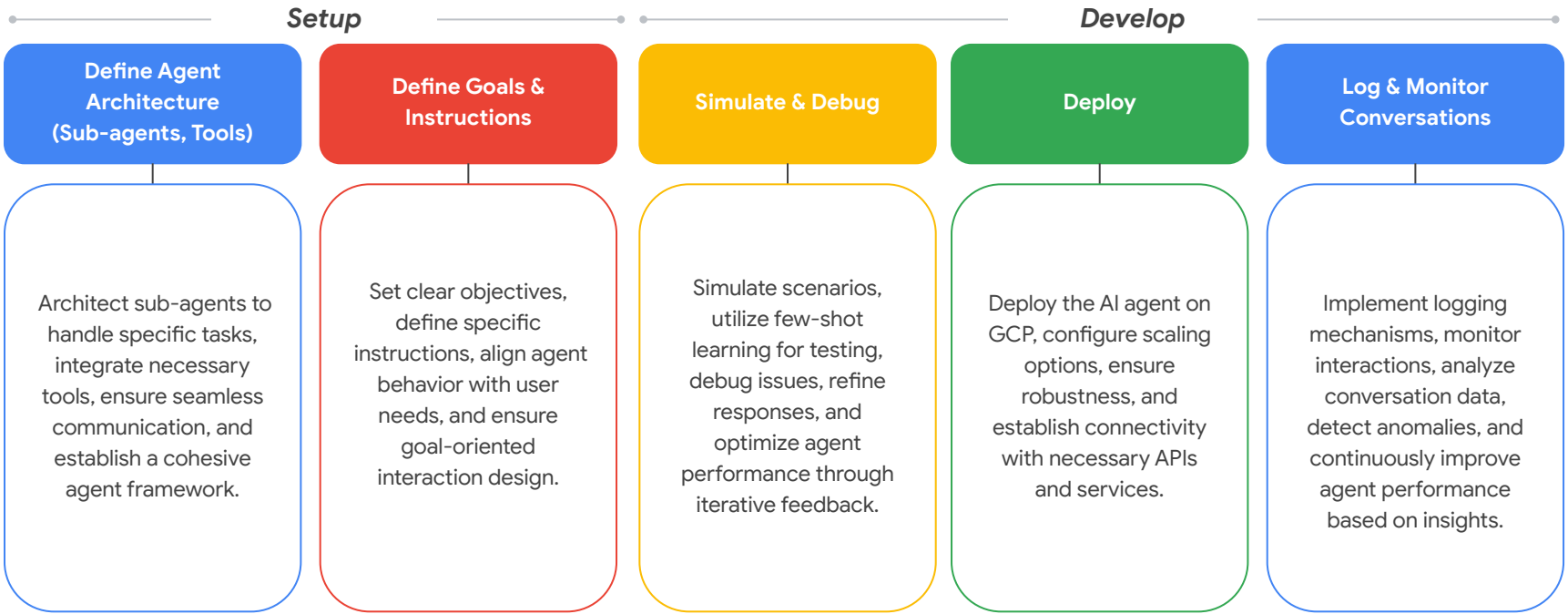
Productionization: Use LangSmith to inspect, monitor and evaluate your chains, so that you can continuously optimize and deploy with confidence.

Deployment: Turn your LangGraph applications into production-ready APIs and Assistants with LangGraph Cloud.



AI Agent Console

GCP provides Life-cycle management covering all aspects of the AI Agent workflow



Evolving from prompt engineering to flow engineering

Zero-shot Prompting

This is the most basic approach. You directly instruct the AI Agent with a natural language prompt, hoping it understands and completes the task. It's flexible but can be unreliable, especially for complex tasks.

Few-shot Prompting

This refines the approach by providing a few examples along with the prompt. These examples guide the AI Agent towards the desired outcome. It offers more control than zero-shot, but requires some effort to create the examples.

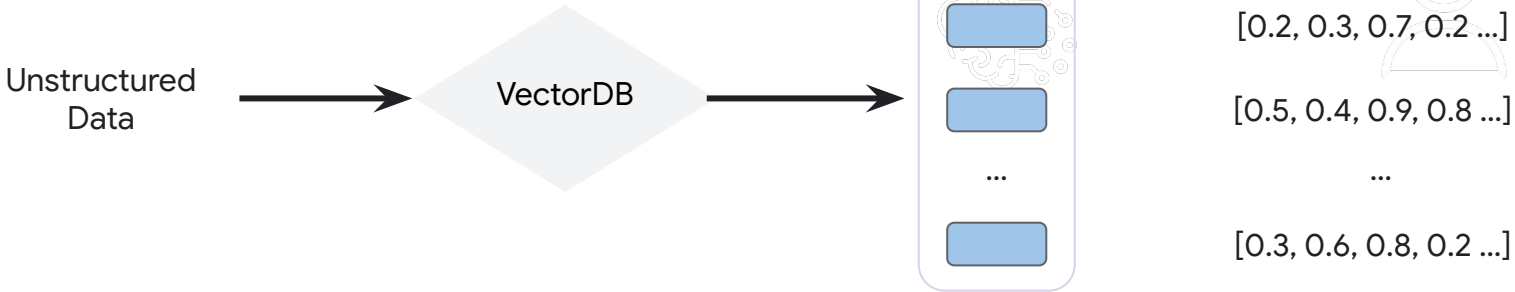
Function Calling

This represents the most structured approach. You define specific functions the AI Agent can call, each performing a well-defined task. This offers the most control and allows for complex functionalities, but requires more development effort to build the functions.

A note on vector databases

Index is a data structure to add in search process, embeddings are the distance in related items (ANN)

- Long-term memory for LLMs
- Semantic and Similarity search
- Many ways to optimize



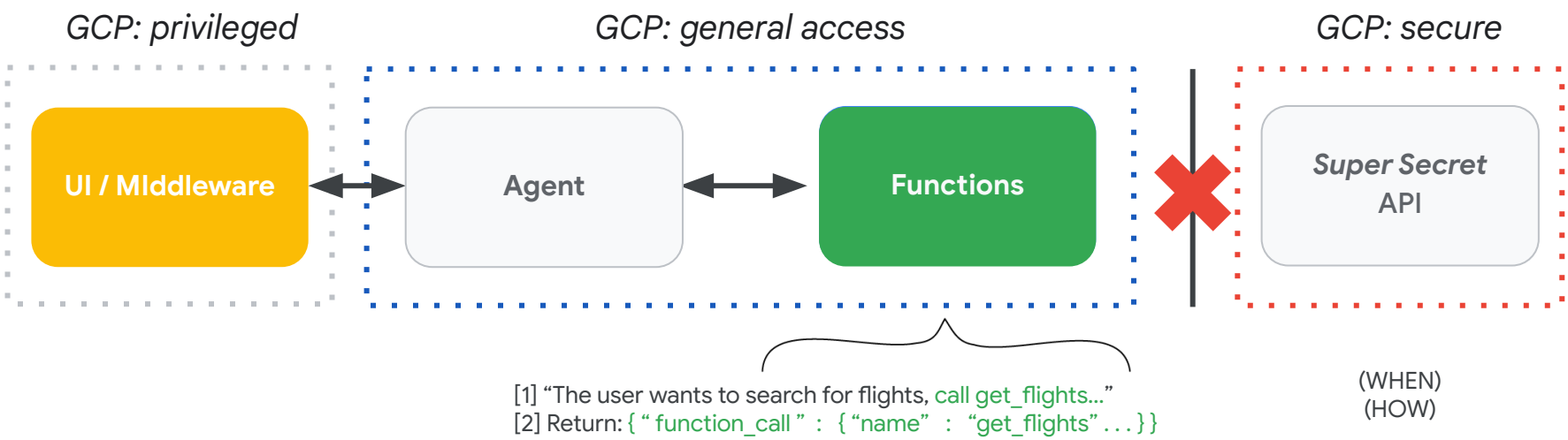
Chunking Strategy

- Documents are broken down into smaller, manageable chunks.
- With smaller chunks, the vector database can quickly retrieve and rank relevant chunks.
- **How to improve?**
- **Use Logical Breakpoints:** Prefer breaking chunks at natural paragraph or sentence boundaries.
- **Adjust Chunk Size and Overlap:** Experiment with different chunk_size and chunk_overlap values to find the best balance.
- **Post-Processing for Smooth Transitions:** After initial chunking, you can implement a post-processing step to merge small chunks with adjacent ones

```
# Partitioning the data with enhanced readability
text_splitter = RecursiveCharacterTextSplitter(
    separators=["\n\n", "\n", ". ", "; ", ", ", " ", ""],  # Added more granular s
    chunk_size=1500,  # Increased chunk size for more content per chunk
    chunk_overlap=300  # Increased overlap to preserve more context
)

# Load and split the documents
splits = text_splitter.split_documents(data_on_pdf)

# Function to merge small chunks with the previous one
def merge_small_chunks(chunks, min_size=500):
    merged_chunks = []
    current_chunk = ""
    for chunk in chunks:
        if len(current_chunk) + len(chunk) < min_size:
            current_chunk += chunk
        else:
            if current_chunk:
                merged_chunks.append(current_chunk)
            current_chunk = chunk
    if current_chunk:
        merged_chunks.append(current_chunk)
    return merged_chunks

# Apply the merging function
optimized_splits = merge_small_chunks(splits)
```

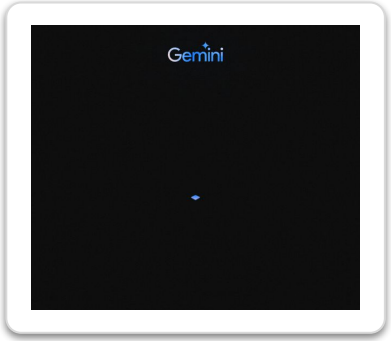
Function Calling

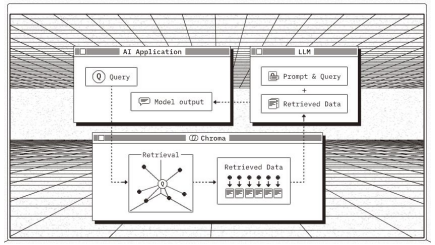
GCP provides Life-cycle management covering all aspects of the AI Agent workflow

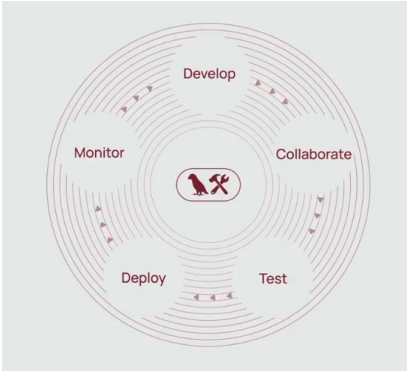


What is in this demo?

Building AI Agents is a blend of components to build next-gen applications







Gemini

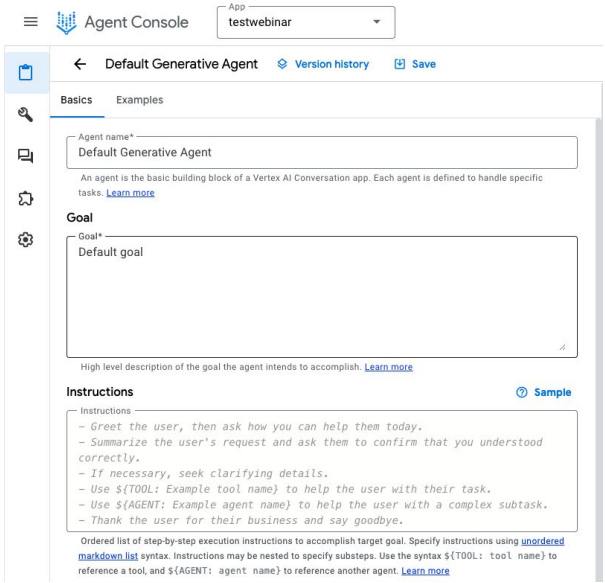
LangChain

05
RAG and Langchain Demo

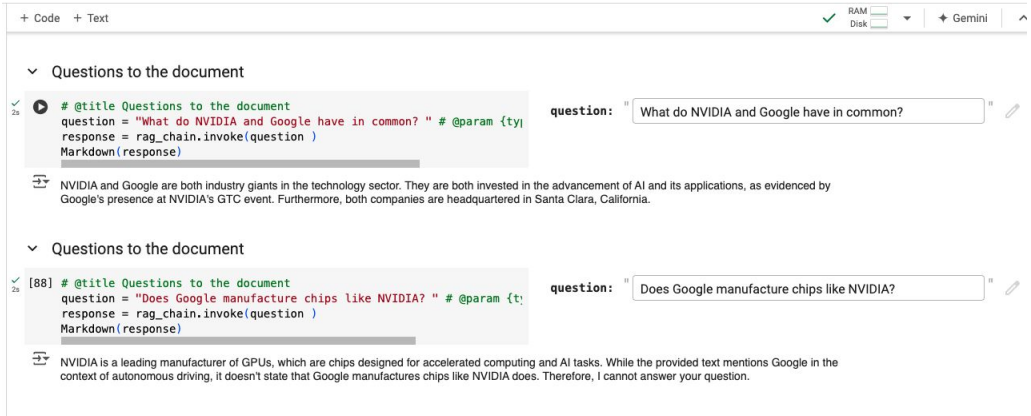
06
Closing

We covered a lot!

AI Agents in GCP

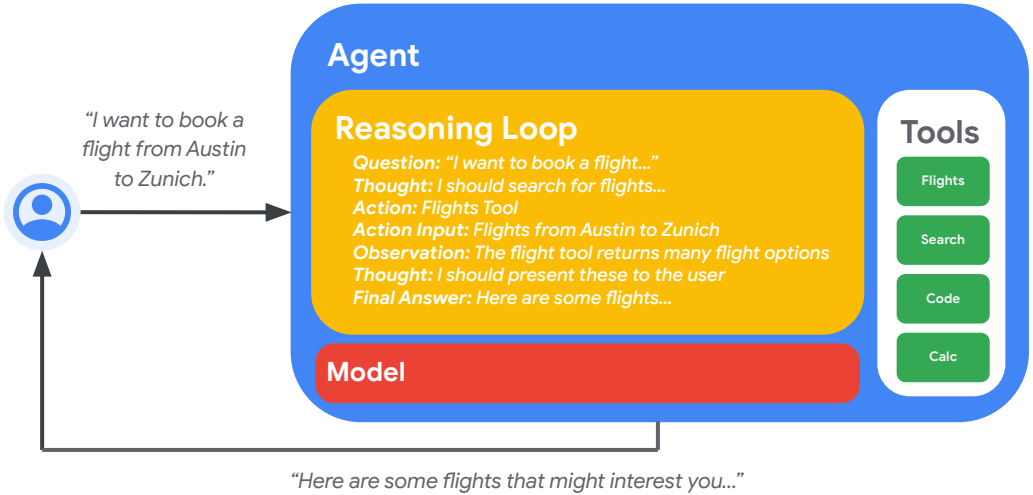


Custom RAG/Langchain App



User Queries or Events Initiate Interactions

An AI agent is software that autonomously performs tasks or decisions, simulating human behavior and intelligence.



Recommendation as you begin your journey

An AI agent is software that autonomously performs tasks or decisions, simulating human behavior and intelligence.

01

Leverage GCP first

Managing infrastructure is very time intensive and tedious.

02

Flows are the future

AI agents will improve with improve flow design and more data.

03

Log + Monitor

Security, ethics and performance all hinge on this facet.

AI Agent Adoption

“I think AI agent workflows will drive massive AI progress this year — perhaps even more than the next generation of foundation models. This is an important trend, and I urge everyone who works in AI to pay attention to it.”

07
Q&A