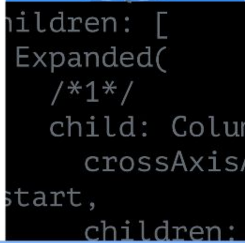
Build Generative AI Agents with Vertex AI Agent Builder



```
hildren: [
Expanded(
  /*1*/
  child: Colu
    crossAxis
start,
    children:
```

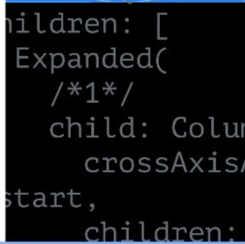
Google Developer Groups
Cloud • Rajkot

Cloud Community Day 2024

Introduction

- AI solutions that optimize business operations and improve decision-making processes.
- Expert in Google Cloud Platform technologies, including Vertex AI with a strong background in building and integrating intelligent conversation AI solutions for enhanced customer engagement
- M.Sc Physics from NIT Surat
- Linkedin Top AI Voice

Yash Kavaiya
AI/ ML Engineer at TCS



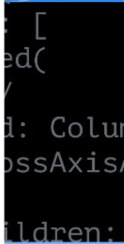
```
ildren: [
 Expanded(
   /*1*/
   child: Colu
     crossAxis
start,
     children:
```

Google Developer Groups
Cloud • Rajkot

Cloud Community Day 2024

Agenda

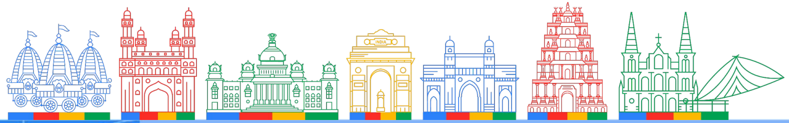
- Introduction to Generative AI
- What are Large Language Models ?
- Neural Networks
- What are Large Language Models ?
- Overview of Vertex AI
- Retrieval-Augmented Generation
- Demo-1 Vertex AI Agent with Agent Builder
- Chainlit UI with Google Gemini API
- PDF Using Streamlit App and Gemini API



```
: [
ed(
/
d: Colu
ssAxis

ildren:
```

Google Developer Groups
Cloud • Rajkot

Cloud Community Day 2024



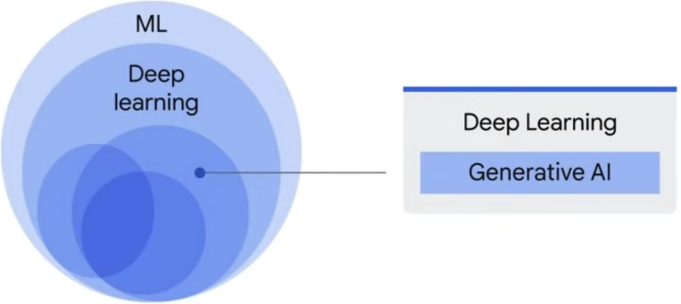
Generative AI

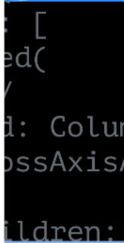
Create new content

(Audio, Code ,Text, Video, Images)

Automatically using computer program

Generative AI is a subset of Deep Learning

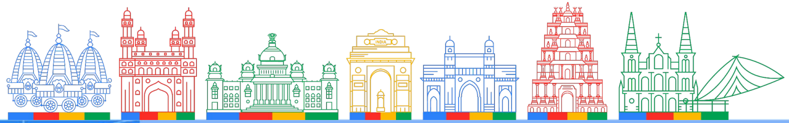




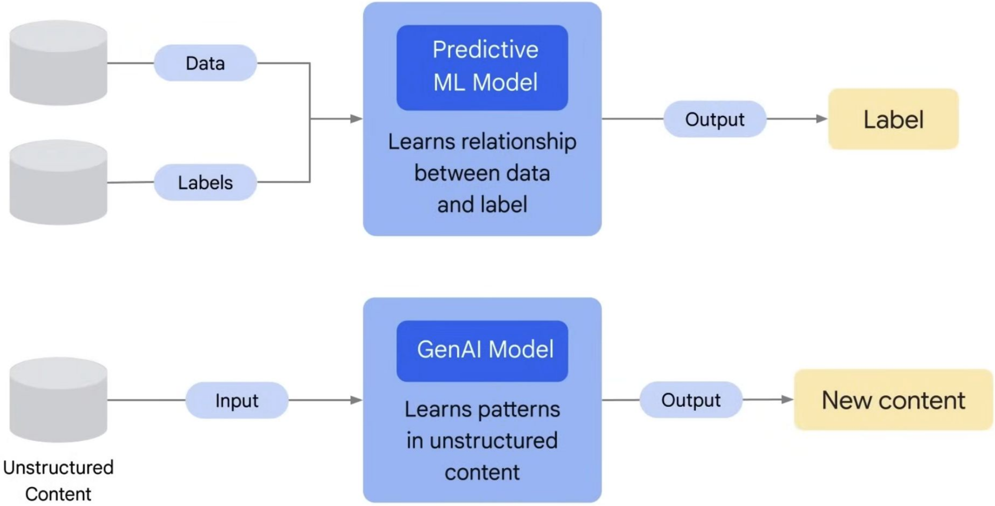
```
: [
ed(
/
d: Colu
ossAxis
ildren:
```

Google Developer Groups
Cloud • Rajkot

Cloud Community Day 2024



Predictive ML vs Generative AI



Data

Labels

Predictive ML Model

Learns relationship between data and label

Output

Label

Unstructured Content

Input

GenAI Model

Learns patterns in unstructured content

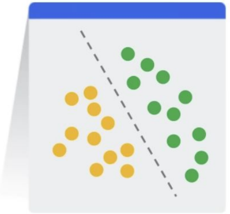
Output

New content

```
: [
ed(
/
d: Colu
ssAxis
ildren:
```

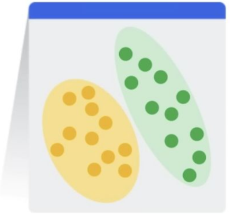
Discriminative AI vs Generative AI

Deep Learning Model Types



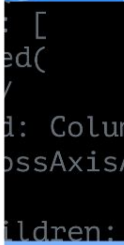
Discriminative

- Used to classify or predict
- Typically trained on a dataset of labeled data
- Learns the relationship between the features of the data points and the labels



Generative

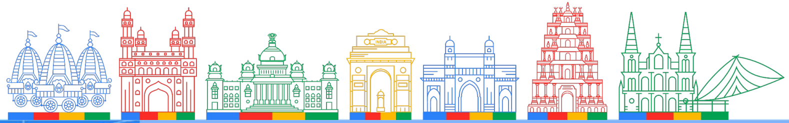
- Generates new data that is similar to data it was trained on
- Understands distribution of data and how likely a given example is
- Predict next word in a sequence



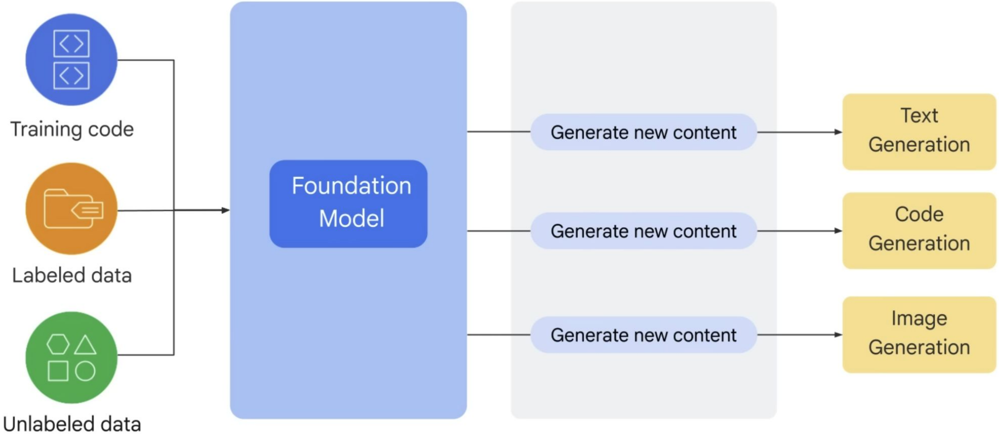
```
: [
ed(
/
d: Colu
ossAxis

ildren:
```

Google Developer Groups
Cloud • Rajkot

Cloud Community Day 2024



Gen AI Output



Gen AI Supervised, Semi-Supervised & Unsupervised Learning

Training code

Labeled data

Unlabeled data

Foundation Model

Generate new content

Generate new content

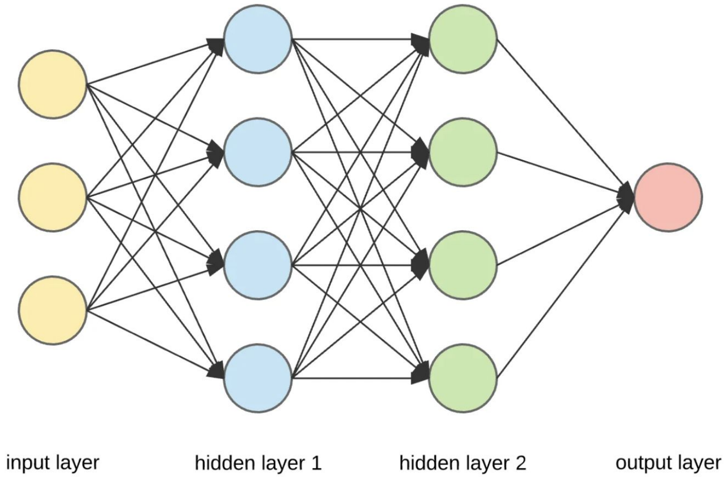
Generate new content

Text Generation

Code Generation

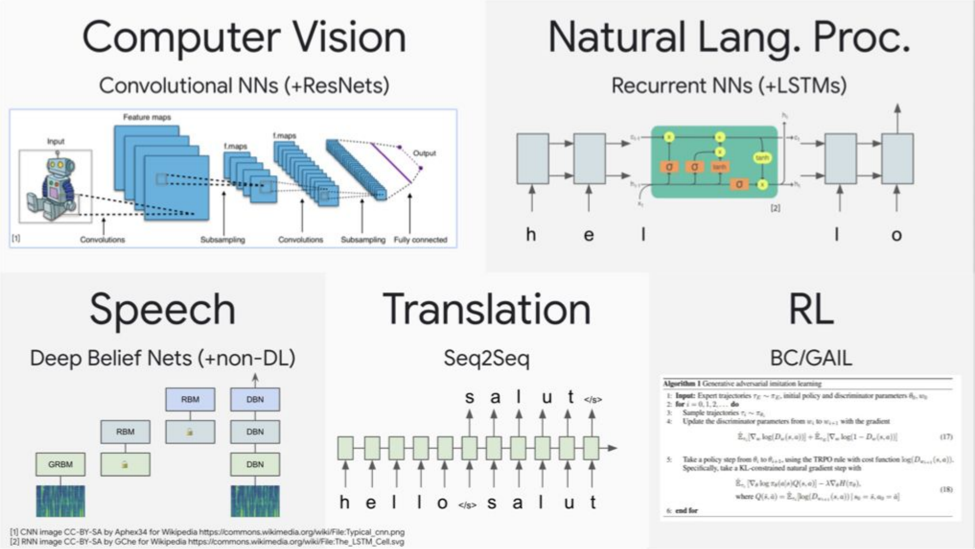
Image Generation

Neural Networks



A neural network is a computational model inspired by the way biological neural networks in the human brain process information. It consists of interconnected nodes (neurons) organized in layers, which work together to recognize patterns and make decisions.

Neural Networks



Computer Vision

Convolutional NNs (+ResNets)

Natural Lang. Proc.

Recurrent NNs (+LSTMs)

Speech

Deep Belief Nets (+non-DL)

Translation

Seq2Seq

RL

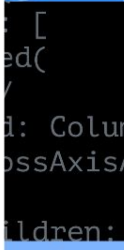
BC/GAIL

Algorithm 1 Generative adversarial imitation learning

```
1: Input: Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters $\theta_0, w_0$
2: for $i = 0, 1, 2, \ldots$ do
3:     Sample trajectories $\tau_i \sim \pi_{\theta_i}$
4:     Update the discriminator parameters from $w_i$ to $w_{i+1}$ with the gradient
           $\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w \log(1 - D_w(s,a))]$   (17)
5:     Take a policy step from $\theta_i$ to $\theta_{i+1}$, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s,a))$.
       Specifically, take a KL-constrained natural gradient step with
           $\hat{\mathbb{E}}_{\tau_i}[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)] - \lambda \nabla_\theta H(\pi_\theta)$,
           where $Q(\bar{s},\bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{w_{i+1}}(s,a)) \mid s_0 = \bar{s}, a_0 = \bar{a}]$   (18)
6: end for
```

[1] CNN image CC-BY-SA by Aphex34 for Wikipedia https://commons.wikimedia.org/wiki/File:Typical_cnn.png

[2] RNN image CC-BY-SA by GChe for Wikipedia https://commons.wikimedia.org/wiki/File:The_LSTM_Cell.svg



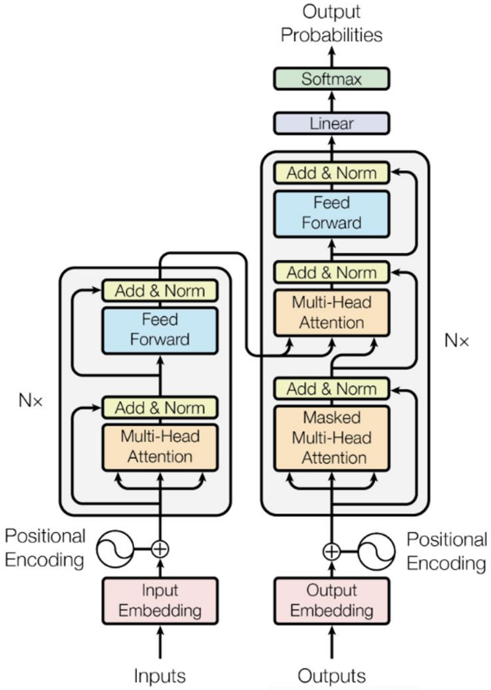
```
: [
ed(
/
d: Colu
ossAxis
ildren:
```

Now : It’s all Transformers

Transformer cartoon (DALL-E)

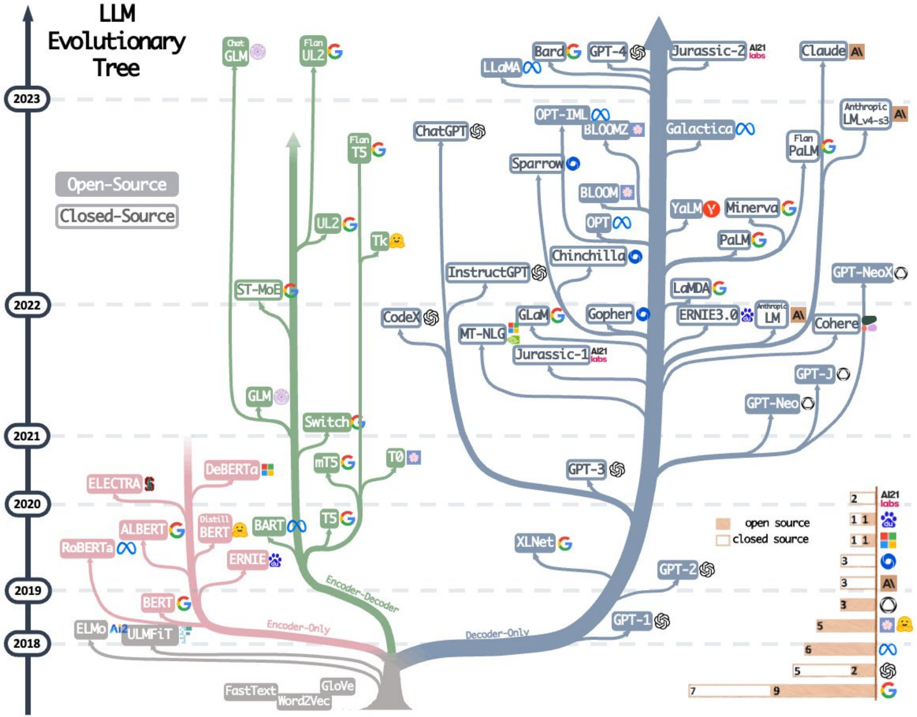
Attention is all you need

- Positional Encoding
- Attention
- Self-attention



What are Large Language Models ?

- Type of machine learning models that are trained on a large data
- Generates outputs for tasks, such as text generation, question answering, and machine translation
- Based on deep learning neural networks, such as the Transformer Architecture
- At the core LLM’s are just language models that can predict the next word in a sentence

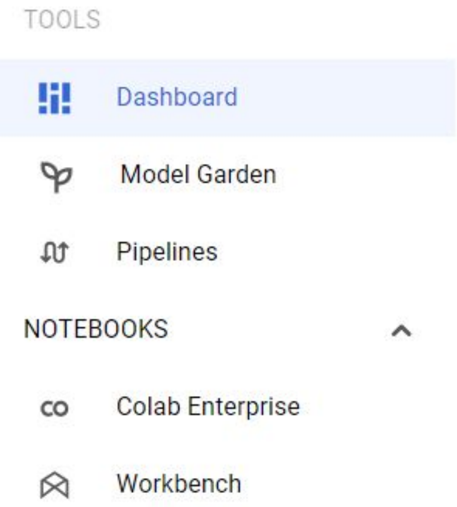


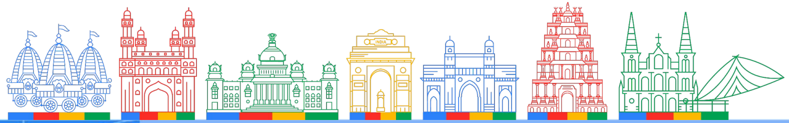
Google Developer Groups
Cloud • Rajkot

Cloud Community Day 2024

What is Vertex AI

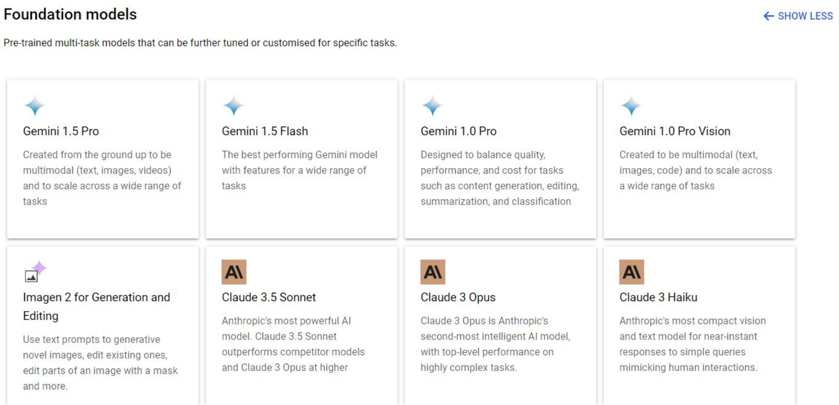
Vertex AI is a Google Cloud platform that simplifies building and using artificial intelligence. It offers tools for every step, from data prep to deployment, and caters to both beginners and experts. Vertex AI includes pre-built models for easy use and also allows you to build your own.





Model Garden

Vertex AI Model Garden is essentially a repository of pre-trained machine learning models that you can easily discover, use, and customize.



Foundation models

Pre-trained multi-task models that can be further tuned or customised for specific tasks.

Gemini 1.5 Pro
Created from the ground up to be multimodal (text, images, videos) and to scale across a wide range of tasks

Gemini 1.5 Flash
The best performing Gemini model with features for a wide range of tasks

Gemini 1.0 Pro
Designed to balance quality, performance, and cost for tasks such as content generation, editing, summarization, and classification

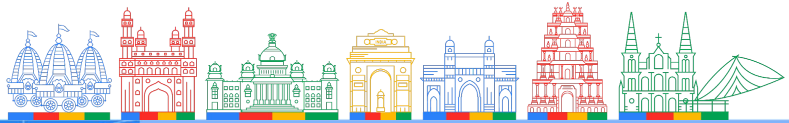
Gemini 1.0 Pro Vision
Created to be multimodal (text, images, code) and to scale across a wide range of tasks

Imagen 2 for Generation and Editing
Use text prompts to generative novel images, edit existing ones, edit parts of an image with a mask and more.

Claude 3.5 Sonnet
Anthropic's most powerful AI model. Claude 3.5 Sonnet outperforms competitor models and Claude 3 Opus at higher

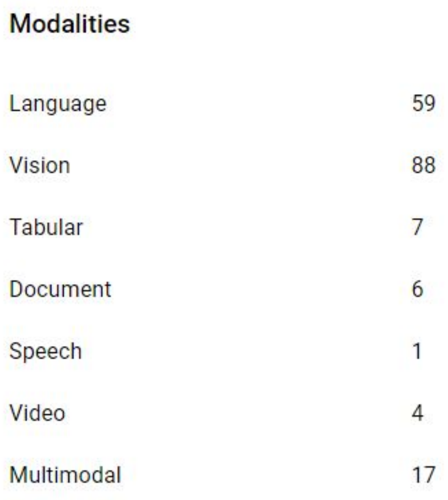
Claude 3 Opus
Claude 3 Opus is Anthropic's second-most intelligent AI model, with top-level performance on highly complex tasks.

Claude 3 Haiku
Anthropic's most compact vision and text model for near-instant responses to simple queries mimicking human interactions.



What's Inside the Garden?

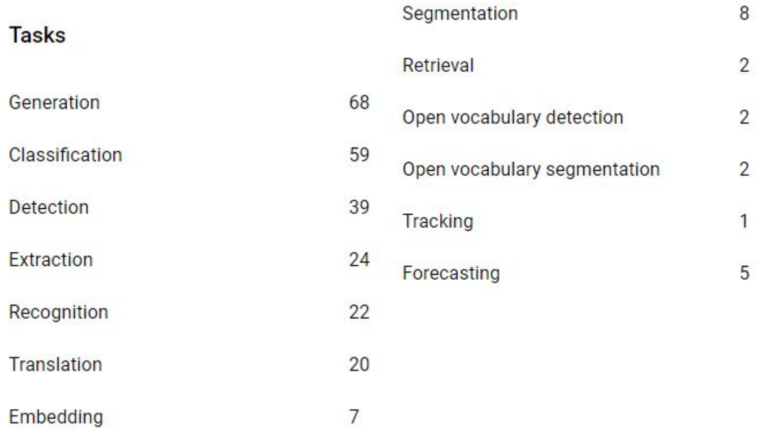
- **Diverse Model Collection:** You'll find a vast array of models covering various domains like computer vision, natural language processing, and more.
- **Foundation Models:** These are powerful models that can be adapted to various tasks with minimal fine-tuning.
- **Task-Specific Models:** These models are tailored for specific tasks, like image classification or sentiment analysis.
- **APIs:** You can directly access and use model functionalities through APIs, without needing to delve into the underlying code.



Modalities	
Language	59
Vision	88
Tabular	7
Document	6
Speech	1
Video	4
Multimodal	17

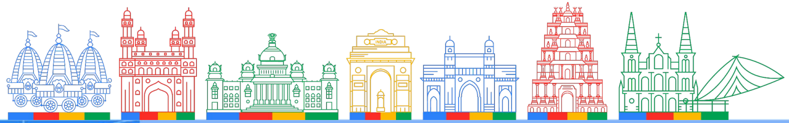
How Does It Benefit You?

- Accelerated Development
- Experimentation
- MLOps Integration
- Customization



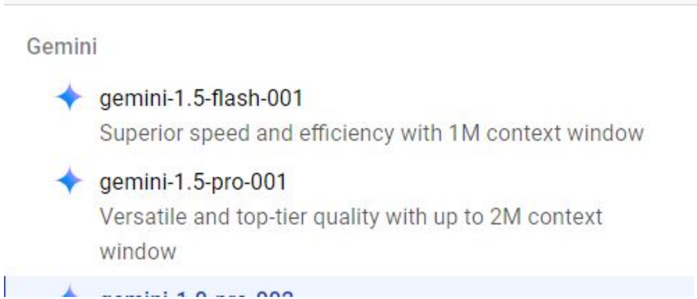
Tasks

Generation	68
Classification	59
Detection	39
Extraction	24
Recognition	22
Translation	20
Embedding	7
Segmentation	8
Retrieval	2
Open vocabulary detection	2
Open vocabulary segmentation	2
Tracking	1
Forecasting	5



Tokens

A token is the smallest unit of text that a language model can process. It could be a word, a punctuation mark, or even a subword unit. For example, the sentence "The quick brown fox jumps over the lazy dog" would be broken down into the following tokens: "The", "quick", "brown", "fox", "jumps", "over", "the", "lazy", "dog".



Gemini

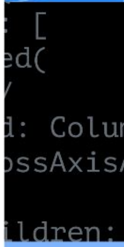
✦ gemini-1.5-flash-001

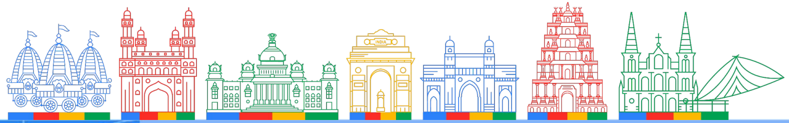
Superior speed and efficiency with 1M context window

✦ gemini-1.5-pro-001

Versatile and top-tier quality with up to 2M context window

✦ gemini-1.0-pro-002

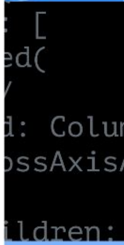




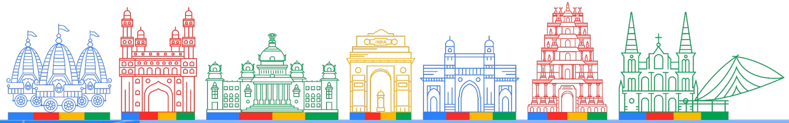
Embeddings

- An embedding is a numerical representation of a token.
- It's a vector (a list of numbers) that captures the semantic and syntactic meaning of the token.
- Words with similar meanings tend to have similar embeddings.

Word Embedding: "What is AI " - [0.42569, 0.82569, 0.385236, ...]



```
: [
ed(
/
d: Colu
ossAxis
ildren:
```

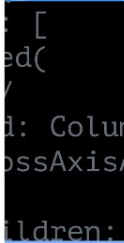


Embeddings

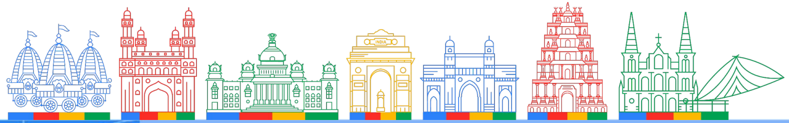
- Let’s consider an example with word embeddings.
- Suppose we have the words "king," "queen," "man," and "woman."
- In the embedding space, these words might have the following vectors:
  "king" == [0.5, 1.2, -0.3]
  "queen" == [0.6, 1.1, -0.2]
  "man" == [0.4, 0.8, -0.5]
  "woman" == [0.5, 0.9, -0.4]

These vectors allow us to perform arithmetic operations to uncover relationships, such as:

"king" − "man" + "woman" ≈ "queen"

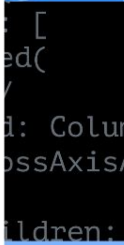


```
: [
ed(
/
d: Colu
ossAxis
ildren:
```



Vector DB

A vector database is a specialized database designed to efficiently store, manage, and retrieve high-dimensional vector data. Unlike traditional databases that primarily deal with structured tabular data, vector databases excel at handling data points represented as vectors in a multi-dimensional space.



```
: [
ed(
/
d: Colu
ssAxis
ildren:
```

Vector DB

Key Functionalities

- Vector Storage: Efficiently stores high-dimensional vectors.
- Indexing: Creates indexes to optimize search performance.
- Similarity Search: Quickly finds vectors similar to a given query vector.
- Scalability: Handles large volumes of data and high query rates.

Milvus
zilliz
Pinecone
vespa
marqo
qdrant
Weaviate
Chroma
Vald
LanceDB

https://www.linkedin.com/pulse/choosing-vector-database-your-gen-ai-stack-abhinav-srivastava/

Temperature

Controls the randomness of the output.

Higher temperature: Increases randomness, leading to more diverse and creative outputs but potentially less coherent text.
Lower temperature: Decreases randomness, resulting in more focused and deterministic outputs but potentially less creative.

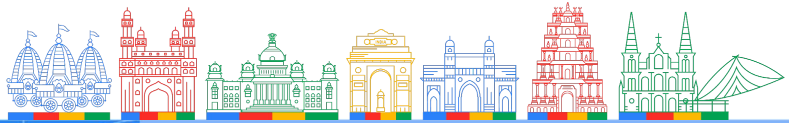
Top_k

Limits the number of possible next tokens to the top k most probable ones at each generation step.

Higher top_k: Increases the diversity of output by considering more options.
Lower top_k: Reduces the diversity but can improve focus and coherence.

Top_p

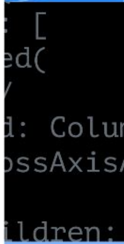
Definition: Selects the next token from the set of tokens whose cumulative probability exceeds a certain threshold (top_p).
Higher top_p: Increases the diversity of output by considering a larger set of tokens.
Lower top_p: Reduces diversity but can improve focus and coherence.



What is Vertex AI Agents

Vertex AI Agent Builder is a suite of tools from Google Cloud that simplifies building and deploying Gen AI agents. It caters to developers of all experience levels, offering

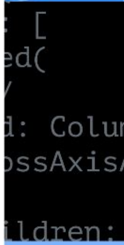
- Easily build no code conversational AI agents
- Ground in Google search and/or your enterprise data with our RAG offerings
- Rapidly create low-code to high-code AI application



```
: [
ed(
/
d: Colu
ssAxis
ildren:
```

Introducing LangChain

- An Open Source modular framework for building applications powered by language models
- Chatbots and virtual assistants
- Text generation and summarization
- Document question answering
- Relies on language models to reason
- Connects a language model to sources of context (prompts, contextual content)
- Combines components (language models, agents,
- memory stores) into complex workflows



```
: [
ed(
/
d: Colu
ssAxis
ildren:
```

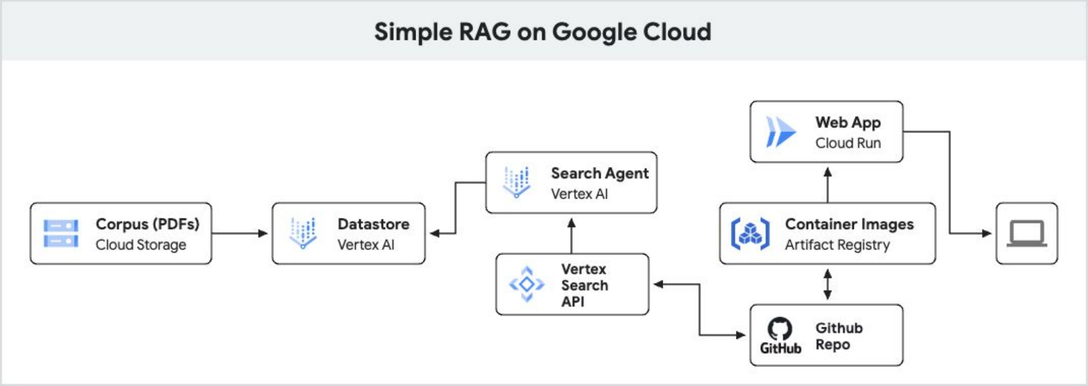
Google Developer Groups
Cloud • Rajkot

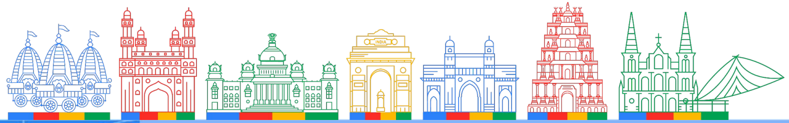
Cloud Community Day 2024

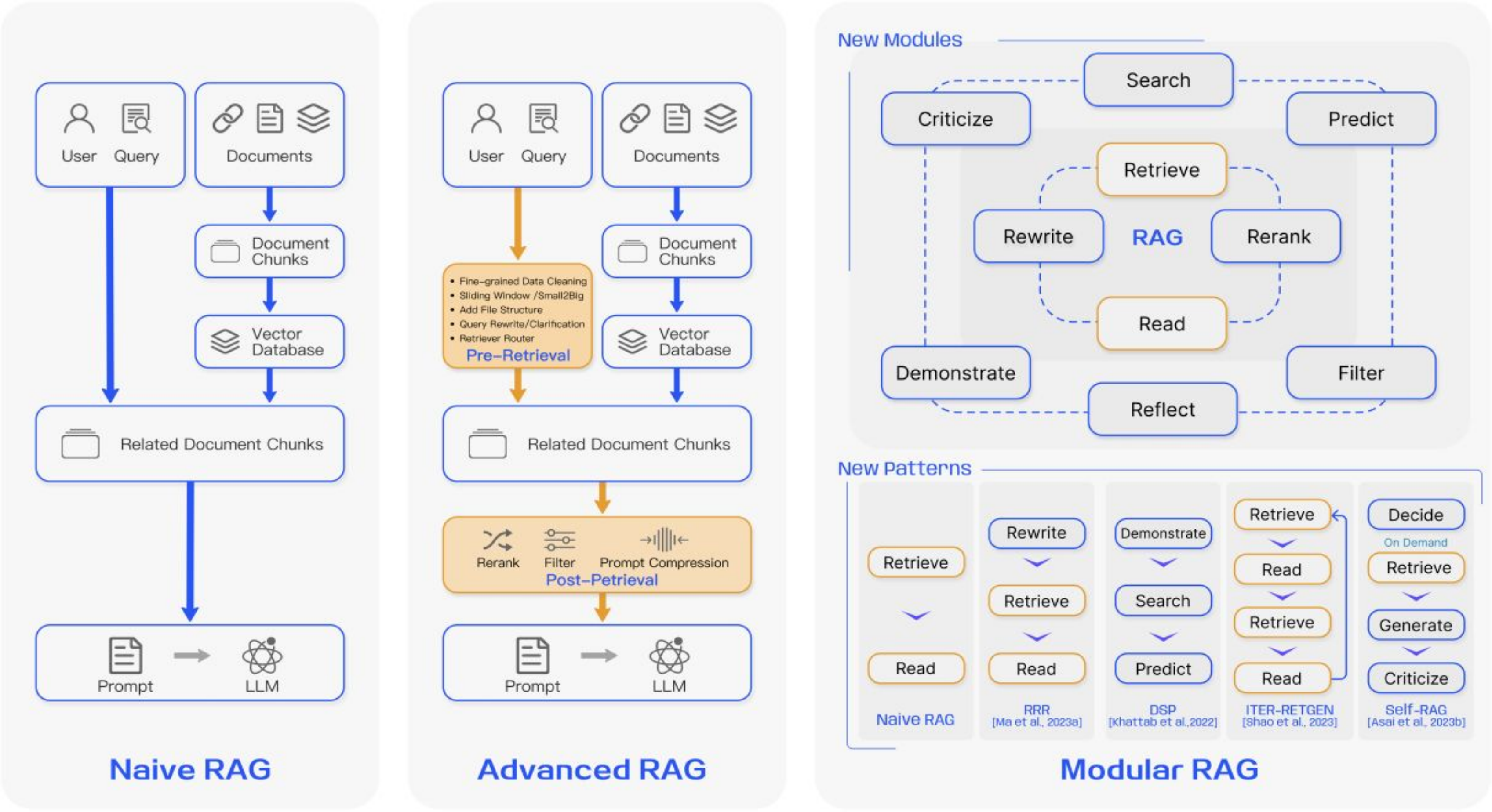
Source:https://www.langchain.com/

Why we need Retrieval-Augmented Generation (RAG)

- Enhanced accuracy
- Contextual responses
- Reduced hallucinations
- Dynamic knowledge updates
- Cost efficiency


Simple RAG on Google Cloud





User Query

Documents

Document Chunks

Vector Database

Related Document Chunks

Prompt

LLM

Naive RAG

User Query

Documents

Document Chunks

Vector Database

- Fine-grained Data Cleaning
- Sliding Window /Small2Big
- Add File Structure
- Query Rewrite/Clarification
- Retriever Router

Pre-Retrieval

Related Document Chunks

Rerank

Filter

Prompt Compression

Post-Petrieval

Prompt

LLM

Advanced RAG

New Modules

Search

Criticize

Predict

Retrieve

Rewrite

RAG

Rerank

Read

Demonstrate

Filter

Reflect

New Patterns

Retrieve

Read

Naive RAG

Rewrite

Retrieve

Read

RRR [Ma et al., 2023a]

Demonstrate

Search

Predict

DSP [Khattab et al.,2022]

Retrieve

Read

Retrieve

Read

ITER-RETGEN [Shao et al., 2023]

Decide

On Demand

Retrieve

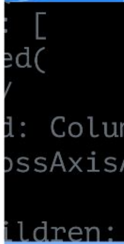
Generate

Criticize

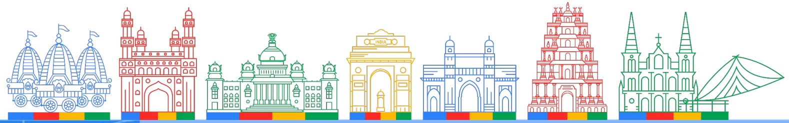
Self-RAG [Asai et al., 2023b]

Modular RAG

Feature	Vertex AI Agent with RAG-based Chatbot	Vector Search Solution	Vertex AI Embedding with Vector DB
Overview	No code -low code tool	Implements a search solution using vector-based retrieval	Combines Vertex AI Embedding with a vector database to enhance search and recommendations
Components Used	Vertex AI Agent, Dialogflow, Cloud Run	Vertex AI Vector Search , Cloud Run	Vertex AI Embedding, Cloud Run / Compute Engine
Scalability	High, handles large datasets and multiple interactions concurrently	High, optimized for large-scale search operations	High, supports large-scale embedding and search operations
Customization	Low, customizable conversation flows and response logic	Moderate, focuses on search and retrieval customization	High, flexible embedding and search parameter tuning
Cost Efficiency	Moderate, depends on usage and integration complexity	High, cost-effective for large-scale search operations	High, optimized for cost-efficient embedding and search solutions



```
: [
ed(
/
d: Colu
ossAxis

ildren:
```



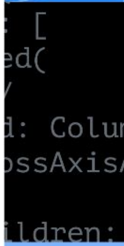
Popular Python Framework for Quick Development

Streamlit

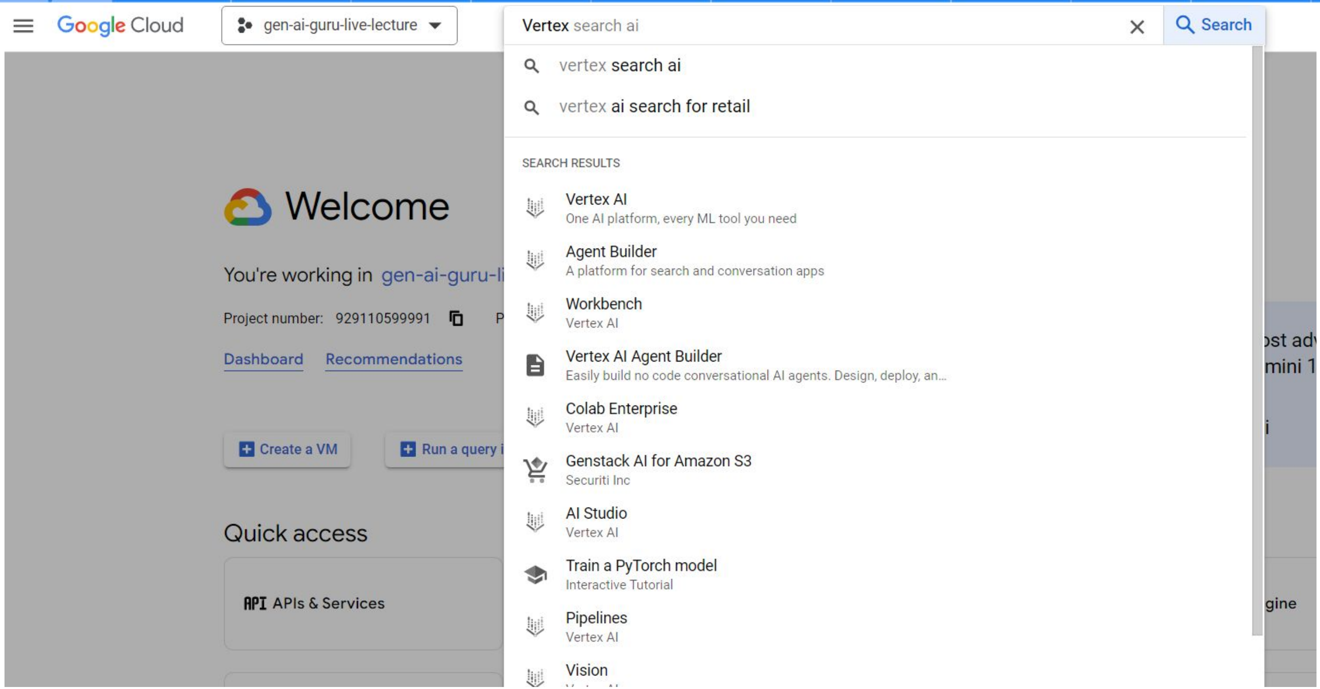
gradio

Chainlit

mesop



```
: [
ed(
/
d: Colu
ossAxis

ildren:
```



Vertex search ai

vertex search ai

vertex ai search for retail

SEARCH RESULTS

Vertex AI
One AI platform, every ML tool you need

Agent Builder
A platform for search and conversation apps

Workbench
Vertex AI

Vertex AI Agent Builder
Easily build no code conversational AI agents. Design, deploy, an...

Colab Enterprise
Vertex AI

Genstack AI for Amazon S3
Securiti Inc

AI Studio
Vertex AI

Train a PyTorch model
Interactive Tutorial

Pipelines
Vertex AI

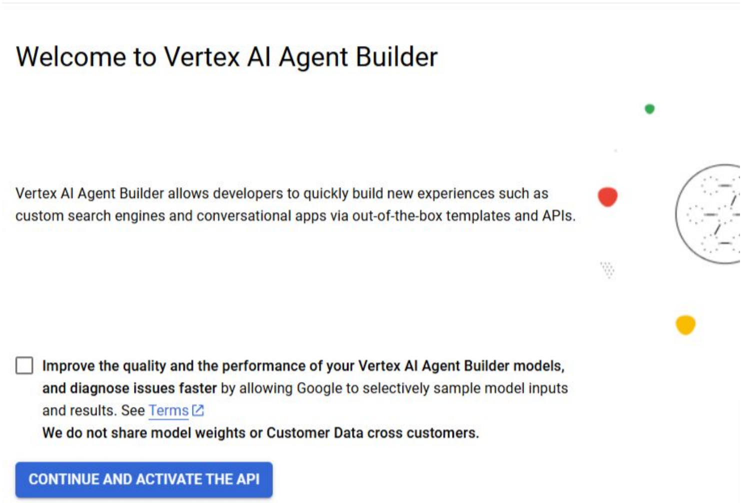
Vision

Welcome

You're working in gen-ai-guru-li

Project number: 929110599991

Enabling Vertex AI API



Welcome to Vertex AI Agent Builder

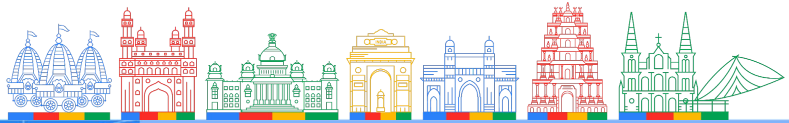
Vertex AI Agent Builder allows developers to quickly build new experiences such as custom search engines and conversational apps via out-of-the-box templates and APIs.

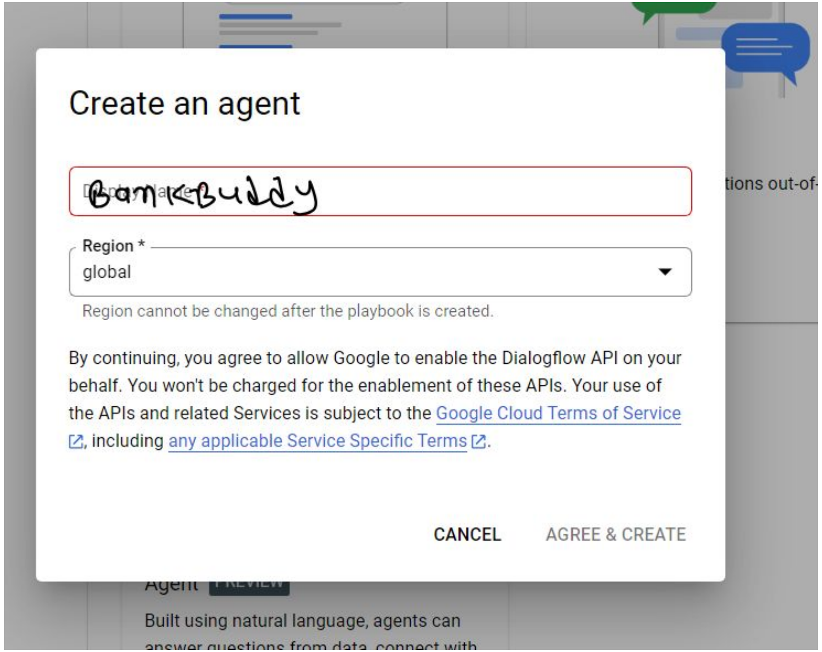
☐ **Improve the quality and the performance of your Vertex AI Agent Builder models, and diagnose issues faster** by allowing Google to selectively sample model inputs and results. See Terms

We do not share model weights or Customer Data cross customers.

Google Developer Groups
Cloud • Rajkot

Cloud Community Day 2024





Create an agent

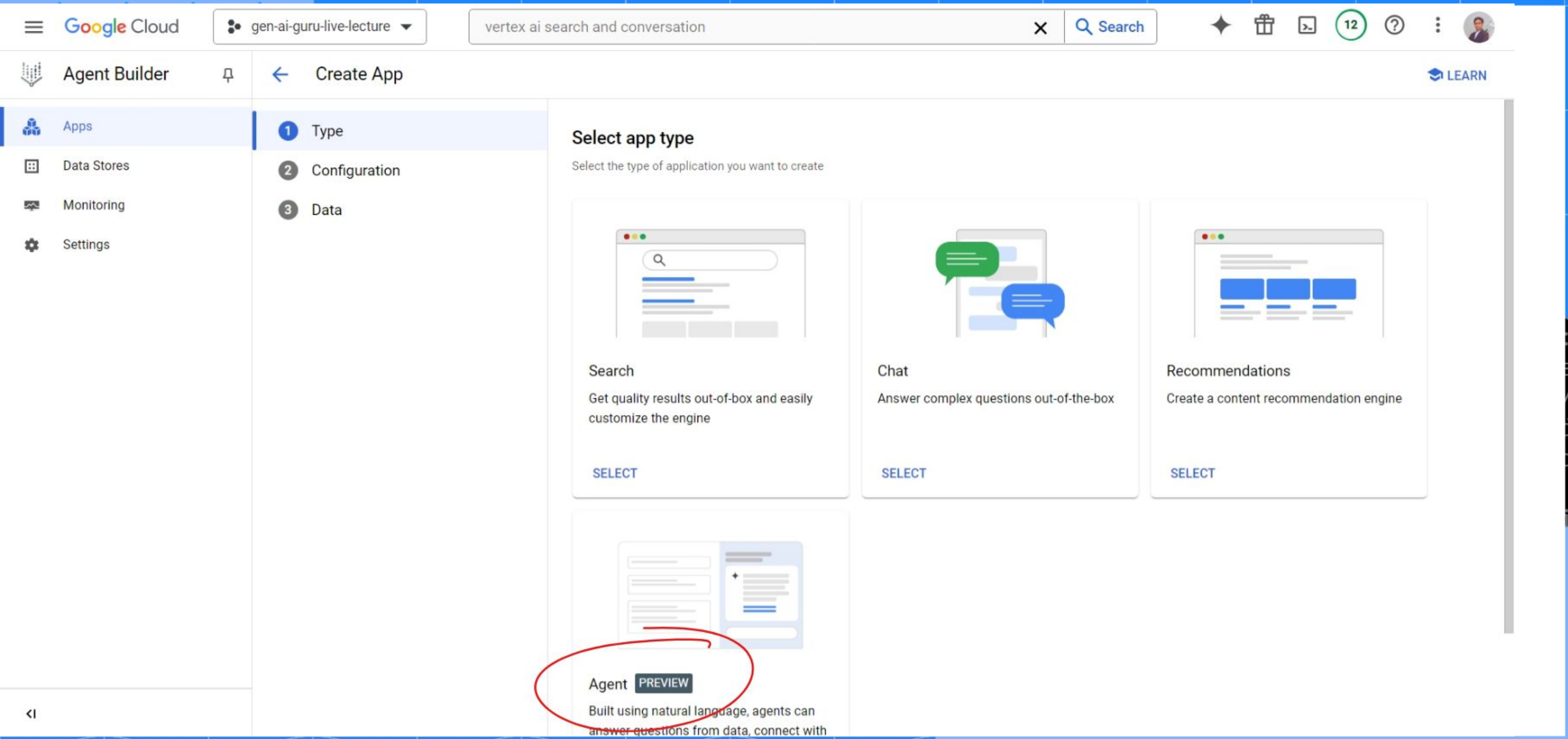
Display name: BankBuddy

Region *: global

Region cannot be changed after the playbook is created.

By continuing, you agree to allow Google to enable the Dialogflow API on your behalf. You won't be charged for the enablement of these APIs. Your use of the APIs and related Services is subject to the Google Cloud Terms of Service, including any applicable Service Specific Terms.

Built using natural language, agents can answer questions from data, connect with



Select app type

Select the type of application you want to create

Search

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

Chat

Answer complex questions out-of-the-box

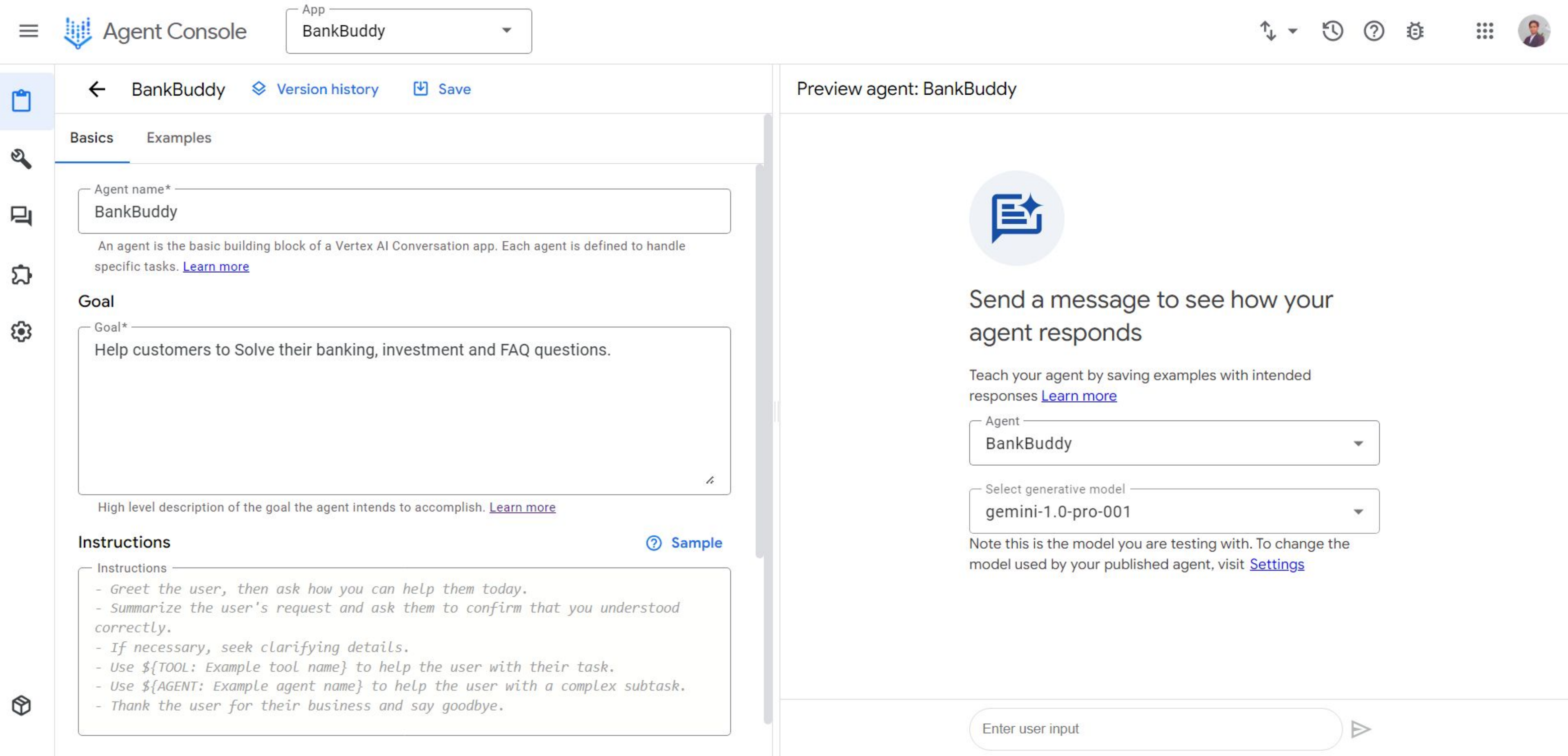
Recommendations

Create a content recommendation engine

Agent PREVIEW

Built using natural language, agents can answer questions from data, connect with

```
: [
ed(
/
d: Colu
ossAxis
ildren:
```



Agent name*

BankBuddy

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*

Help customers to Solve their banking, investment and FAQ questions.

High level description of the goal the agent intends to accomplish. Learn more

Instructions

Instructions

```
- Greet the user, then ask how you can help them today.
- Summarize the user's request and ask them to confirm that you understood
correctly.
- If necessary, seek clarifying details.
- Use ${TOOL: Example tool name} to help the user with their task.
- Use ${AGENT: Example agent name} to help the user with a complex subtask.
- Thank the user for their business and say goodbye.
```

Preview agent: BankBuddy

Send a message to see how your agent responds

Teach your agent by saving examples with intended responses Learn more

Agent

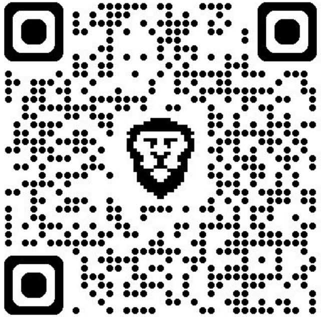
BankBuddy

Select generative model

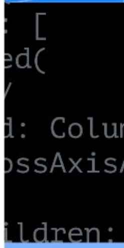
gemini-1.0-pro-001

Note this is the model you are testing with. To change the model used by your published agent, visit Settings

Live Demo



Code



```
: [
ed(
/
d: Colu
ossAxis

ildren:
```

Google Developer Groups
Cloud • Rajkot

Cloud Community
Day 2024

Demo -1

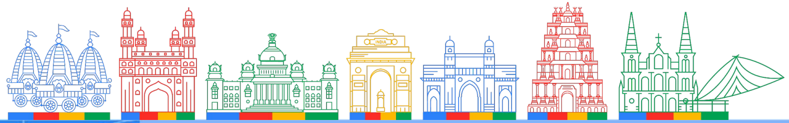
Vertex AI Agent with Agent Builder

Google Developer Groups
Cloud • Rajkot

Cloud Community Day 2024

Demo -2

Chainlit UI with Google Gemini API

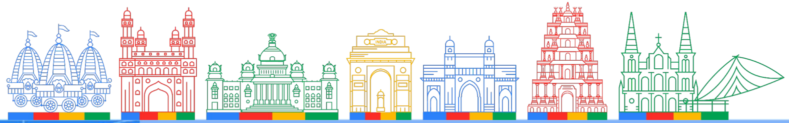


Demo -3

PDF Using Streamlit App and Gemini API

Google Developer Groups
Cloud • Rajkot

Cloud Community
Day 2024



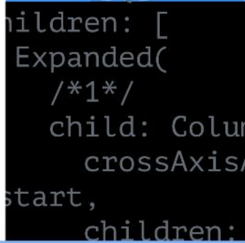
Google Developer Groups

Cloud • Rajkot

Cloud Community Day 2024

Thanks

https://linktr.ee/yashkavaiya



```
hildren: [
 Expanded(
   /*1*/
   child: Colu
     crossAxis
start,
     children:
```