

Breakfast & Learn

Generative AI on Google Cloud #go-further

Ready to take your GenAI journey to the next level?

Explore demo and code snippets for:

- Interacting with PaLM APIs, Vertex AI, and do prompt engineering
- Mastering foundation model tuning to elevate your ideas
- Developing a Search feature based on your own knowledge base!



Florent Machen
Consultant



Julien Landuré
CTO



Mardi 5 Décembre
9h30 - 11h



Station F, in-person
Google for Startups Room



French

Julien Landuré



CTO Zenika

Google Cloud 4x certified



Official Trainer

GDE Cloud



@jlandure



Florent Machen



Consultant Formateur

Google Cloud 2x certified



Official Trainer



@fmachen



Le Partenariat

Une offre complète !

Consulting / Professional Services :

- Consulting / Audit / Réalisation
- Consultants certifiés
- Possibilité de bénéficier de crédits dans certains cas

Training :

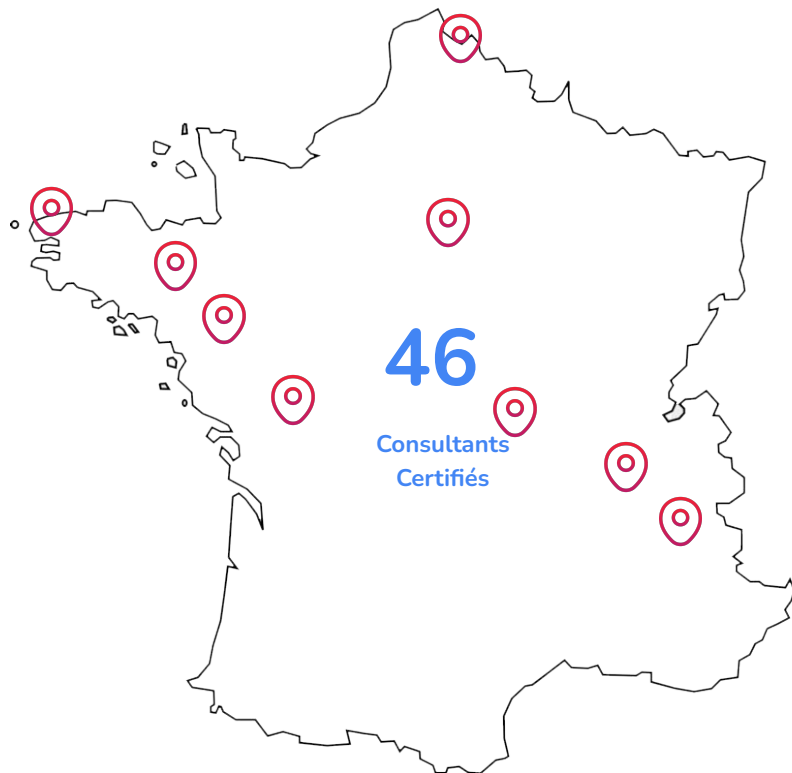
- Authorized Training Partners FR
- Formations officielles Google
- Support de l'éditeur

Resell :

- Offre de valeur associée au rattachement de votre billing account à Zenika
- Complémentaire à une offre projet



ZENIKA X GOOGLE CLOUD



Domaines d'expertise : Consulting x Formation

Move To cloud (Cloud Foundation)

Infrastructure (Kubernetes, VPC)

Cloud Provisioning (Terraform)

CI/CD (Build, CircleCI, GitHub, GitLab)

Mobile apps (Firebase/Firestore)

Data Expertise (Kafka, Elastic, BigQuery)

Gen AI (Vertex, BQ)

UN ACCOMPAGNEMENT GLOBAL DES PROJETS



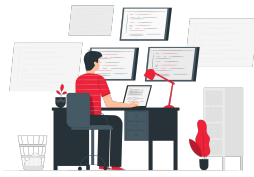
CONSEIL

Accompagner nos clients grâce à **ton expertise terrain** et à **une veille** réalisée en amont.



LAB

Faire partie d'une **équipe pluridisciplinaire et autonome, avec des infrastructures et des outils ainsi qu'un réseau de partenaires** pour créer un cadre qui va permettre de kickstarter les projets de nos clients et construire un MVP.



RÉALISATION

Mettre en oeuvre ton **savoir-faire technique** pour concrétiser les ambitions de nos clients.



FORMATION

Aider nos clients dans le développement de leurs compétences grâce à un **calendrier inter/intra entreprise** complet et proposer **des cycles personnalisés et sur mesure** en fonction de leurs besoins.



PROJET

Travailler sur des stacks innovantes, dans des environnements à la page. Tu pourras apporter **ton expertise** tout en évoluant dans un **environnement agile** et parmi un **large panorama de clients**.

ORGANISME DE FORMATION AGRÉÉ QUI DÉLIVRE LES FORMATIONS OFFICIELLES.

Authorized Training Partner

L'un des partenaires training autorisés en France. Reconnus par Google comme un organisme de formation de référence.

Nos consultants sont certifiés sur les cursus dispensés, et validés par Google comme **Authorized Trainer**



Google Cloud
Partner

BÉNÉFICE CLIENT

- | Échangez avec les formateurs, **experts sur le produit et les solutions GCP**
- | Bénéficiez **d'un accès privilégié aux contenus officiels Google**
- | Profitez d'un **accès post-formation à la plateforme Google dédiée**
- | Formez-vous **avec un partenaire de confiance**
- | Bénéficiez de retours d'expérience concret, et **d'un accompagnement post-formation**



Cloud AI gen AI Overview

October, 2023

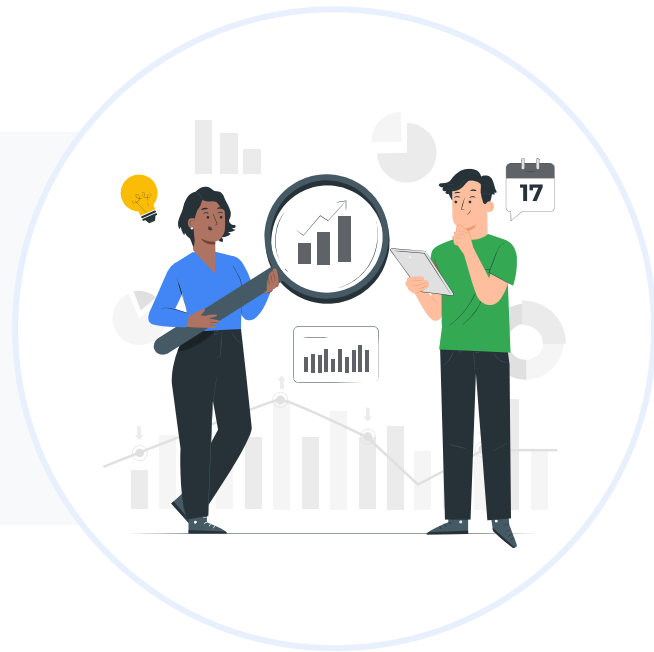


Table of Contents

- Primer on LLM and Generative AI 01
- What are Google's offerings? 02
- Labs 03
- What's next 04





PRIMER ON LARGE LANGUAGE MODELS & GENERATIVE AI

Artificial Intelligence

Any technique which enables computers mimic human behaviour

Chatbots,
Rule-based systems,
Robotics

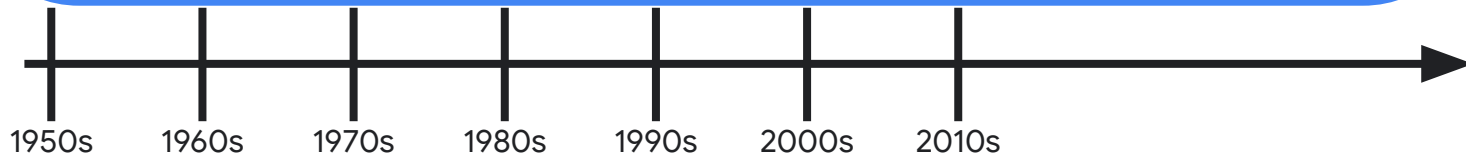
Machine Learning

AI techniques that give computers the ability to learn **without being explicitly programmed** to do so

Decision trees,
K-nearest neighbors,
Logistic regression

Deep Learning

A subset of ML, which make the computation of multi-layer neural networks feasible



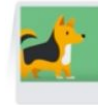
Generative AI is a subset of Deep Learning



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

Discriminative technique



Classify

Discriminative model
(classify as a dog or a cat)



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

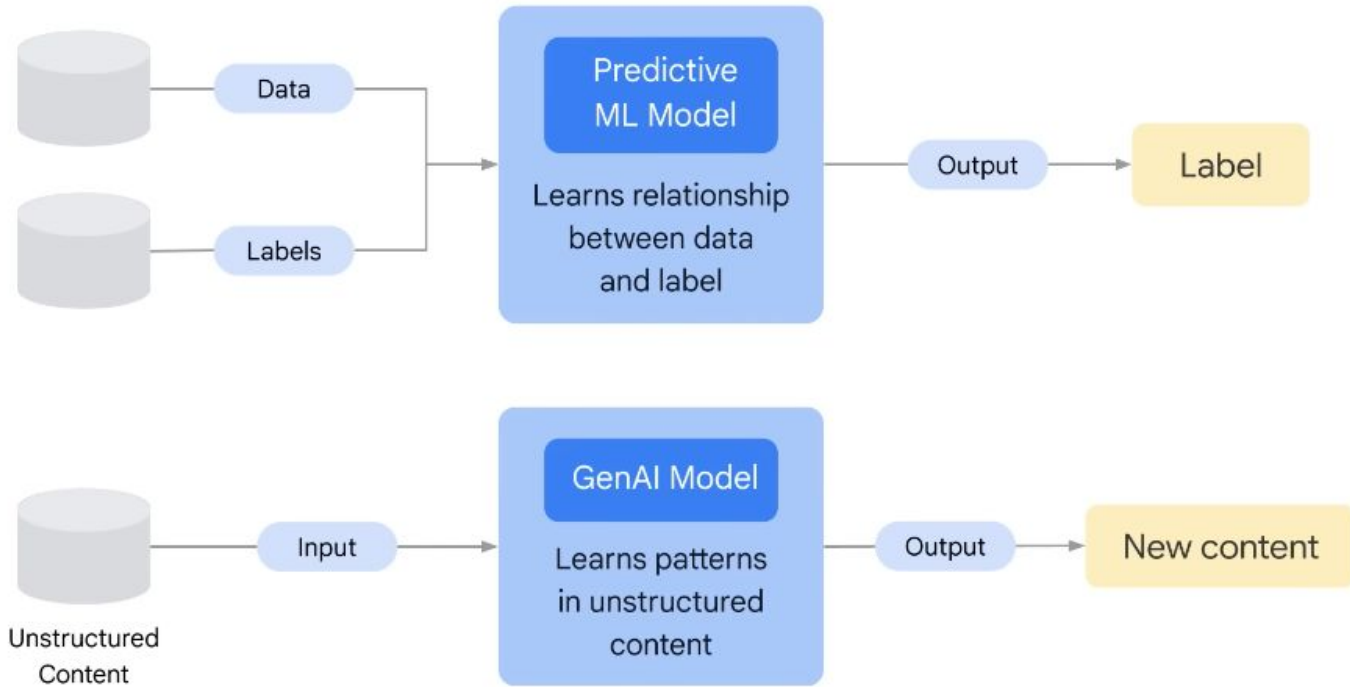
Generative technique



Generate

Generative model
(generate dog image)



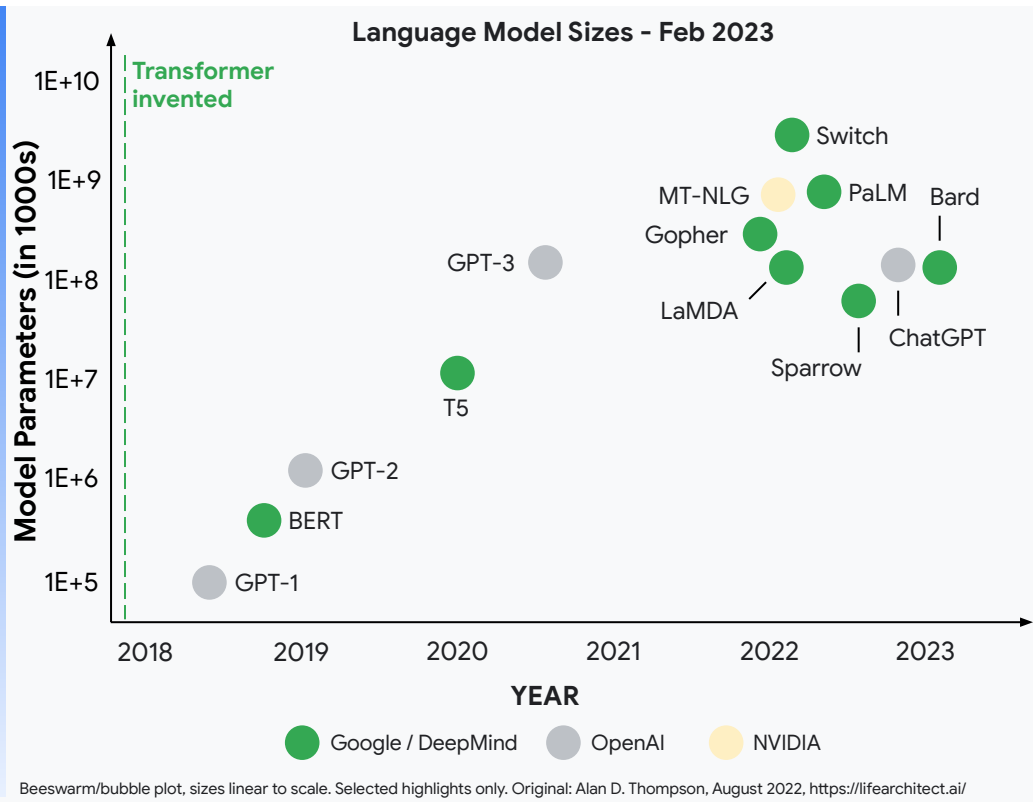


This revolution started at Google and we continue to innovate

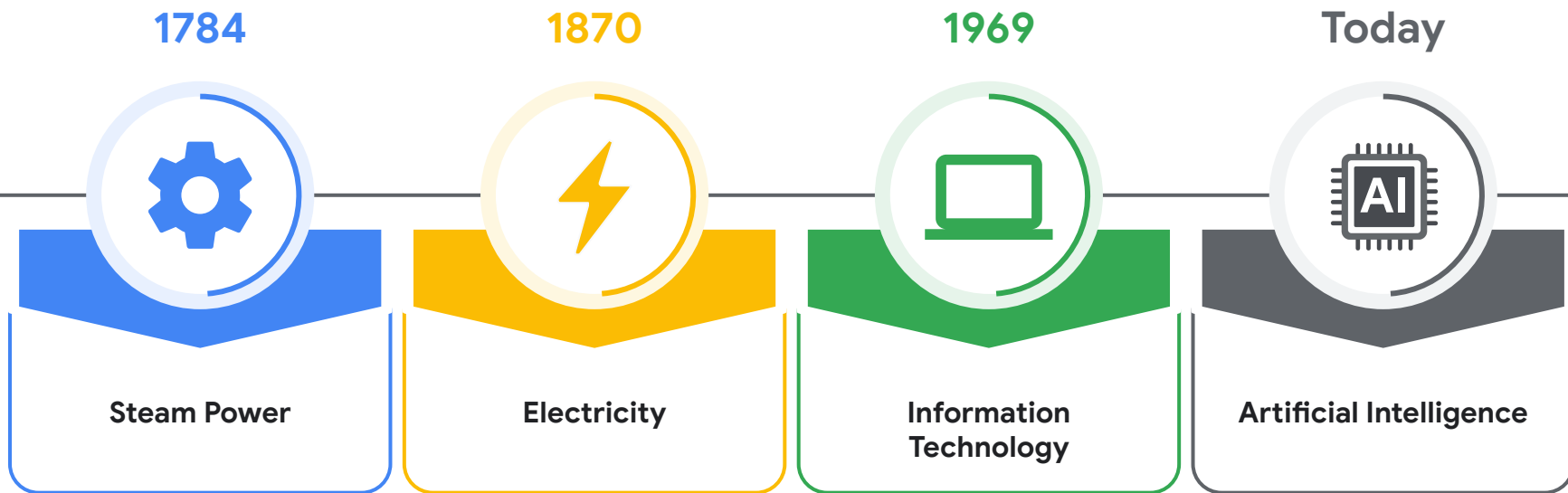
Google invented the Transformer architecture that powers Large Language Models in 2017.

Large Models already power Cloud AI products, including Contact Center AI, Document AI, NL Content Classification, and Google Search

3,000+ Google / Deepmind researchers and 7,000+ scientific publications continue to **drive our research to product pipeline**



We're in an AI-driven revolution



What are large language models?



ML algorithms that can **recognize, predict, and generate** human languages



Pre-trained on petabyte scale text-based datasets resulting in large models with **10s to 100s of billions of parameters**



LLMs are normally **pre-trained on a large corpus of text** followed by fine-tuning on a specific task



LLMs can also be called **Large Models** (includes all types of data modality) and **Generative AI** (a model that produces content)



Go read this huuuuuuge pile of books.



So, you've learned about cats and millions of other concepts ... what's a cat?

A cat is a small, domesticated carnivorous mammal.



Generative language models

LaMDA, PaLM, GPT-3, etc.

Why are large language models different?



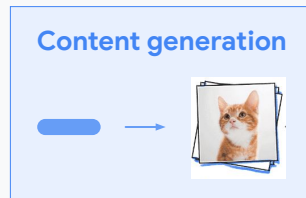
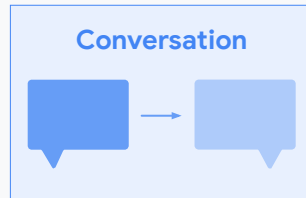
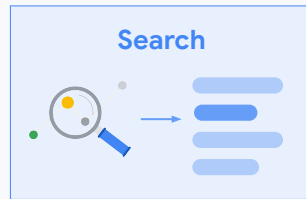
LLMs are characterized by **emergent abilities**, or the ability to perform tasks that were not present in smaller models.



LLMs contextual understanding of human language **changes how we interact** with data and intelligent systems.



LLMs can find patterns and connections in **massive, disparate data corpora**.



This revolution started at Google and we continue to innovate



2017
Transformer

Google invents
Transformer
kickstarting LLM
revolution



2018
BERT

Google's
groundbreaking
large language
model, BERT



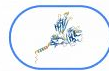
2019
T5

Text-to-Text
Transfer Transformer
LLM 10B P model open
sourced



2020
LaMDA

Google LaMDA
model trained to
converse



2021
AlphaFold

AlphaFold predicts
structures of all
known proteins



2022
PaLM

Industry leading
large language
model



2023
Bard

A conversational AI
Service powered by
LaMDA.

Responsible AI at the foundation

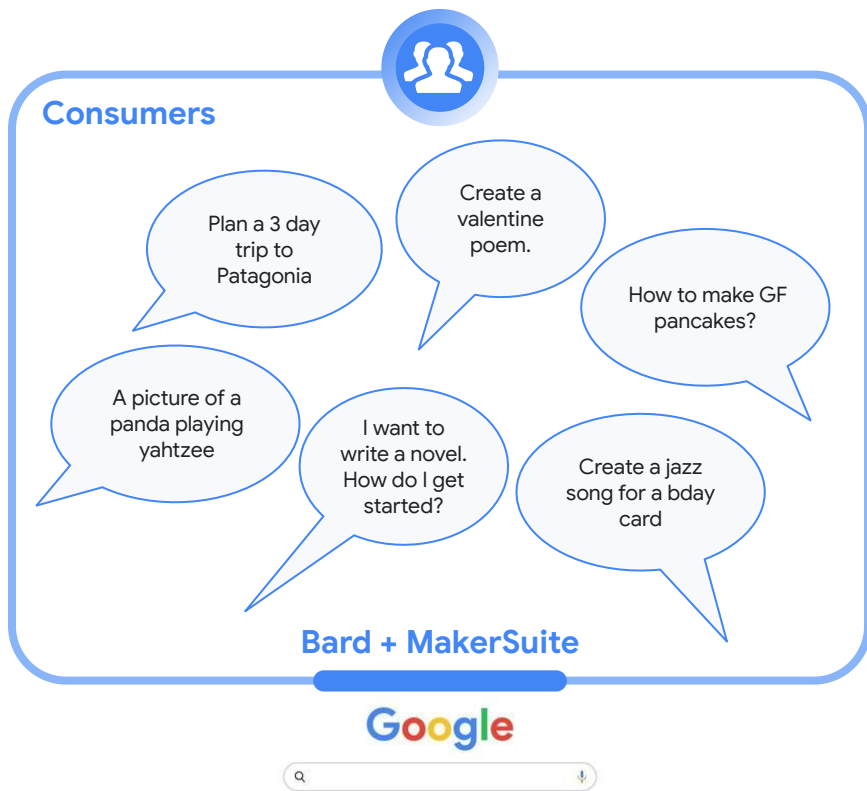


WHAT ARE GOOGLE'S OFFERINGS?

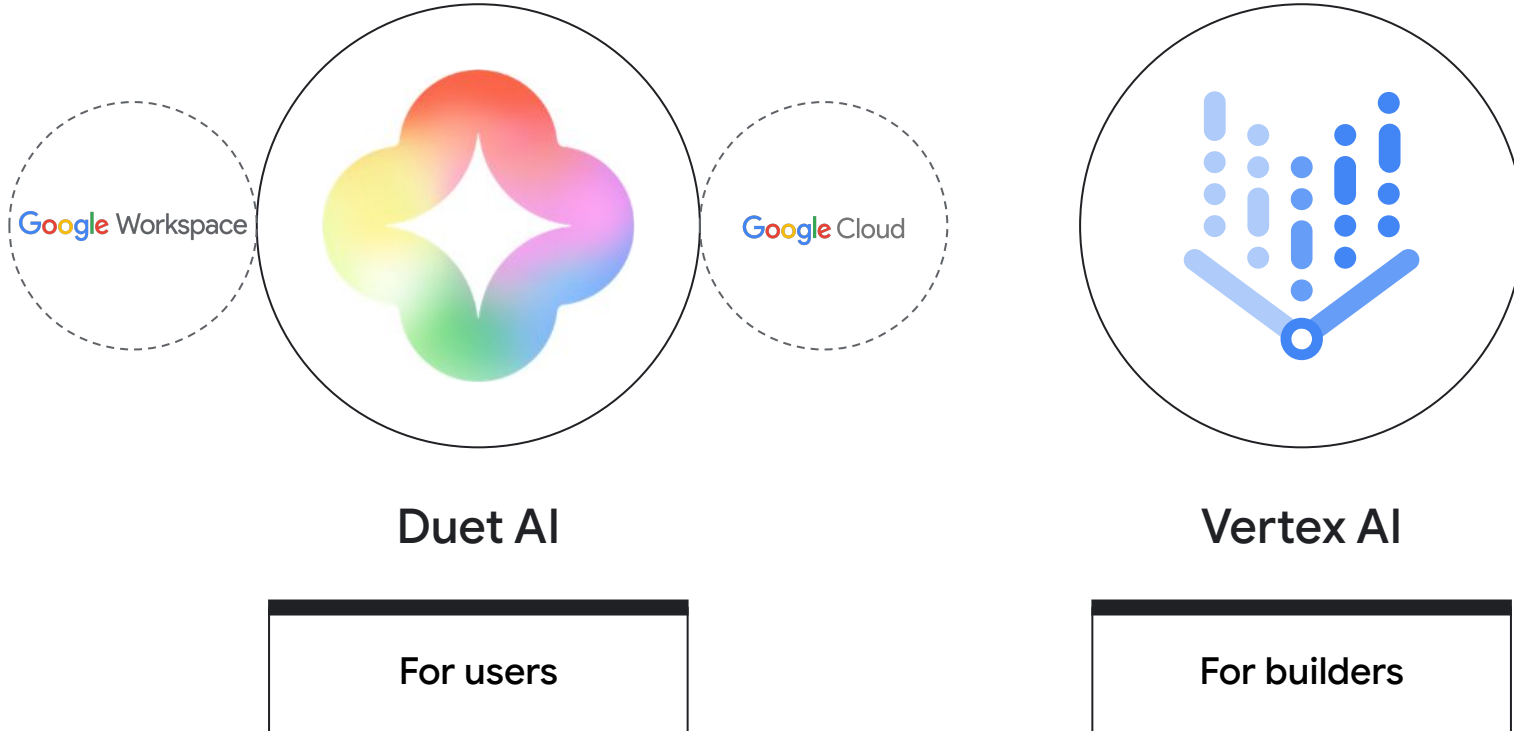
Google's AI Ecosystem



Consumers & enterprises have different needs....



Google's AI Ecosystem for Enterprises



Duet AI for Google Workspace Enterprise



Helps you write  
in Gmail and Docs

Duet AI works behind the scenes to help you write — whether it's refining existing work or helping you get started



Helps you visualize
in Slides 

With Duet AI, you can easily create images for presentations and meetings from a simple prompt



Helps you organize
in Sheets 

Duet AI is here to help you organize, classify and analyze your data faster than ever before



Helps you connect
in Meet 

Duet AI helps you look and sound your best on video calls so you can focus on the conversation

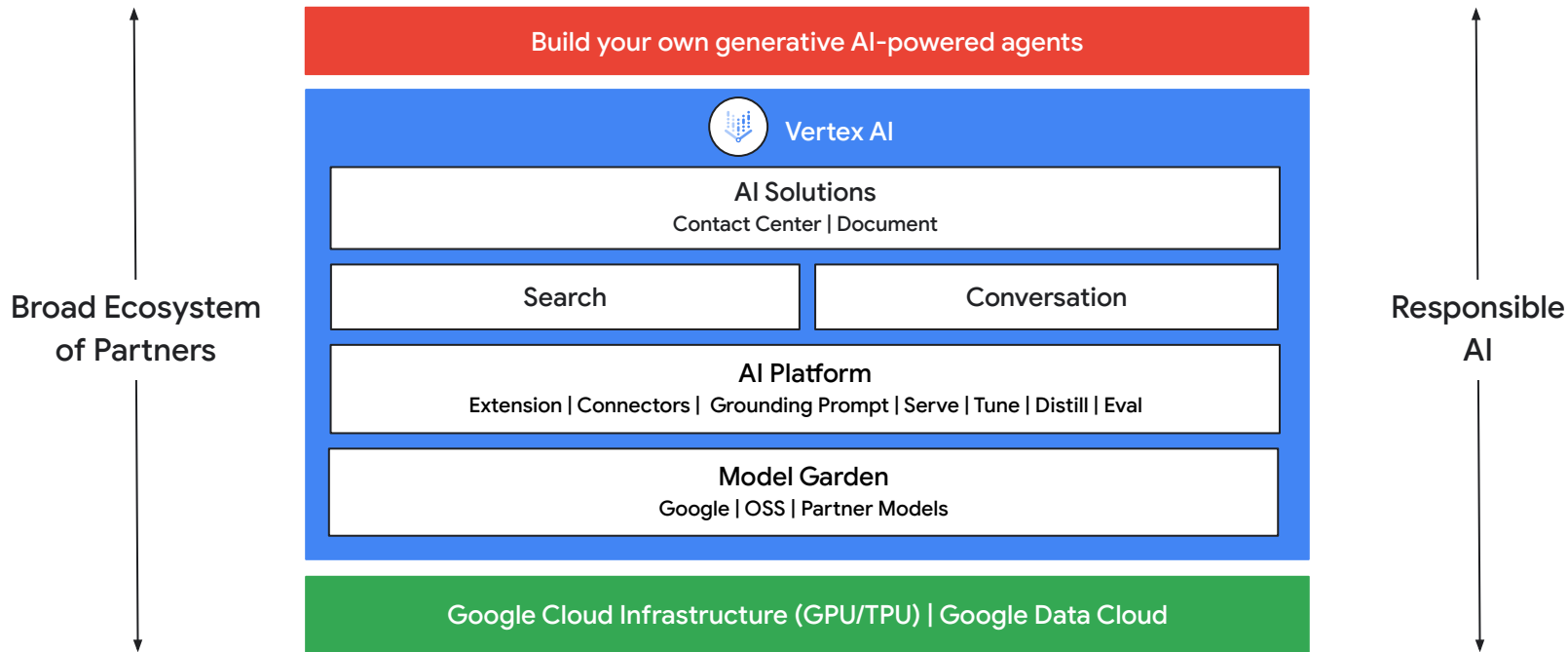
 Help me write an engaging headline

 DEMO



Google Cloud's Generative AI Portfolio

To support the needs of **Generative AI** centric enterprise development





Vertex AI

Foundation Models and Model Garden



Foundation Models

Across a variety of model sizes to address use cases



Now in GA

PaLM for Text

Custom language tasks



Now in GA

PaLM for Chat

Multi-turn conversations with session context



Now in GA

Imagen for Text to Image

Create and edit images from simple prompts



Now in GA

Embeddings API for Text and Image

Extract semantic information from unstructured data



Now in GA

Chirp for Speech to Text

Build voice enabled applications



Now in GA

Codey for Code Generation

Improve coding and debugging

PaLM for Text and Chat

GA

Custom language tasks in a few sentences

Zero/few-shot prompting with a gallery of pre-built prompts

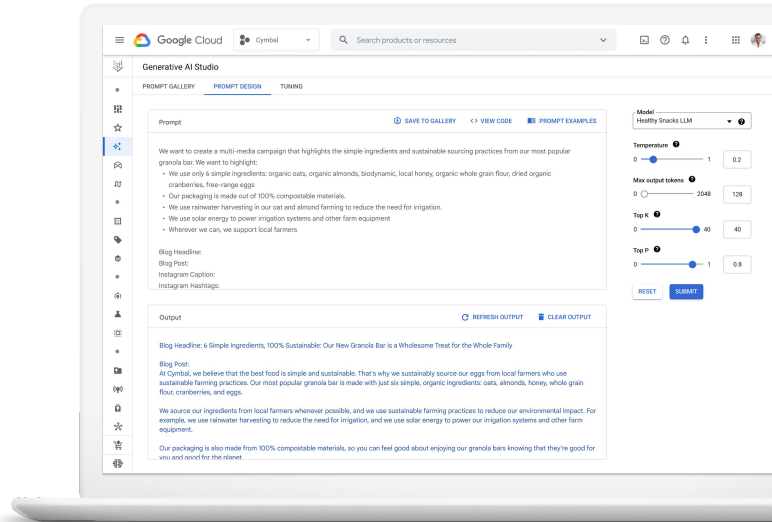
Multi-Turn Conversations w/ Preserved Context for Chat

Customize & Optimize PaLM on your data for specific tasks

With as few as 100 examples or human feedback

Use Cases:

Classification, sentiment analysis, entity extraction, extractive question answering, summarization, text rewrite in different style, ad copy generation, concept ideation, dialog, chatbots



New version

32k context windows (Public Preview) and language expansion (38 in GA)

Text-bison streaming & batch predictions (GA)

Adapter Tuning for Text (GA) and Chat (Public Preview) + RLHF (Public Preview)

Embeddings API for Text

GA

Extract semantic information from unstructured data

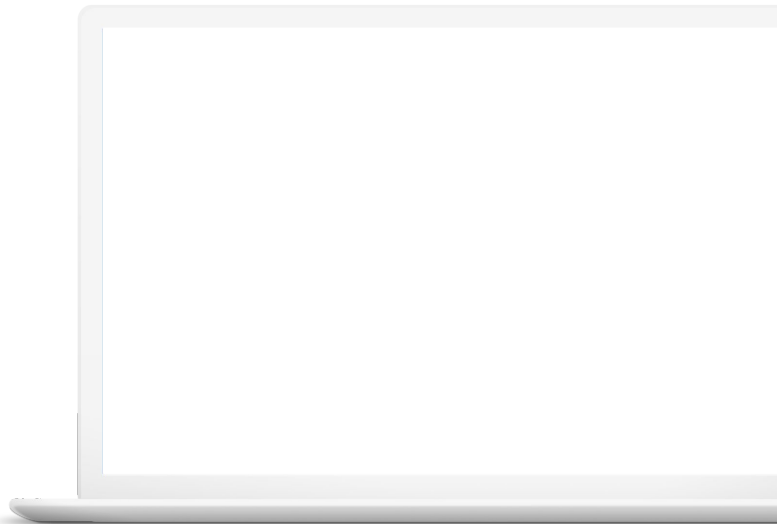
Variety of data modalities

Generate high quality embeddings for text (GA) and multi-modal data formats (Public Preview)

Combine with PaLM and Matching Engine to build sophisticated applications

Use Cases:

Improve recommendation engines, ad targeting systems, image classification, image search, & more



New version

Improved semantic matching for retrieval
Batch predictions (Private Preview)

Codey API for Code Assistance

GA

Accelerate software development

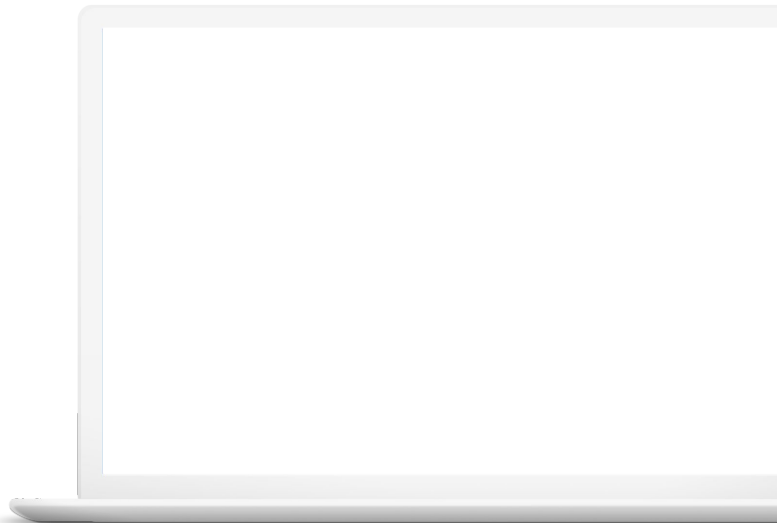
Real-time Code Completion and Generation

Code chatbot

Data privacy + recitation checks

Use Cases:

Text-to-Code tuned with your own codebase, improve developer velocity, troubleshoot and debug, and improve code quality



New version

Up to 25% quality improvement in major supported languages

32K context windows for longer form code (Public Preview)

Text2SQL for better SQL performance on top of the code models (Private Preview)

Adapter Tuning (Private Preview) with just a few hundred examples from your code base

Imagen for Text-to-Image

GA

Low latency and high quality image generation and editing

Generation, editing, captioning & classification

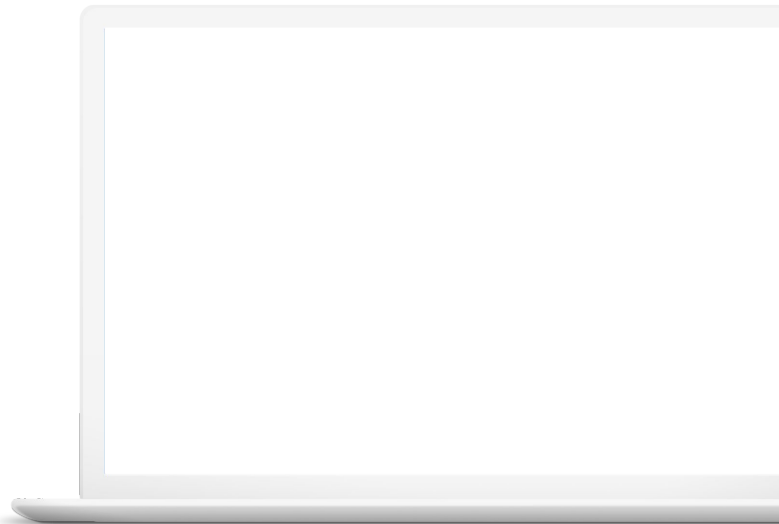
Style & Subject tuning to your brand (allowlisting)

Ready for enterprise-use

Images generated are the customer's data.

Use Cases:

Prototyping, marketing, avatars, ads



New version

Improvement in visual appeal for images, and latency

Experimental

Digital Watermarking

Label and verify AI-generated images produced by Imagen

 DEMO



Chirp for Speech to Text

GA

Build voice-enabled applications for global audiences

Apply the power of large models to Speech

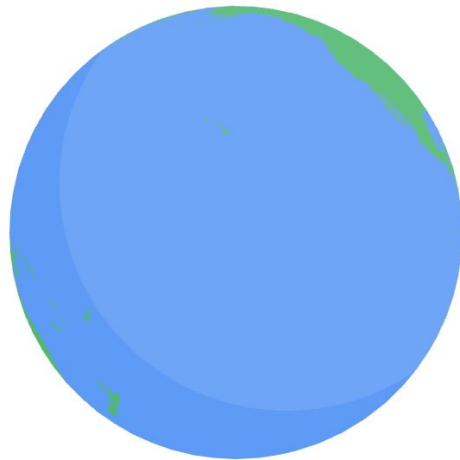
Self supervised training on millions of hours of audio to create the first 2B parameter speech model.

Improve speech recognition quality

Achieves 98% accuracy on English and over 300% relative improvement in tail languages.

Use Cases:

Captioning videos, voice assistants



Announcing at Next '23 - Experimental on Vertex AI

Experimental

Digital watermarking

Label and verify AI-generated images produced by Imagen

Invisible watermarking

Digital watermarking produces an invisible watermark on images generated by Imagen, which does not modify visible content or lower image quality

Tamper resistant watermarking

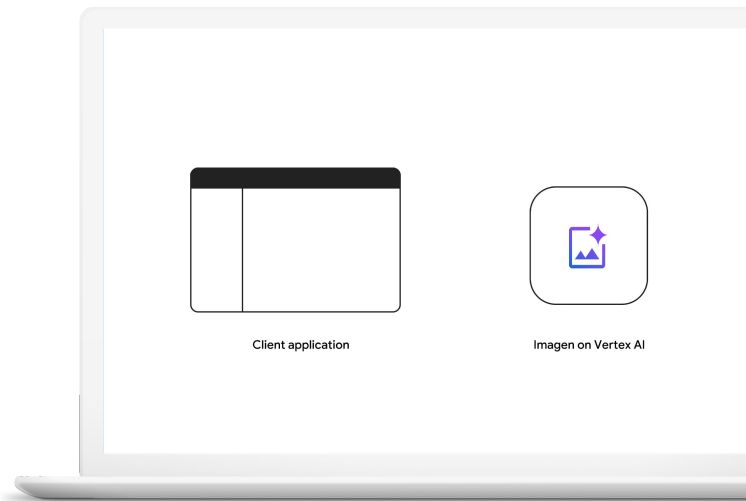
Vertex AI's watermarking technology is the first enterprise cloud offering to be embedded directly into the image pixels, making it difficult to tamper with

Watermark Verification

Digital watermarking provides a verification tool to check whether your image was generated by Imagen

Powered by Google Deepmind

Digital watermarking is powered by SynthID, a state-of-the-art technology from Google Deepmind



Google Cloud Vertex AI

The only cloud provider delivering open source and 1st/3rd party foundation models on a single AI platform





Vertex AI

Model Garden



Vertex AI Model Garden

One stop shop for ML and GenAI journeys

GA

Use **Foundation Models directly** for many use cases with pre-built templates

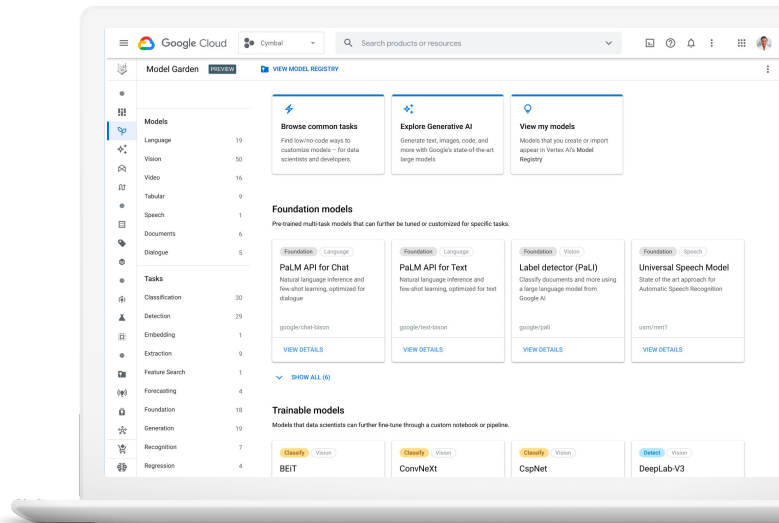
Tune Foundation Models with data & prompts for targeted industry or use cases

Customize popular open source models via Data Science Notebooks, Vertex AI Pipelines and Vertex AI MLOps toolkit

API access to task-specific solutions

New models!

Deploy and tune Llama 2 and Code Llama
Pre-announcing Anthropic Claude 2 model!





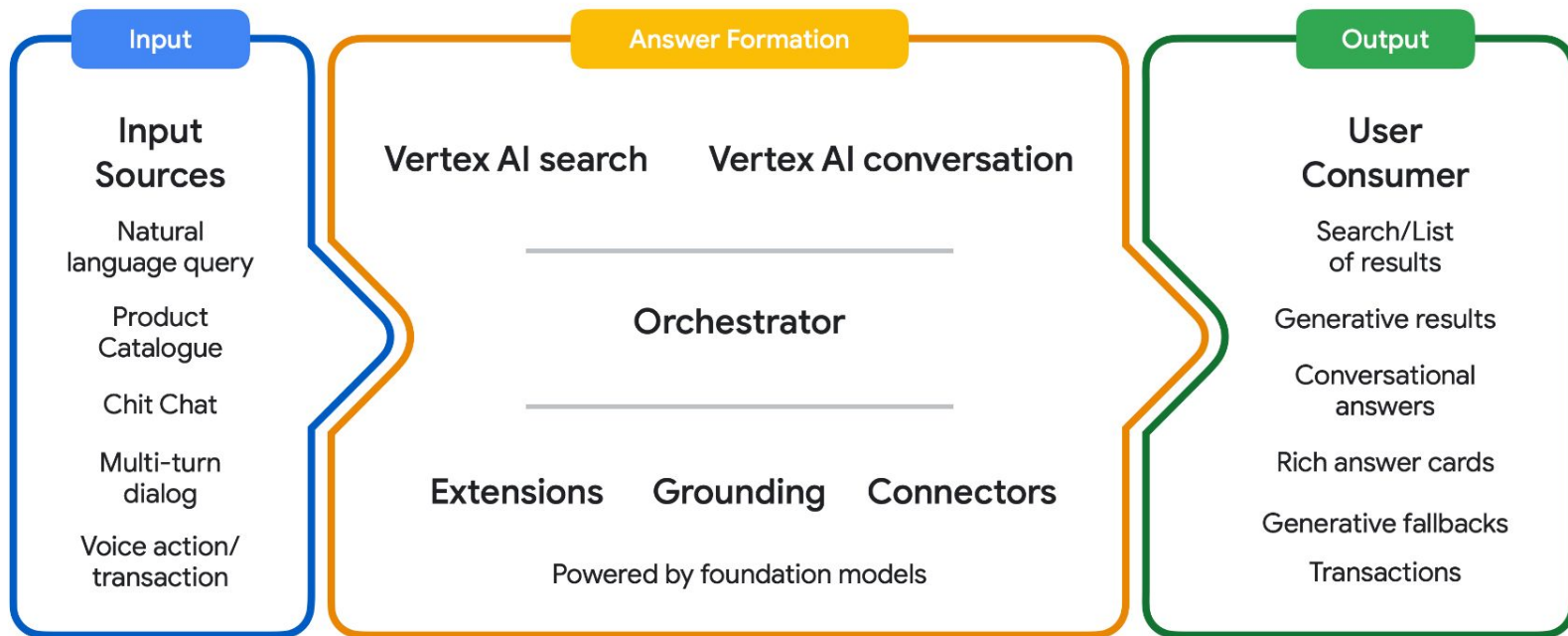
Vertex AI

Search and Conversation



Vertex AI Search & Conversation

General
availability



Announcing at Next '23 - Generally Available on Vertex AI

GA

Vertex AI Search

Deliver relevant, personalized search experiences in minutes, for enterprise apps or consumer-facing websites

Quickly set up Google quality **multi-modal**, **multi-turn search** experiences for customers and employees

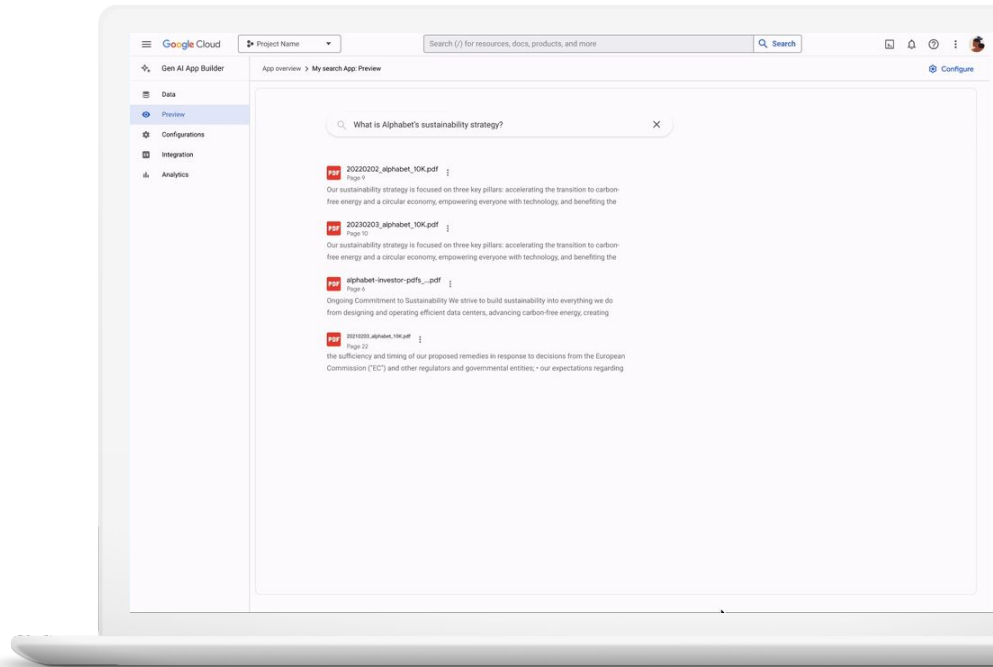
Ground search results in enterprise data or combine with LLM knowledge

Enterprise access controls ensure information is surfaced only to users who have access

Data connectors and extensions connect to backend systems to read data and take actions with 3rd party applications like Salesforce, etc.

Vector Search for highly scalable vector embedding search across different types of data

Helpful features like **summarization**, **citations**, **relevance scores**, **flexible LLM style prompting** help in knowledge discovery and customize search results



Announcing at Next '23 - Generally Available on Vertex AI

Vertex AI Conversation

GA

Build custom chat and voice bots powered by generative AI and grounded in your enterprise data

Combine **deterministic workflows with generative AI** to make conversations dynamic and personalized

Multi-modal support image and audio in addition to text

Train on a variety of data from websites, documents, FAQs, email and agent conversation history

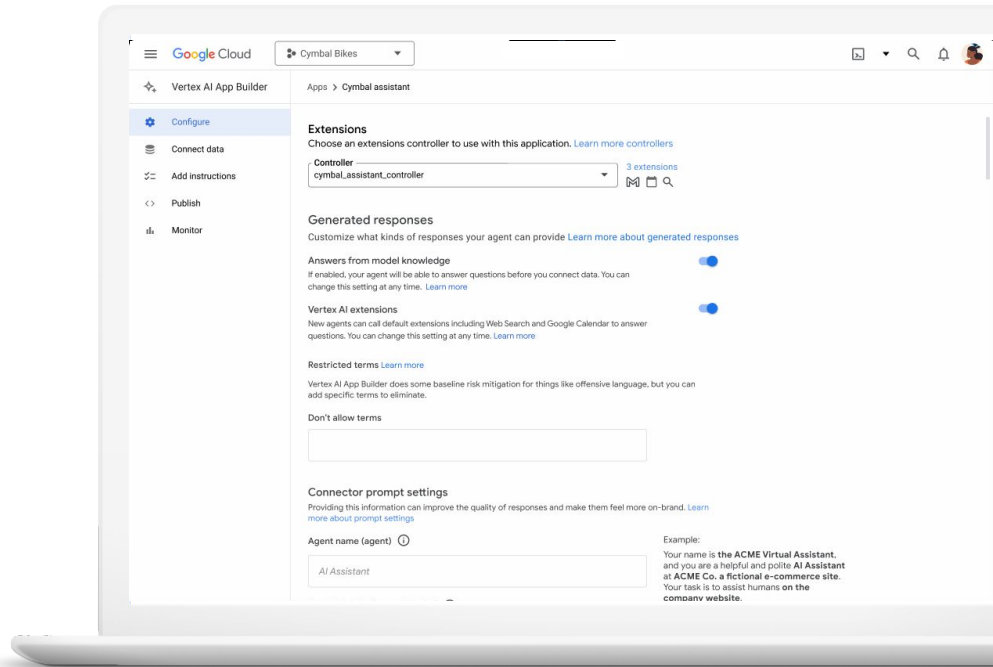
Enable transactions to connect to backend systems or 3rd party applications for complete transactions (bookings, payments etc)

Generators and **generative fallbacks** inject actions and pre-programmed conversations into workflows

Helpful features for customer service workflows summarization, generative FAQ, citations and agent assist



Fully integrates with existing
Contact Center AI offerings





WHAT'S NEXT?



Labs GenAI with Vertex AI

Generative AI with Vertex AI: Getting Started

bit.ly/stationf-genai-lab-dec-2023



