Google Developer Experts

Harnessing the powers of Gemini

Aashi Dutt
Google Developer Expert ML (Gen AI)

$whoami

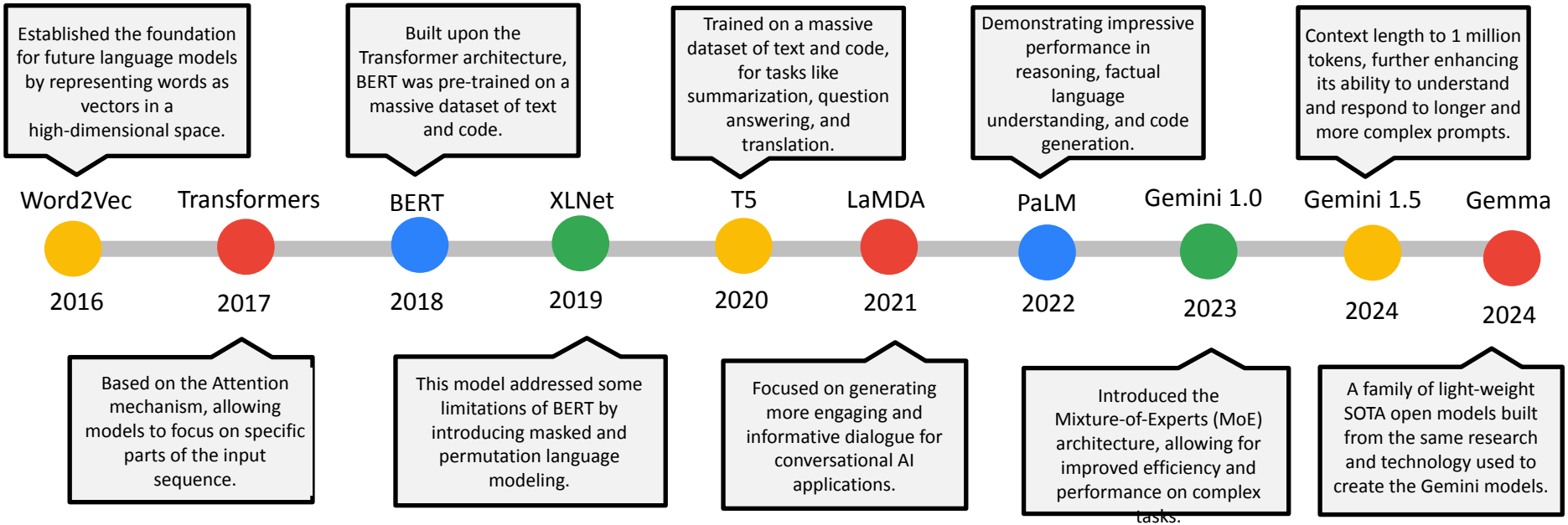
- Post-Grad Student at Georgia Institute of Technology
- Google Developer Expert in Machine Learning (Gen AI)
- Organizer @ TensorFlow User Group Chandigarh
- Co-founded, Health-tech startup
- 3X Kaggle Expert
- MIT Bootcamps Alumni
- Writer, blogs on AI/ML on Medium

Content

- What is Gemini?
- Architecture of Gemini
- Hands-on: Gemini on Colab and Google AI Studio
- Examples built with Gemini
- Resources

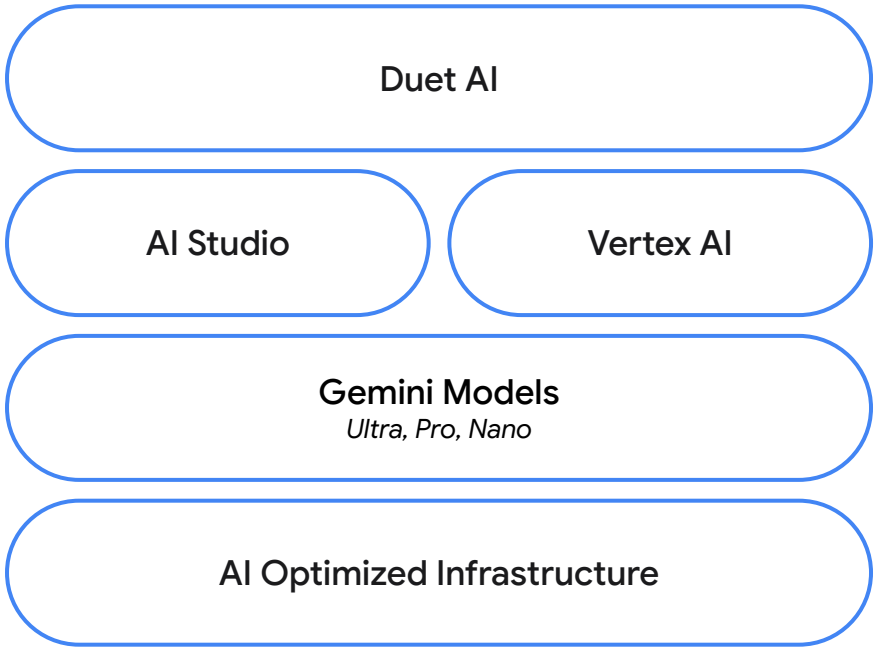
Slides courtesy- Nitin Tiwari ML GDE

Language models over the years



Experts

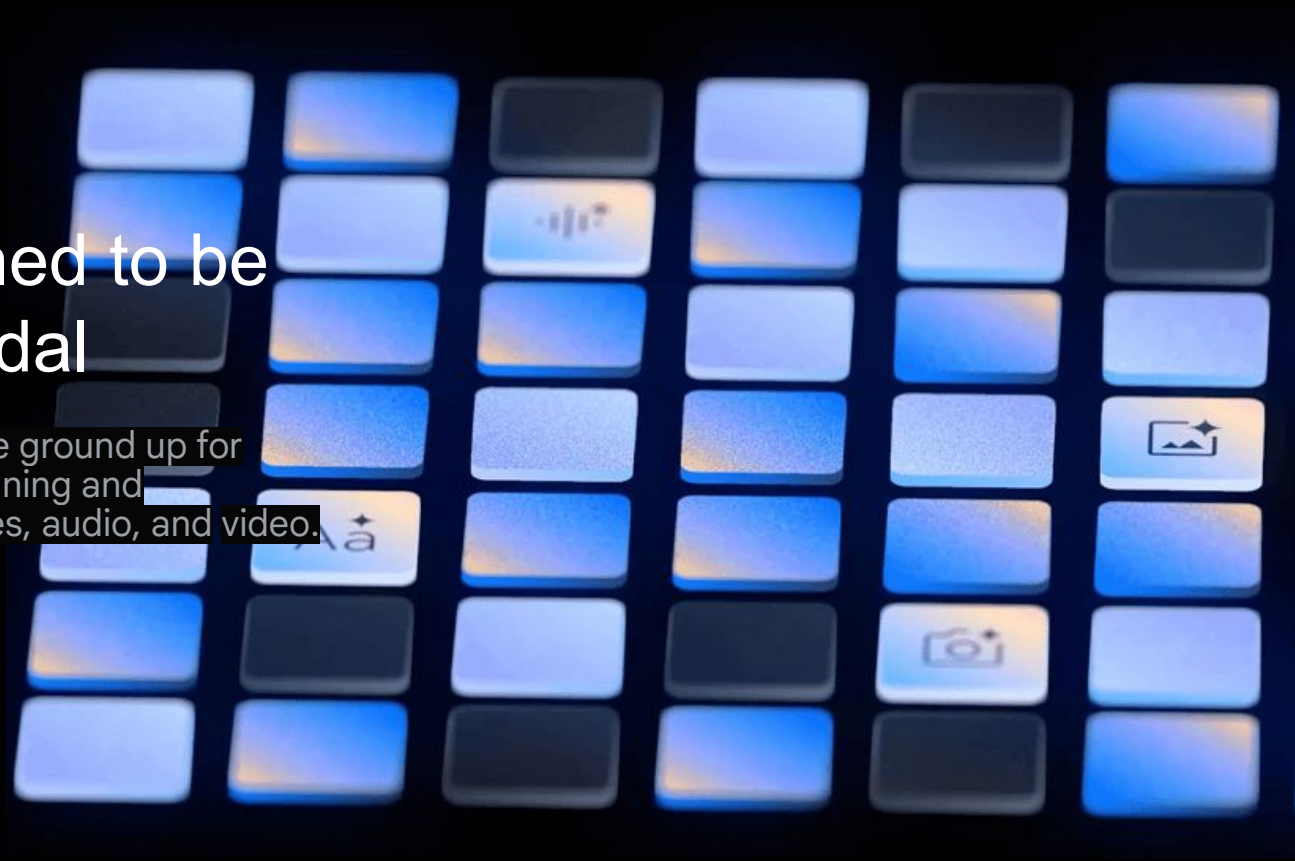
Generative AI is built on a vertically integrated technology stack



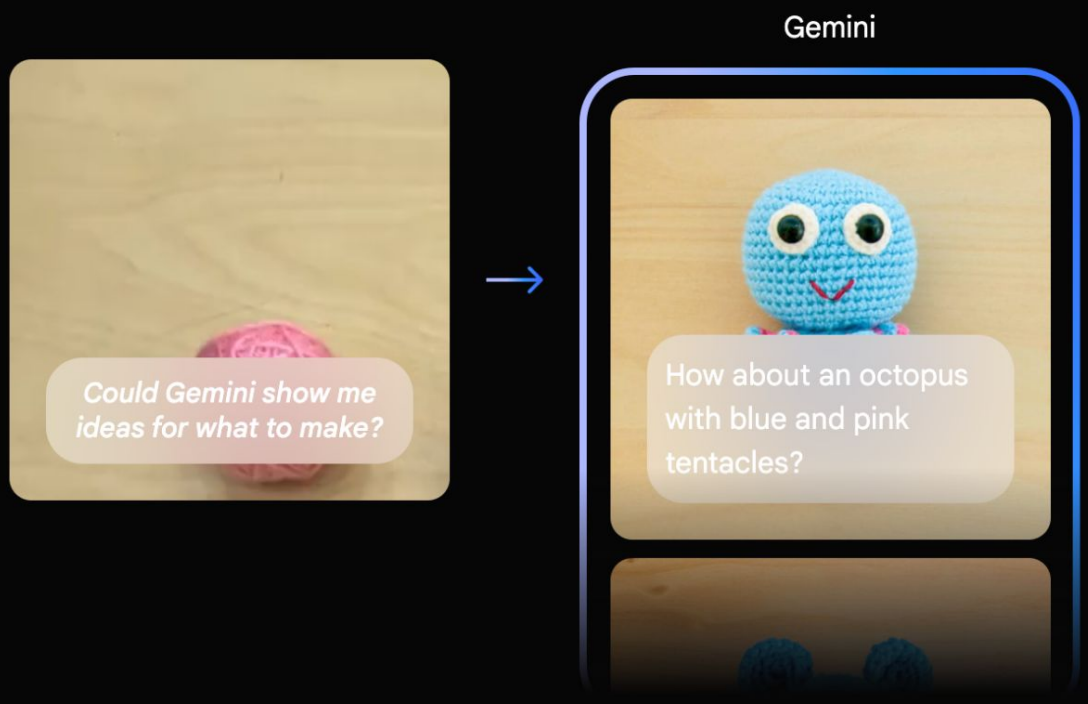

Gemini
Natively multimodal
Sophisticated reasoning
Advanced Coding

Gemini is designed to be natively multimodal

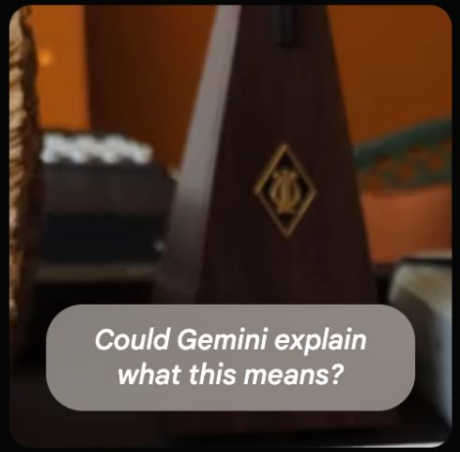
Gemini models are built from the ground up for multimodality, seamlessly combining and understanding text, code, images, audio, and video.



Gemini models can generate text and images, combined.



Gemini models can understand and perform tasks involving several different written languages.

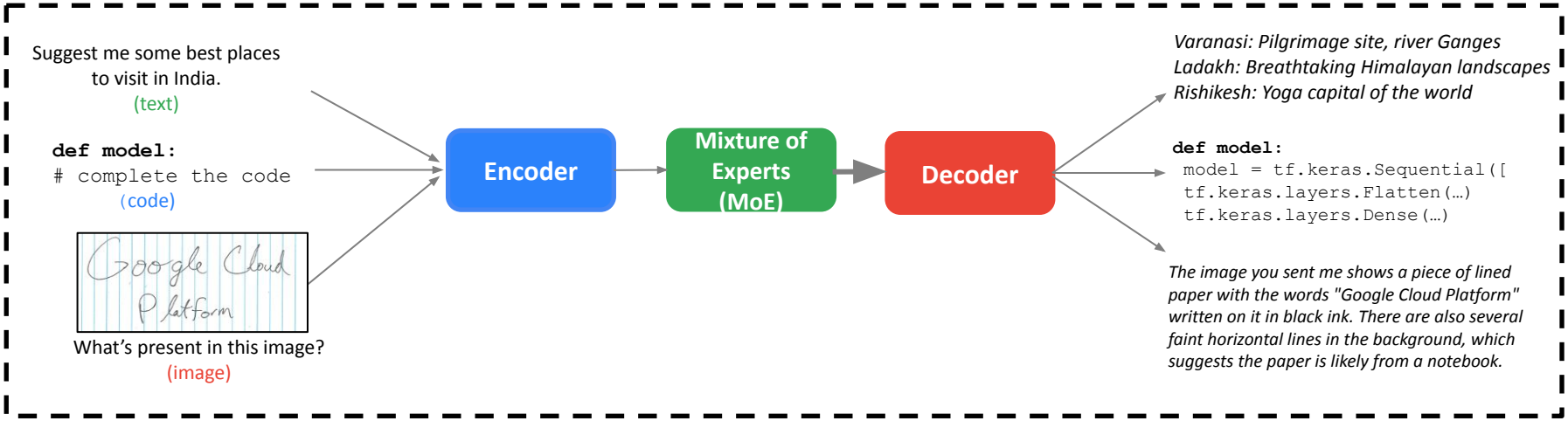


Gemini

I see the time signature is 6/8. This means there are 6 eighth notes in each measure.

The dynamic marking is piano, which means to play softly. Andante grazioso means to play at a graceful walking pace.

Architecture of Gemini



- **Encoder:** Takes input formats like text, code, images, or audio and converts them into a common internal representation that the decoder can understand.
- **Decoder:** Based on the encoded information and the specific task at hand, the decoder generates outputs in different modalities, such as text, code, or translated languages.
- Mixture of Experts: This component employs a network of smaller, specialized models ("experts") instead of a single, monolithic model. The input is routed to the most relevant expert based on its characteristics for improved efficiency and performance.

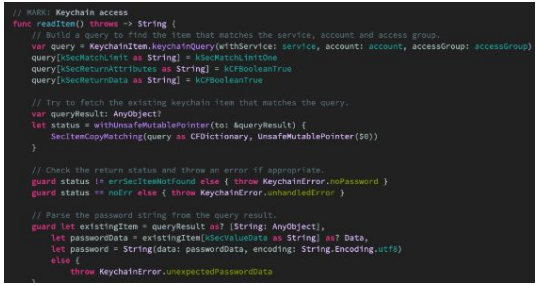
Benchmarks of Gemini 1.5

- Enhancement to Gemini 1.0, with a larger token context window (1,28,000) and can go up to 1 million tokens.

How much data is 1 million tokens?

Videos up to 1 hour

Audio up to 11 hours



~ 30,000 lines of code

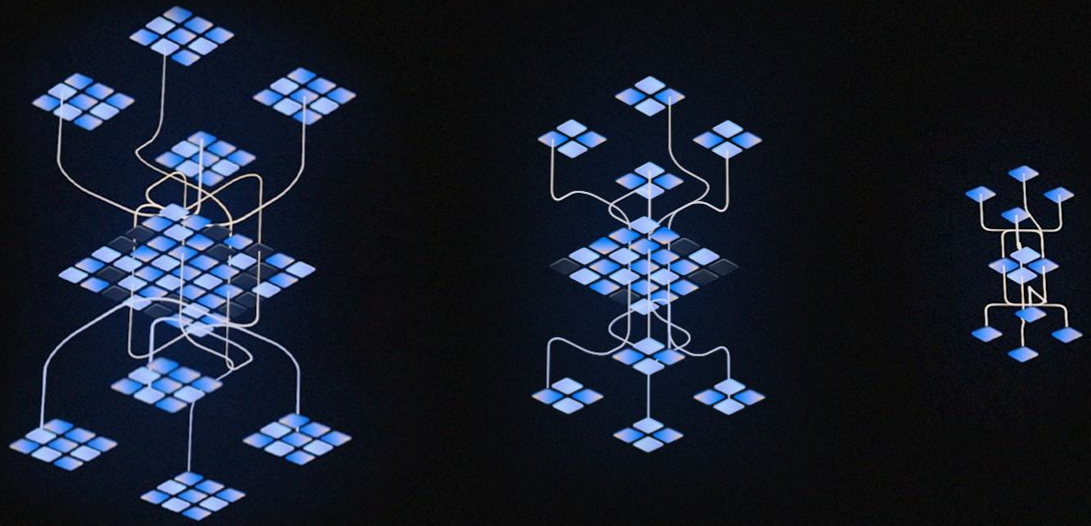
- Needle in a Haystack evaluation: 99% times Gemini Pro 1.5 found a small piece of text containing a particular fact or statement purposely placed within a long block of text of about 1 million tokens.

Learn more:
https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

Experts

Media Source: Google

Gemini is also the most flexible model Google ever built



Ultra 1.0 Pro 1.5 Nano 1.0

Each Gemini model is built for its own set of use cases, making a versatile model family that runs efficiently on everything from data centers to on-device.

Our most capable and largest model for highly-complex tasks.

Our best model for scaling across a wide range of tasks.

Our most efficient model for on-device tasks.

Ultra capable

State-of-the-art performance across a range of industry benchmarks.

Multimodal reasoning

Natively understands and reasons across sequences of audio, images, and text.

Complex coding

Excels at coding and achieves state-of-the-art performance when integrated into AlphaCode 2.

Mathematical reasoning

Advanced analytical capabilities and strong performance on competition-grade problem sets.

Gemini Pro

Longer context

1.5 Pro introduces a breakthrough context window of up to two million tokens — the longest context window of any large scale foundation model yet. It achieves near-perfect recall on long-context retrieval tasks across modalities, unlocking the ability to accurately process large-scale documents, thousands of lines of code, hours of audio, video, and more.

Nano size, mega utility

We're focused on making Nano the most powerful on-device model available. Later this year, Pixel and Android users will be the first to experience its new multimodal capabilities.

Image understanding

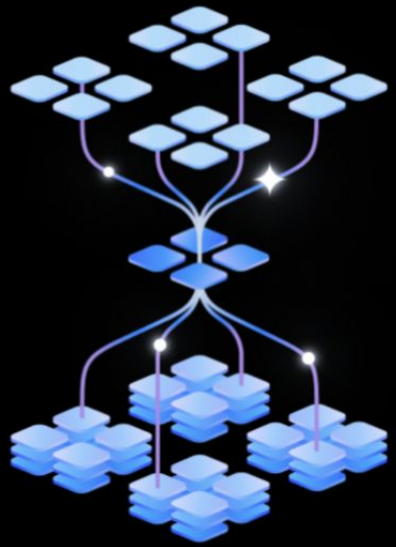
Richer and clearer descriptions of images and what's in them.

Speech transcription

Understands what you're saying so it's possible to talk rather than type.

Text summarization

Distil messages, emails, and documents into concise, readable summaries.



Flash 1.5

Light weight model optimized for speed and efficiency

Performance in a flash

Designed to be fast and efficient to serve at scale.

Built for speed

Sub-second average first-token latency for the vast majority of developer and enterprise use cases.

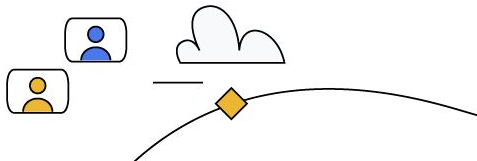
Quality at lower cost

On most common tasks, 1.5 Flash achieves comparable quality to larger models, at a fraction of the cost.

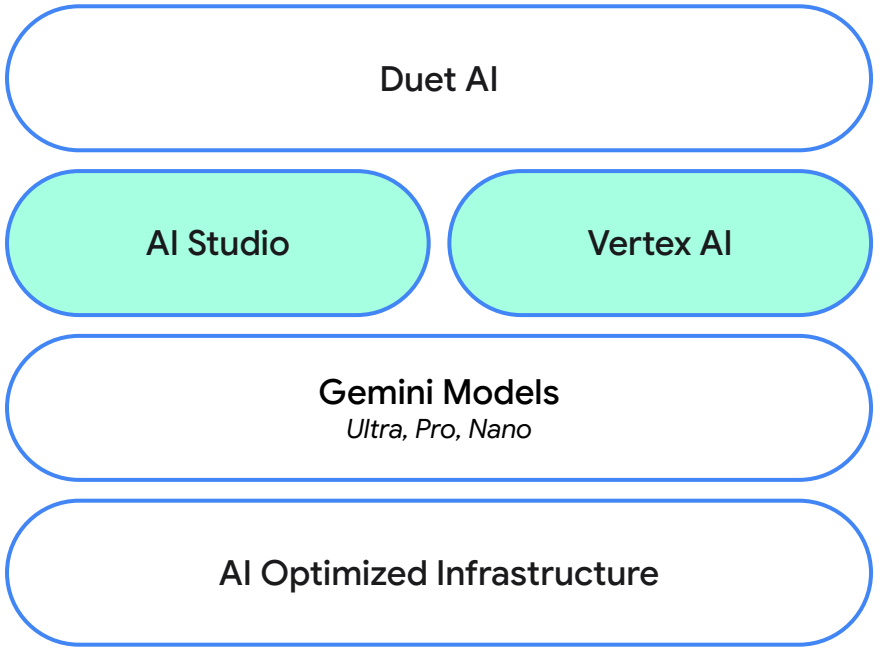
Long-context understanding

Process hours of video and audio, and hundreds of thousands of words or lines of code.

Build Gemini-powered applications



Available for use via AI Studio and Vertex AI



Gemini API

Gemini Pro: Optimized for text prompts.

Gemini Pro Vision: Optimized for text and image prompts.

```
# Install SDK
pip install -q -U google-generativeai
import google.generativeai as genai

# Configure API Key
genai.configure(api_key=<your-API-key>)

# Configure the model
model = genai.GenerativeModel('gemini-pro-vision')

# Give the input image and prompt, get response.
img = PIL.Image.open('image.jpg')
response = model.generate_content(["Describe the image", img],
stream=True)
```

Hands-on with Gemini 1.0 Pro: Colab

- Step 1: Go to Google AI Studio and generate an API key.
- Step 2: Clone the repository: https://github.com/AashiDutt/Gemini-Demo
- Step 3: Upload the notebook file on Google Colab.
- Step 4: Execute the cells in the notebook by following the instructions.

Hands-on with Gemini 1.5 Pro: Google AI Studio

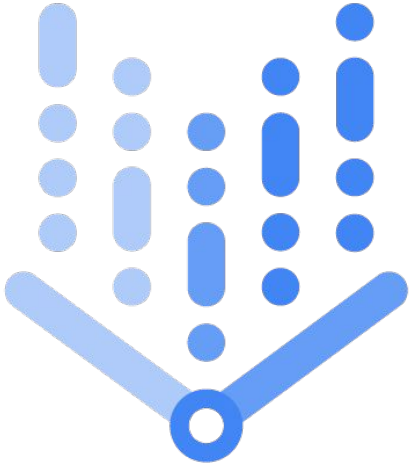
- Step 1: Go to Google AI Studio and generate an API key.
- Step 2: Choose the Gemini 1.5 Pro model.
- Step 3: Try prompts with multi-modal data such as images, videos, texts, PDFs, etc.

Demo Time- Scan and load

Experts

Vertex AI

Vertex AI is a machine learning (ML) platform that lets you train and deploy ML models and AI applications, and customize large language models (LLMs) for use in your AI-powered applications. Vertex AI combines data engineering, data science, and ML engineering workflows, enabling your teams to collaborate using a common toolset and scale your applications using the benefits of Google Cloud.

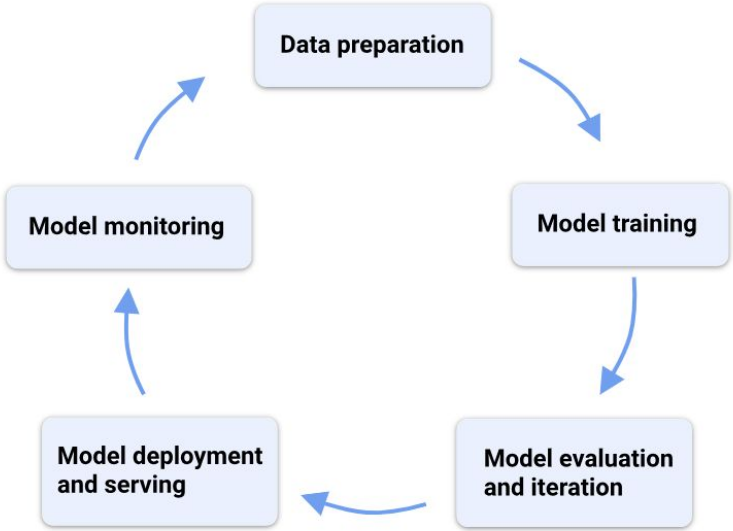


Vertex AI provides several options for model training and deployment:

- AutoML lets you train tabular, image, text, or video data without writing code or preparing data splits.
- Custom training gives you complete control over the training process, including using your preferred ML framework, writing your own training code, and choosing hyperparameter tuning options.
- Model Garden lets you discover, test, customize, and deploy Vertex AI and select open-source (OSS) models and assets.
- Generative AI gives you access to Google's large generative AI models for multiple modalities (text, code, images, speech). You can tune Google's LLMs to meet your needs, and then deploy them for use in your AI-powered applications.

After you deploy your models, use Vertex AI's end-to-end MLOps tools to automate and scale projects throughout the ML lifecycle. These MLOps tools are run on fully-managed infrastructure that you can customize based on your performance and budget needs.

Machine learning workflow



https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform

Vertex AI

Dashboard

LEARN

TOOLS

- Dashboard
- Model Garden
- Pipelines

NOTEBOOKS

- Colab Enterprise
- Workbench

VERTEX AI STUDIO

- Overview
- Multimodal NEW
- Language
- Vision
- Speech

DATA

MODEL DEVELOPMENT

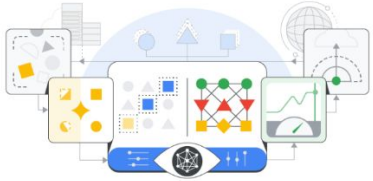
DEPLOY AND USE

Marketplace

Get started with Vertex AI

Vertex AI empowers machine learning developers, data scientists, and data engineers to take their projects from ideation to deployment, quickly and cost-effectively. Learn more about Vertex AI

ENABLE ALL RECOMMENDED APIS

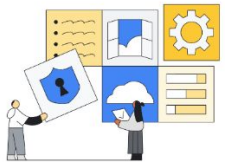


Tutorials

Try an interactive tutorial to learn how to train, evaluate, and deploy a Vertex AI AutoML or custom-trained model.

VIEW TUTORIALS

SHOW API LIST



Colab Enterprise

A new notebook experience with enterprise-grade privacy and security. Start coding in a couple clicks.

Go to Colab Enterprise

Model Garden

Browse, customize, and deploy machine learning models. Choose from Google or popular open-source models.

Try now

Vertex AI Studio

Test and customize large language and generative image models.

Try now

Try our most advanced model, Gemini 1.5 Pro Experimental, which supports up to a 1 million token context window TRY GEMINI

Vertex AI Studio

Vertex AI Studio lets you quickly test and customize generative AI models so you can leverage their capabilities in your applications. Learn more

DOCUMENTATION API REFERENCE

Multimodal
Powered by Gemini NEW

Try Gemini, the latest Google model family capable of processing text, image, video, audio, and documents with a 1M token context window. Learn more about Gemini

TRY GEMINI MULTIMODAL HOME VIEW CODE

Language
Powered by Gemini NEW

Write natural language and code prompts for tasks like classification, summarization, code generation, chatbots and more, with PaLM 2 or Gemini.

OPEN VIEW CODE

Vision
Powered by Imagen NEW

Write text prompts to generate new images or new areas of an existing image.

OPEN VIEW CODE

Speech

Convert speech into text or synthesize speech from text using Google's Universal Speech Model (USM).

OPEN VIEW CODE

Model Garden

EXPLORE GENERATIVE AI | VIEW MY MODELS | DEPLOY FROM HUGGING FACE

Search models

Browse, customize, and deploy machine learning models with **Model Garden**. Choose from models created by Google and other providers.

Sort by: **Trending** Newest Last Update Alphabetical

Foundation models

→ SHOW ALL (70)

Pre-trained multi-task models that can be further tuned or customized for specific tasks. Models marked with ✦ are available in Vertex AI Studio.

Gemini 1.0 Pro

The best performing Gemini model with features for a wide range of 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 1.0 Pro Vision

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

Gemma

Open checkpoint variants of Google Deepmind's Gemini model suited for a variety of text generation tasks

CodeGemma

Open code m… riants Gemma model suited for t… text and text-to-code tasks

Fine-tunable models

→ SHOW ALL (44)

Modalities

Language	49
Vision	85
Tabular	6
Document	5
Speech	1
Video	4
Multimodal	6

Tasks

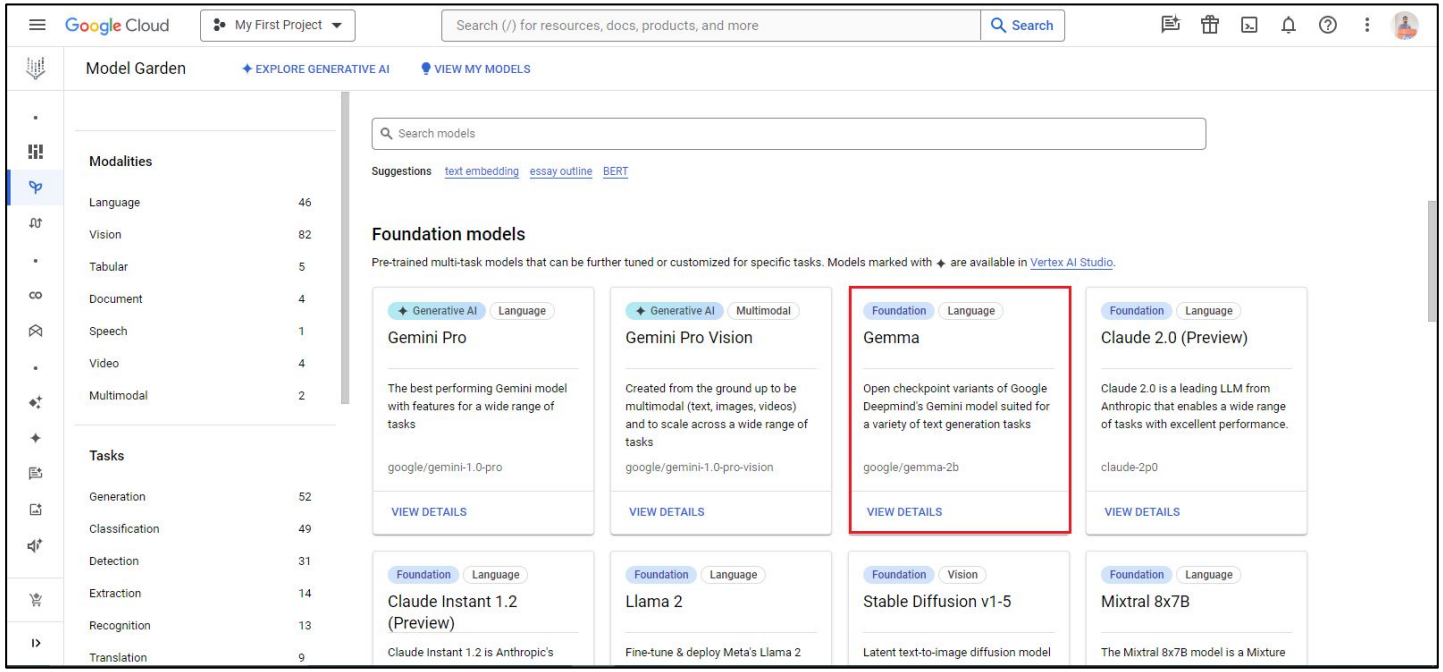
Generation	55
Classification	49
Detection	31
Extraction	15
Recognition	15
Translation	10
Embedding	4
Segmentation	7
Retrieval	1

Bonus: Trying the new Gemma models on Vertex AI

- Gemma is a family of lightweight, state-of-the-art open models built from research and technology used to create Google Gemini models.
- They are text-to-text, decoder-only large language models, available in English, with open weights, pre-trained variants, and instruction-tuned variants.
- Gemma is available in two variants – Gemma 2B and Gemma 7B on Vertex AI and Hugging Face.

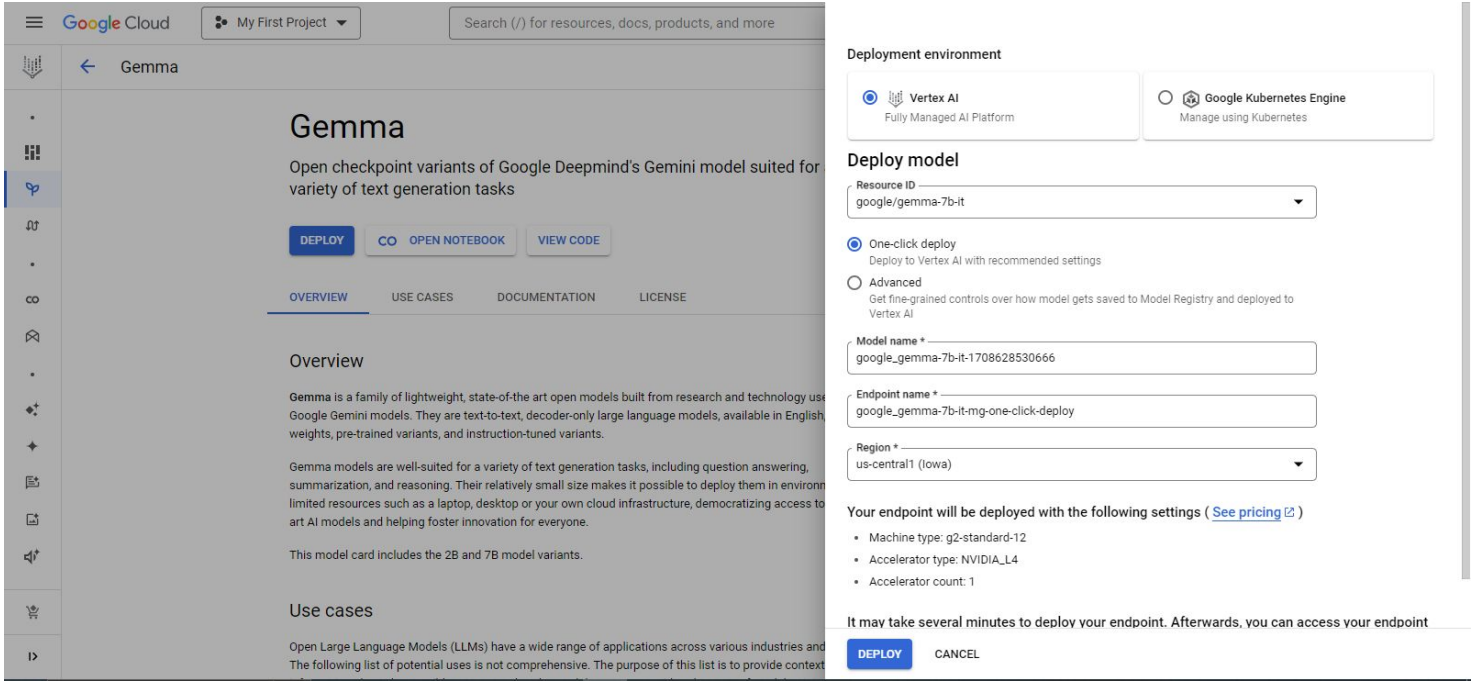
Bonus: Trying the new Gemma models on Vertex AI

- Step 1: On GCP, open the Vertex AI Model Garden and choose the Gemma model.



Bonus: Trying the new Gemma models on Vertex AI

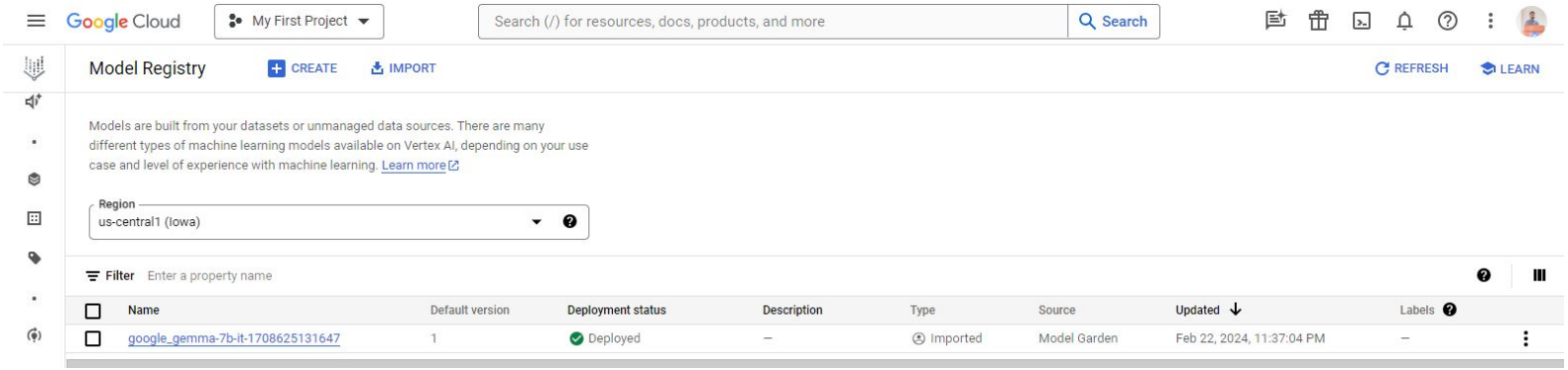
- Step 2: Choose any model variant to deploy it to a Vertex AI endpoint. It takes approximately 20-30 minutes for the deployment.



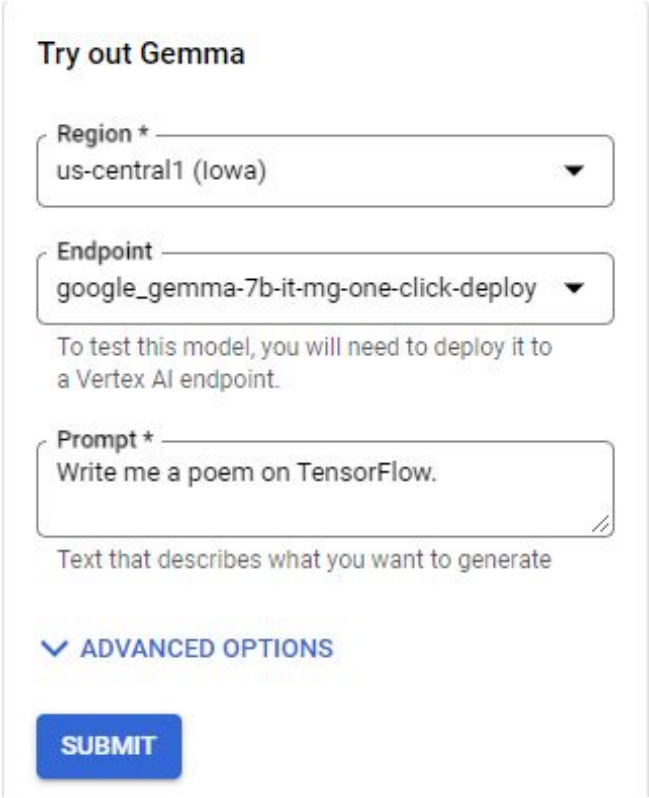
Experts

Bonus: Trying the new Gemma models on Vertex AI

- Step 3: Once the model is deployed, you should be able to see it in your Model Registry.



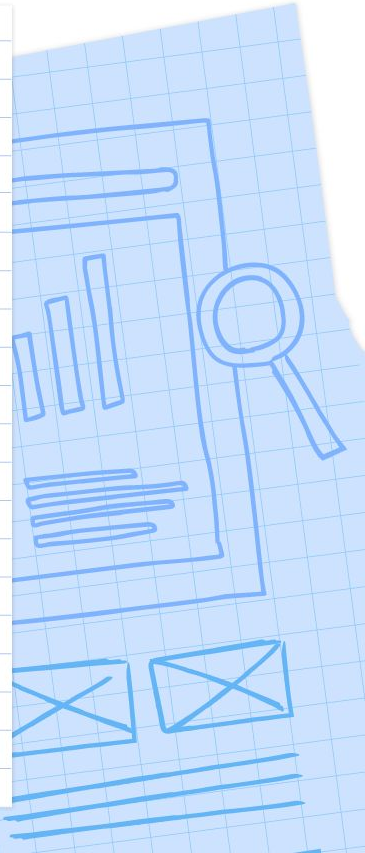
Bonus: Trying the new Gemma models on Vertex AI



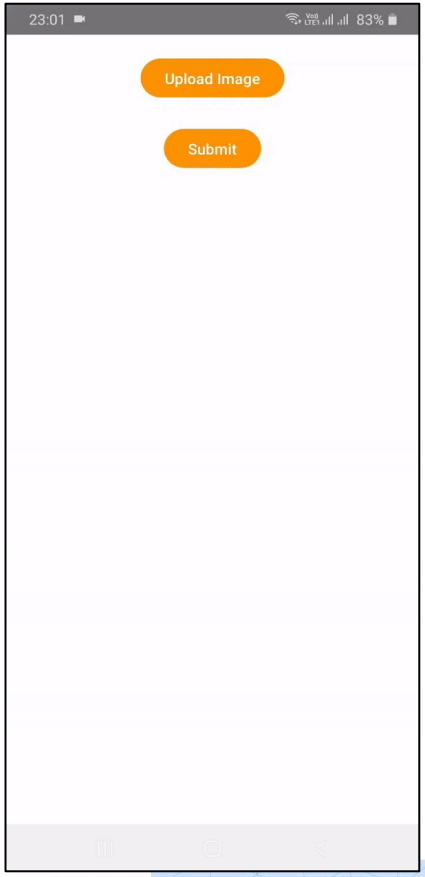

```
Prompt:
Write me a poem on TensorFlow.
Output:
A framework born for neural nets,
Where data flows and models meet.
TensorFlow, the name so proud,
A power tool for learning, out of
the crowd.

With Python's grace, it takes its
flight,
Data transformation, a swift
delight.
Layers stack up, an intricate art,
The model learns, a hidden start.

From image recognition to
language,
It finds solutions, beyond our
range.
Medical strides, a helping hand,
Unveiling secrets, across the
land.

So hail TensorFlow, the machine
learning king,
A force for good, with endless
sting.
May it empower, bring new light,
And unlock solutions, day and
night.
```

Experts

Examples built with Gemini



💊 PharmaScan : Extract medicine details using Gemini



PharmaScan is an Android application that leverages the Gemini Pro Vision model to identify medicines and provide their details such as usage, dosage, diagnosis, etc. on the go.

Check out on GitHub:
https://github.com/NSTiwari/Medicine-Scan-with-Gemini

Try it on Hugging Face 🤗:
https://huggingface.co/spaces/Aashi/Medicine-Prescription-with-Gemini

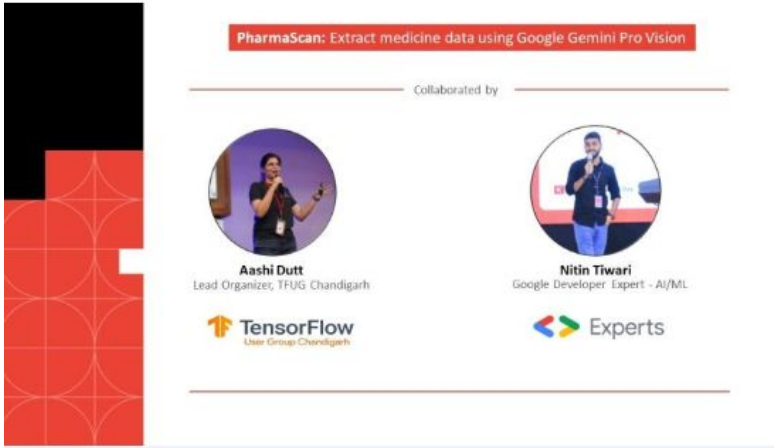
Watch the complete demo on YouTube:
https://www.youtube.com/watch?v=Q06ABLwFGTQ

The project was featured on the official handle of Google Developer Expert

Google Developer Experts
399,623 followers
1w •

Picture it: you pull out some meds and realize you don't have the packaging. Now, you're missing the uses, dosage info, and directions.

ML GDE Nitin Tiwari and Aashi Dutt have created a solution with PharmaScan. Using Gemini Vision Pro, the app provides instant info about your meds. Learn more → https://goo.gle/3w90RqS

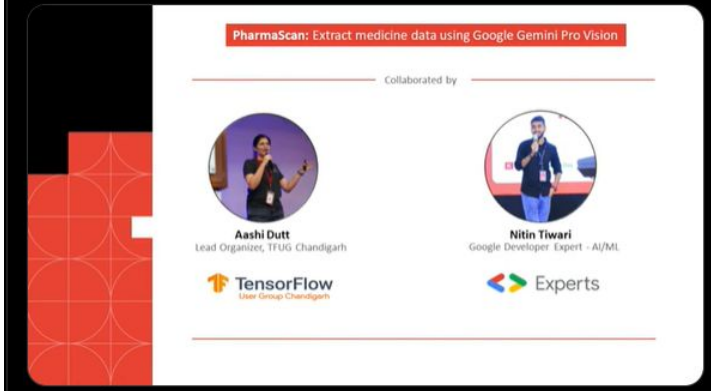


PharmaScan: Extract medicine data using Google Gemini Pro Vision
youtube.com

@GoogleDevExpert
@GoogleDevExpert

Picture it: you pull out meds and realize you're missing the packaging, with its uses and directions.

ML GDE Nitin Tiwari and Aashi Dutt's app PharmaScan solves the issue with Gemini Vision Pro, providing instant info about your meds. Learn more ↓



From youtube.com

12:30 AM · Feb 16, 2024 · **2,310** Views

6 16 1

Experts

👩‍🌾 FarmScan : Farmer's Digital Assistance built with Gemini

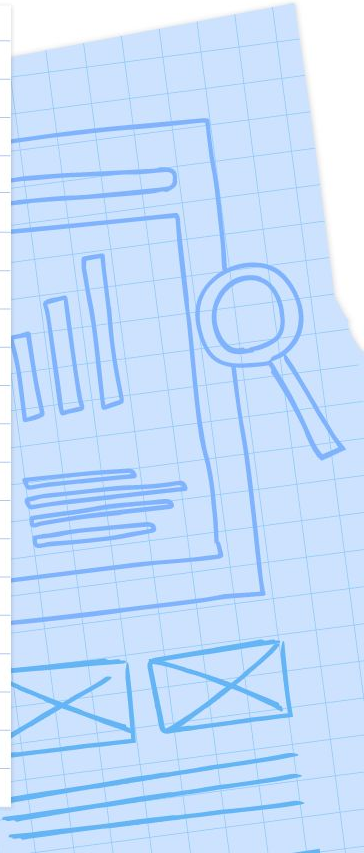
FarmScan is an implementation of the Google Pro Vision model API on Android to recognize the freshness of fruits/vegetables, their approximate market value, shelf life, and a lot more insights to help farms plan the cultivation/selling of crops better.

Check out on GitHub:
https://github.com/NSTiwari/FarmScan-using-Gemini

Try it on Hugging Face 🤗:
https://huggingface.co/spaces/Aashi/FarmScan

Experts

Resources



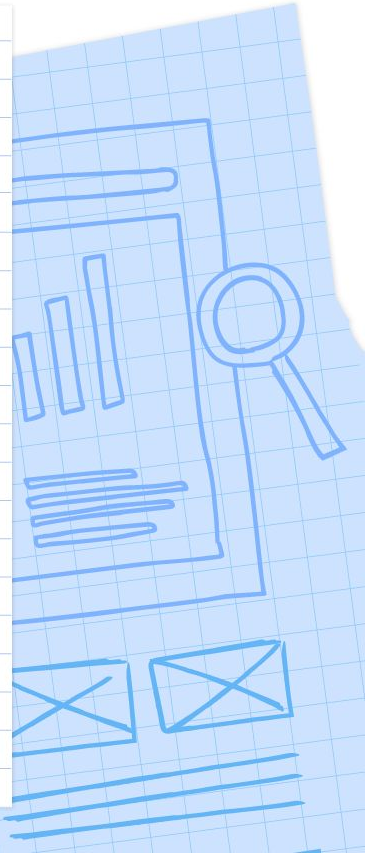
Resources

- Google AI Studio: https://aistudio.google.com
- Get started with examples on https://ai.google.dev/examples
- Google AI SDK for JavaScript: https://github.com/google/generative-ai-js
- Getting started with Gemini Nano (on-device): https://ai.google.dev/tutorials/android_aicore?hl=en
- Gemini API: Quickstart with Python: https://ai.google.dev/tutorials/python_quickstart?hl=en
- Android SDK for Gemini API: https://ai.google.dev/tutorials/android_quickstart?hl=en
- Gemini on Google Cloud Platform: https://github.com/GoogleCloudPlatform/generative-ai/

Experts

We’re excited to see what you build with Gemini.

Share on Twitter (X) @AashiDutt



Experts

Thank You.

medium.com/@aashi-dutt3

twitter.com/@AashiDutt

github.com/AashiDutt