

Google Developer Groups
Chennai

Build Your Social Media Brain: Content Creation with Google ADK Agents

cloud
community
days ✨ 2025 *



Google Developer Groups

About Me – Geeta Kakrani

Startup Mentor | AI Consultant | GDE in AI/ML | Founder @ kanishkaIT

- Helping early-stage startups build AI-first products
- Expertise in LLMs, Multi-Agent Systems & Generative AI
- Speaker at IITs, IIMs & global tech forums
- Focused on turning AI ideas into scalable, real-world solutions



cloud
community
days ✨ 2025 *



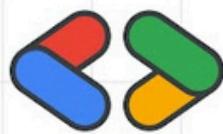
Google Developer Groups

Low Internate

🚦 "Waiting for LLM to respond like it's a red light in Mumbai traffic."

cloud community
days ✨ 2025 *





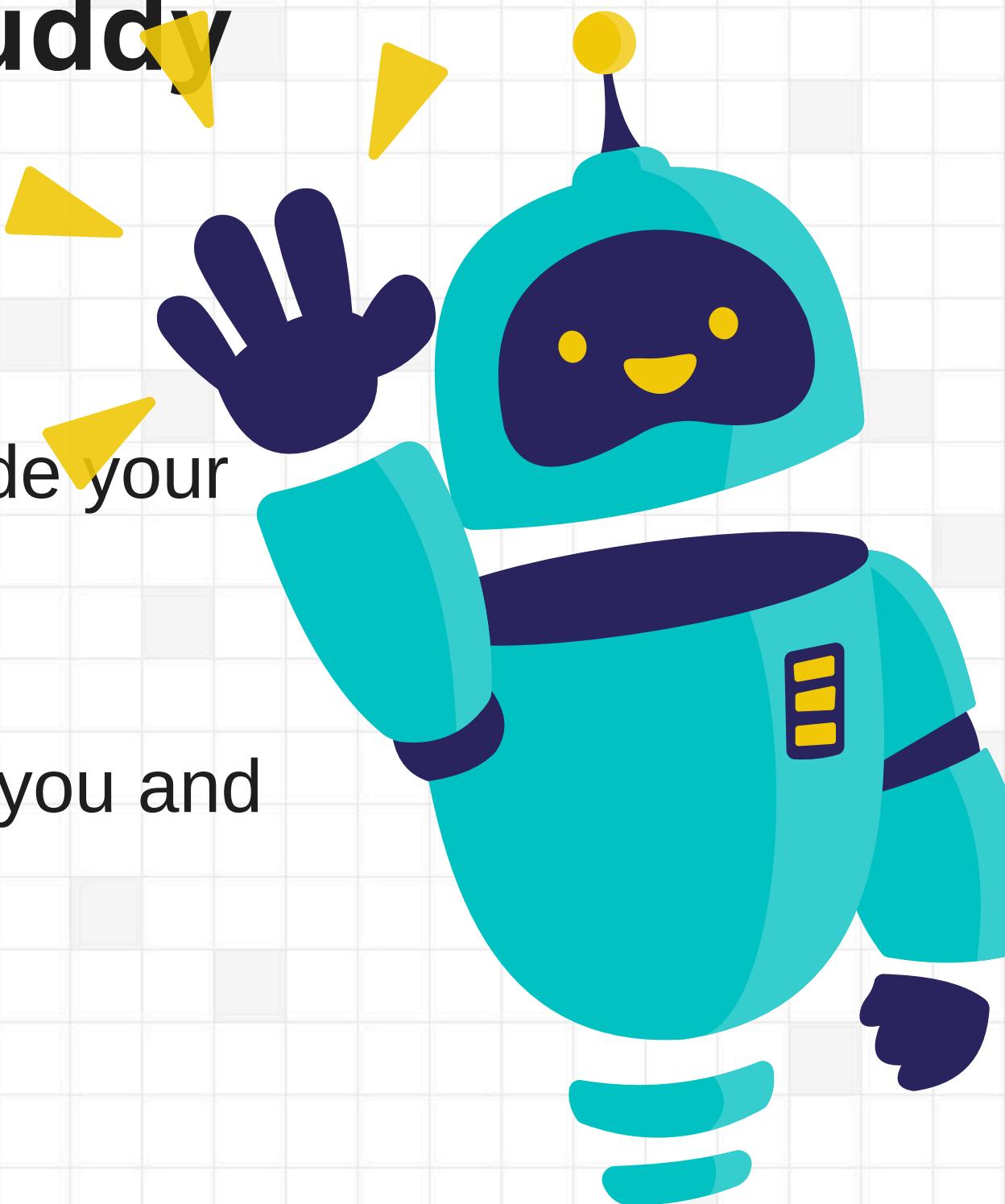
Google Developer Groups

Meet Gemma – Your On-Device LLM Buddy

Gemma says:

- I'm Gemma – an open-source LLM from Google.
- I don't need the internet to think... because I live right inside **your** device!"
- Fast responses. No network lags. No waiting."
- And guess what? I can be updated and fine-tuned just for you and your startup's needs.

cloud community
days ☀ 2025 *





Gemma Model Sizes – Choose What Fits Your Need

Gemma is available in multiple sizes to balance speed, accuracy, and device compatibility:

Model Size Like a... Best For

1B 🚲 Scooter Super fast, low memory apps (IoT, mobile)

4B 🚗 Sedan Balanced performance (Laptops, Edge devices)

12B 🚙 SUV Smarter AI, deeper tasks (Chatbots, education)

27B 🚛 Heavy Truck Most powerful (Cloud, R&D, multi-agent systems)

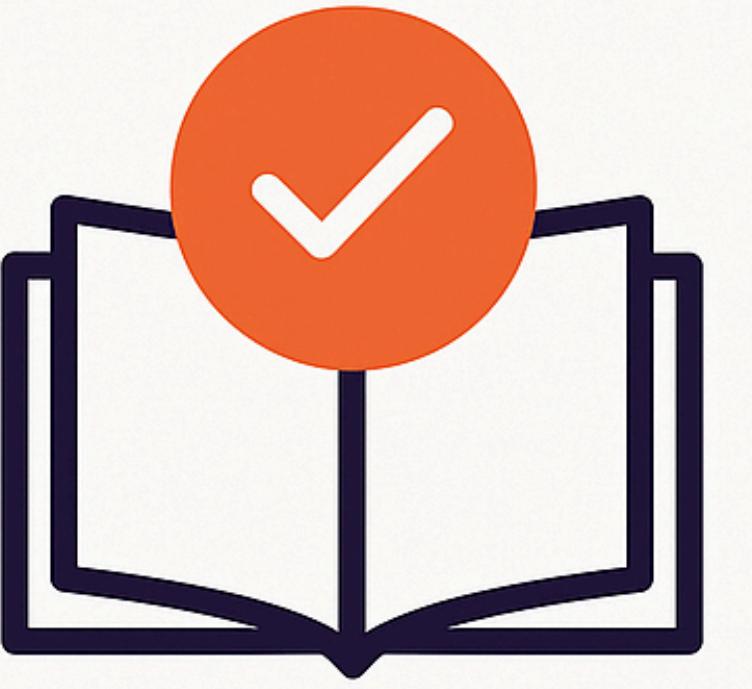


Tip:

Smaller models = faster, run locally

Larger models = more intelligent, require better hardware.

👉 Start small, scale up as needed!



Fine-Tune Gemma





Fine-Tuning vs. RAG (Retrieval-Augmented Generation)

Fine-Tuning

- Teaches the model new knowledge by updating its internal weights
- Requires training on custom data (e.g., legal, medical, domain-specific)
- Knowledge is permanent
- Slower to update, needs compute resources
- Ideal when you want the model to write or behave in a specific way



RAG (Retrieval-Augmented Generation)

- Keeps the model frozen and fetches external info at runtime
 - Uses tools like vector DBs (FAISS, Chroma, etc.)
 - Knowledge is temporary and easily updatable
 - No retraining needed – just update your documents
 - Ideal for real-time Q&A, chatbot over PDFs, large knowledge base



Summary

- **Fine-Tuning = long-term memory**
- **RAG = dynamic, real-time lookup brain**

LoRA Fine-Tuning – Lightweight Model Customization

What is LoRA?

LoRA (Low-Rank Adaptation) is a smart, efficient way to fine-tune large models by adding small trainable adapter layers – keeping the main model frozen.

Key Benefits:

-  Lightweight – Only a few parameters are trained
-  Fast & Cost-Efficient – No need to retrain the full LLM
-  Modular – Easily switch or remove fine-tuned layers
-  Great for personalization – Teach the model your domain-specific style or data

Example:

Want to make Gemma understand legal or medical content?

→ Use LoRA to fine-tune it quickly and affordably, without touching the base model.

Gemma (Base & Instruct)

General-purpose text model

→ For chat, Q&A, summarization



CodeGemma

Trained for coding tasks

→ Code generation, completion, debugging



PaliGemma

Multimodal model (text + vision)

→ Image captioning, visual Q&A

Real-World Use Cases of Gemma (and Variants)

1. Farmer Assistant (Local Language Chatbot)

Gemma-Instruct can power a voice/text bot that:

Explains crop care, weather alerts, and fertilizer usage

Speaks in Hindi, Marathi, Tamil, etc.

Works offline on edge devices using 2B Gemma

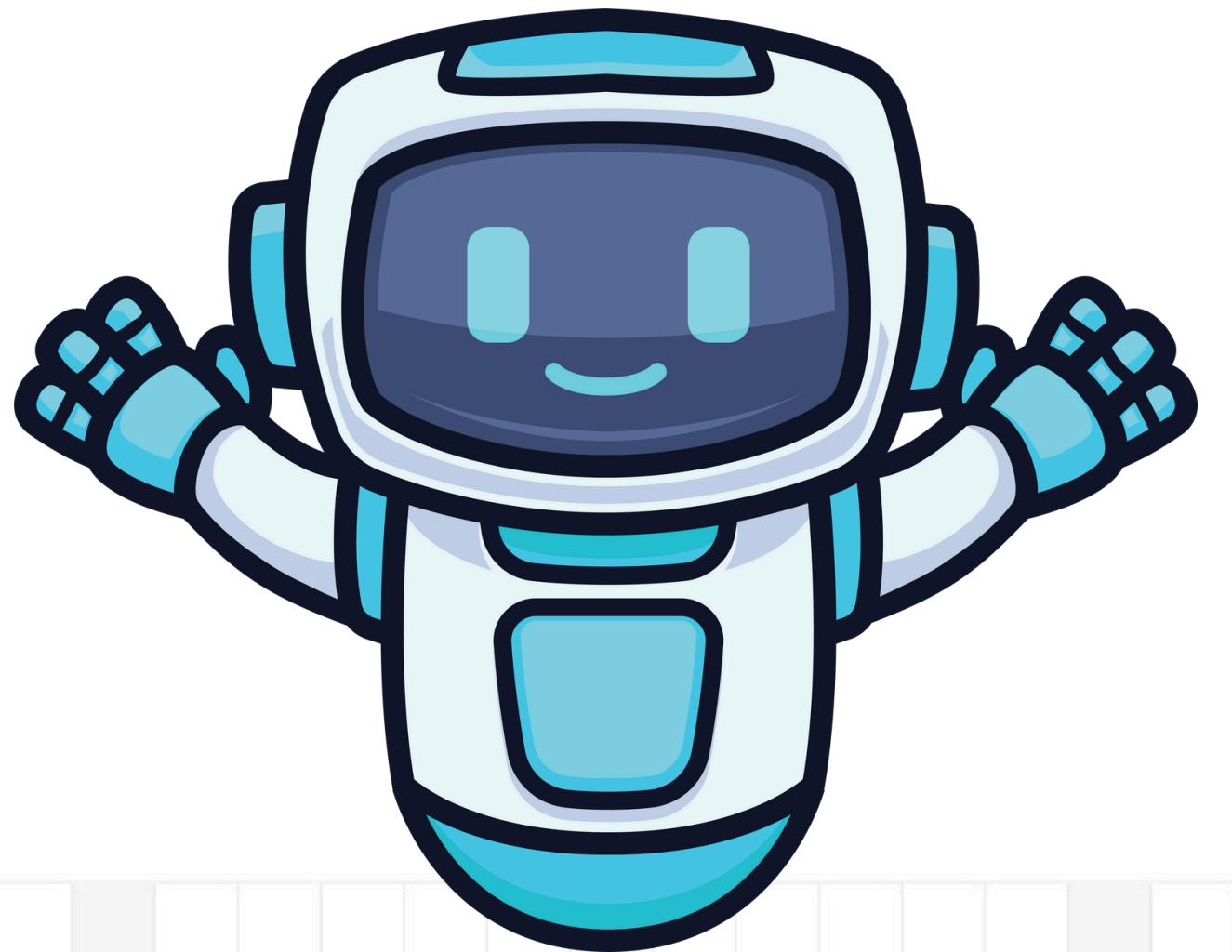
2. Rural Healthcare Q&A Bot (with MediGemma or fine-tuned Gemma)

- Answers health-related queries from rural users
- Uses LoRA to fine-tune on Indian medical guidelines
- Can run on low-end devices in local clinics



Google Developer Groups

How Use Me ?



cloud
community
days ✨ 2025 *

Kaggle

kaggle.com/models/google/gemma-3

Search Sign In Register

GOOGLE · CREATED ON 2025.01.17

315 Download Code :

Gemma 3

google/gemma-3



Gemma is a family of lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models.

Model Card Code (53) Discussion (3) Competitions (5)

You've consented to the license for Gemma 3 View License Consent

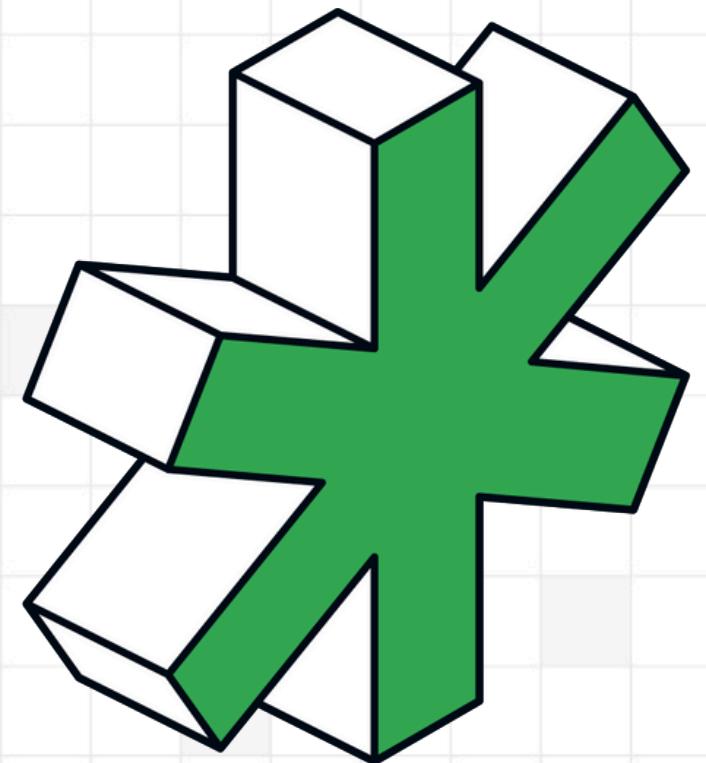
Model Details

Gemma 3 model card

Model Page: Gemma

Downloads
19.8K 
3285 in the last 30 days

Tags



Vertex AI

Google Cloud

Overview Solutions Products Pricing Resources

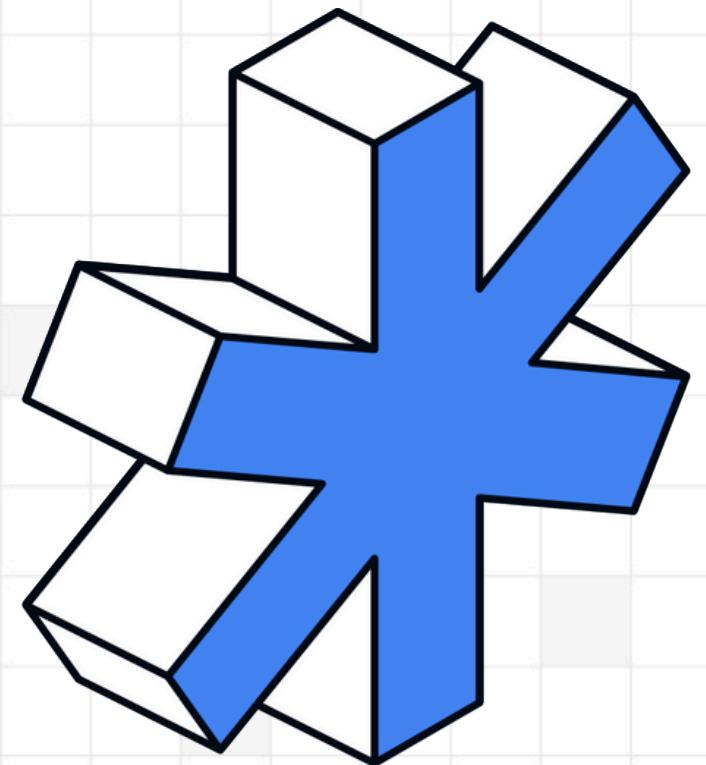


Docs Support Console



Contact Us

[Try Gemini 2.5, our most intelligent model now available in Vertex AI](#)



Model Garden on Vertex AI

Jumpstart your ML project with a single place to discover, customize, and deploy a wide variety of models from Google and Google partners.

[Browse Model Garden](#)

[Contact sales](#)

Morden Garden

Google Cloud GemmaAgent Search (/) for resources, docs, products, and more Search ⚡ Model Garden Explore Generative AI View my endpoints & models Deploy from Hugging Face View release notes

Tasks

Task	Count
Text generation	66
Text classification	44
Entity extraction	31
Translation	31
Image classification	35
Object detection	31
Image segmentation	13
Image generation	33
Image understanding	20
Text embeddings	10
Tabular classification	11
Document processing	12
Image retrieval	1
Video classification	2
Open vocabulary detection	2
Open vocabulary segmentation	2

Foundation models

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

Show all (132)

PaliGemma 1 & 2

Lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models

Gemma

Lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models

CodeGemma

Lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models

Llama

Access a fully

All partners

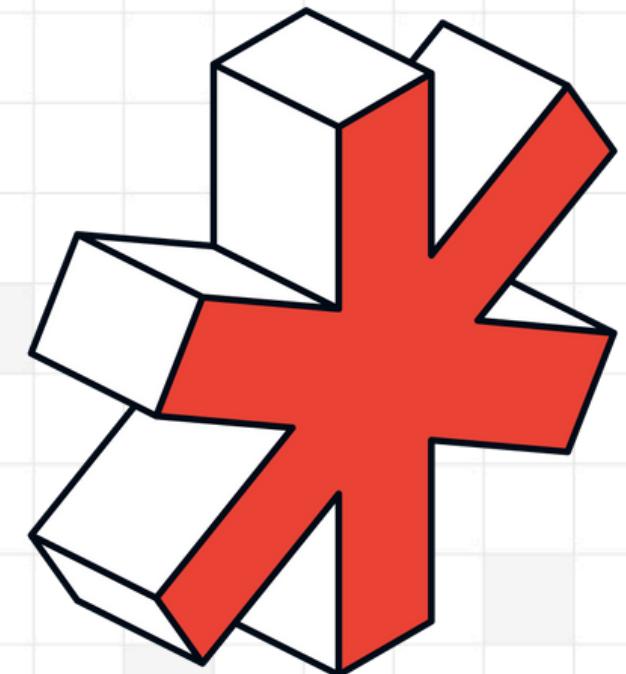
Show all (8)

ANTHROPIC

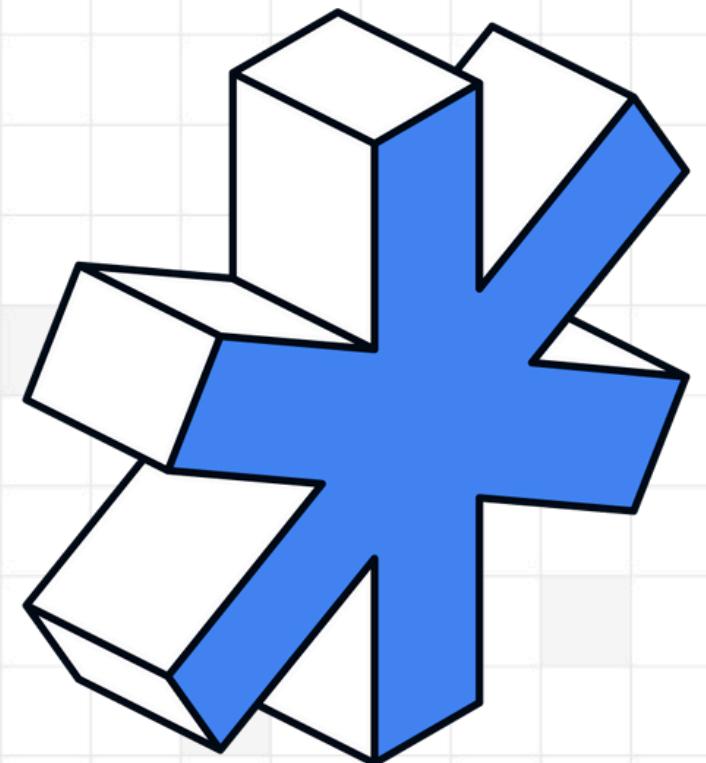
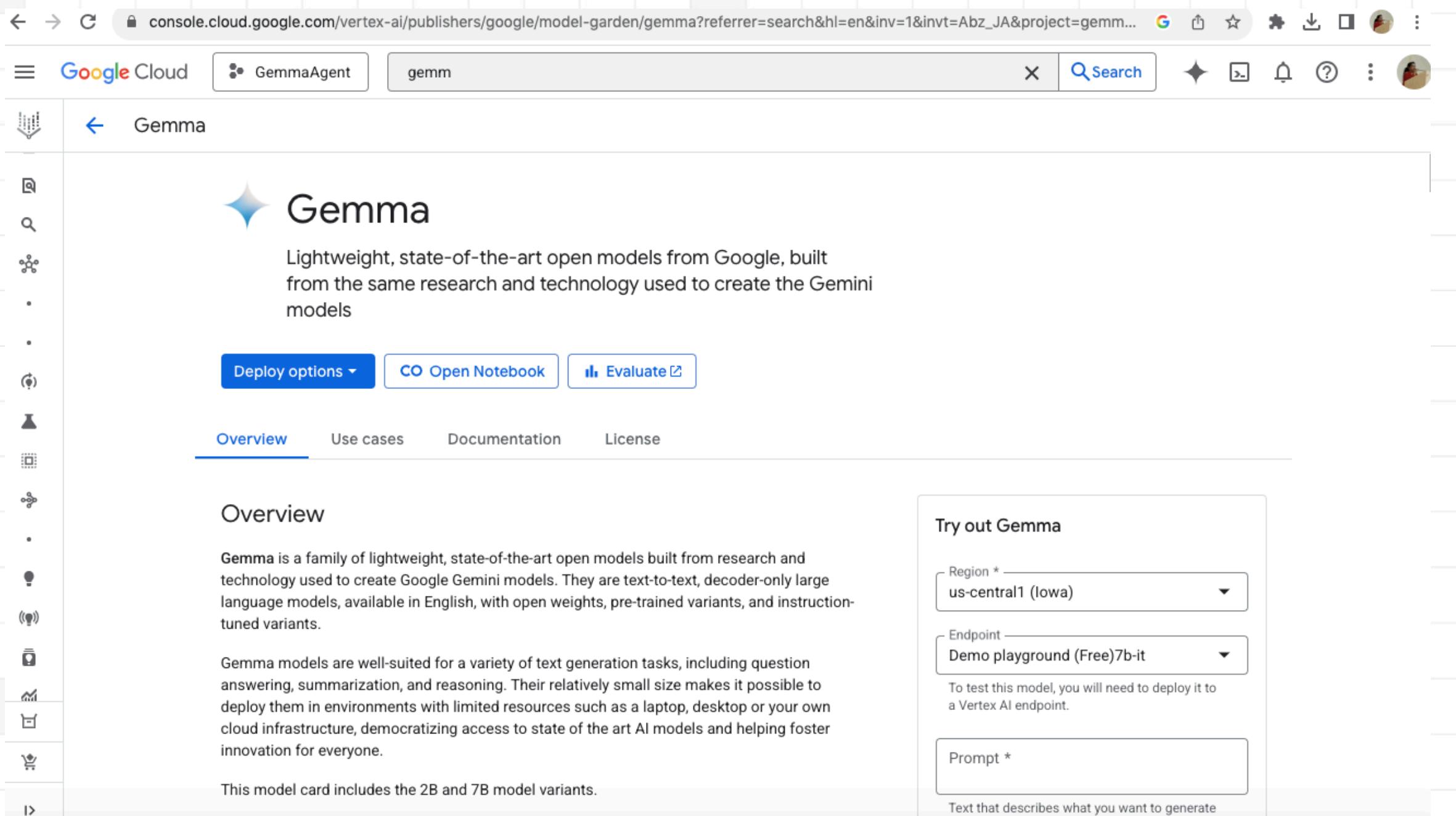
Meta

Hugging Face

is



Deploye Gemma



Create End Point for use in Scalable mode

← → C console.cloud.google.com/vertex-ai/models?hl=en&inv=1&invlt=Abz_JA&project=gemmaagent-462506

Google Cloud GemmaAgent Search (/) for resources, docs, products, and more [Search](#)

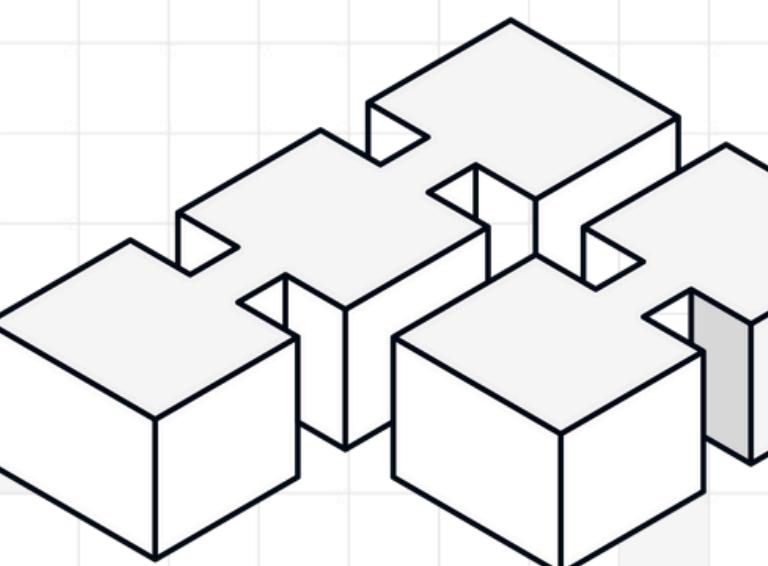
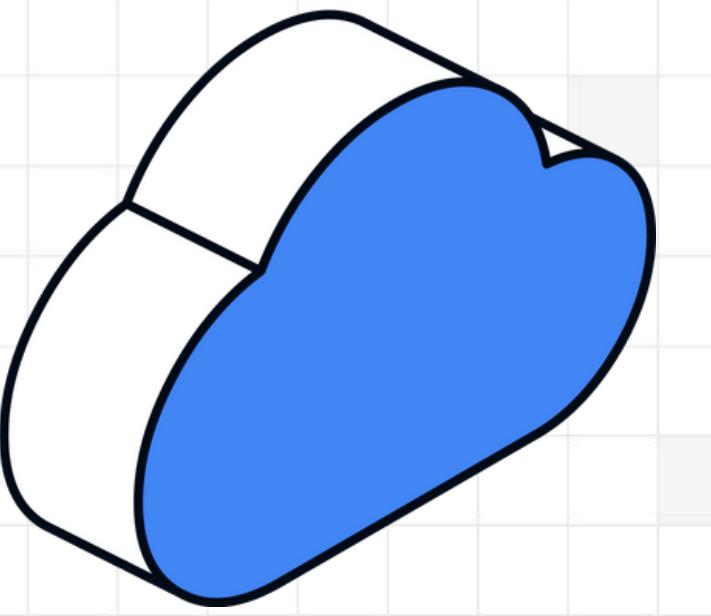
Model Registry [Create](#) [Import](#) Refresh Learn

Models are built from your datasets or unmanaged data sources. There are many different types of machine learning models available on Vertex AI, depending on your use case and level of experience with machine learning. [Learn more](#)

Region: us-central1 (Iowa)

Filter: Enter a property name

Name	Default version	Deployment status	Description	Type	Source	Updated	Labels
gemma-3-1b-it-	1	Deployed	-	Imported	Model Garden	Jun 10, 2025, 12:46:01 PM	-



THANK YOU

You can connect with me on:



<https://www.linkedin.com/in/geetakakrani/>



@GKakrani



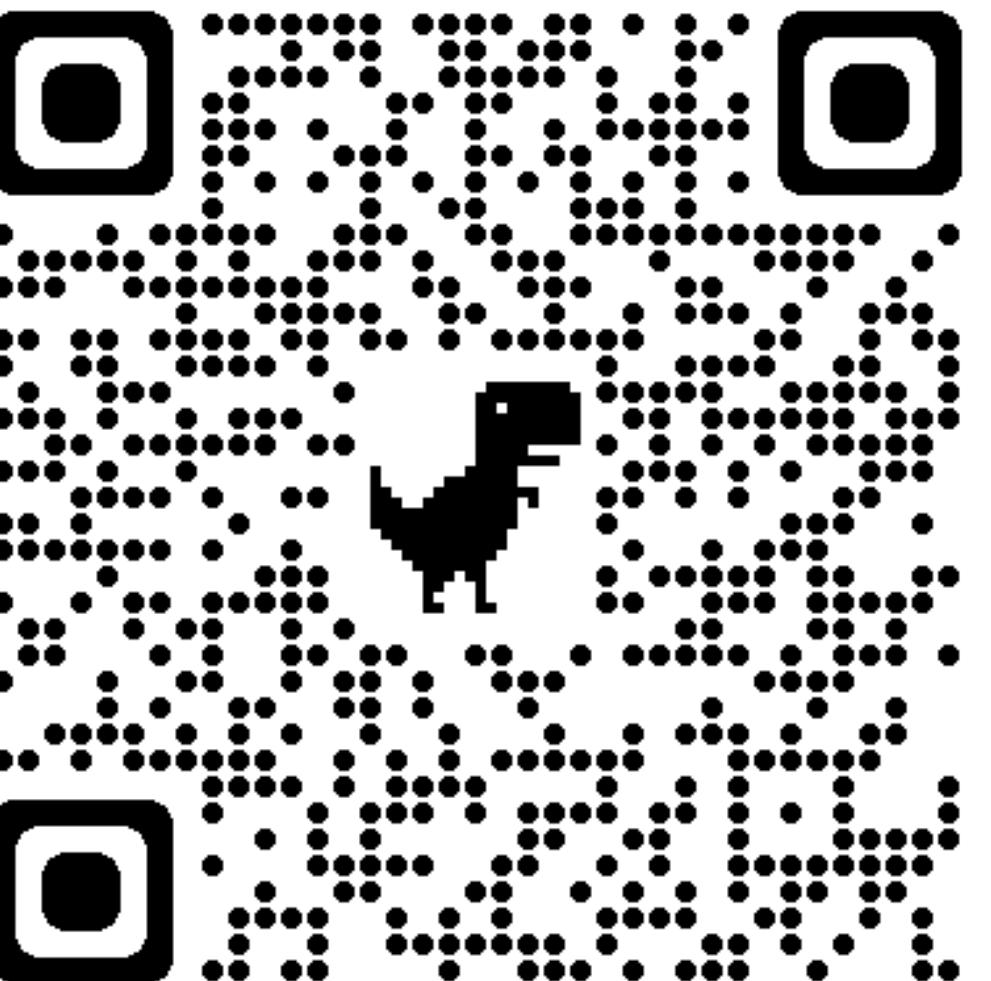
kanishkaitpvtltd@gmail.com



<https://www.youtube.com/@kanishkait>



<https://medium.com/@Geetakakrani>



<https://kanishkait.in>

??



- Q & A -

TIME