

# Build with Gemma





## Manikandan Krishnamurthy

Google Cloud Architect 🏗️ | 5x GCP Customer  
Delivery 🌐 | 2x GCP Certified 🏆 | 6x GCP He...

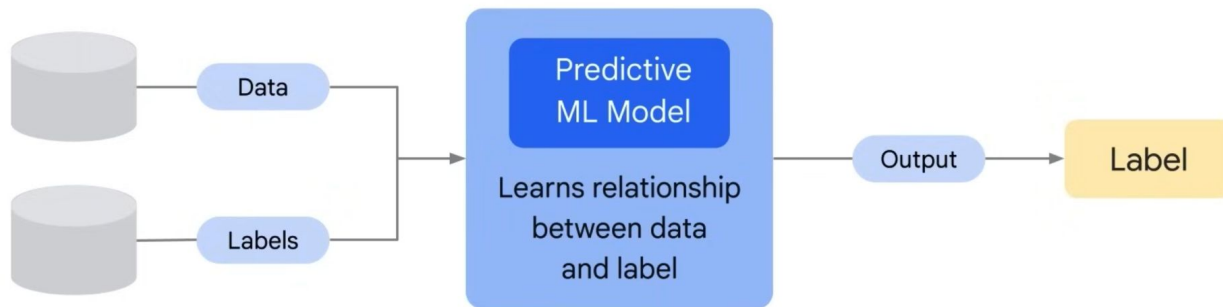




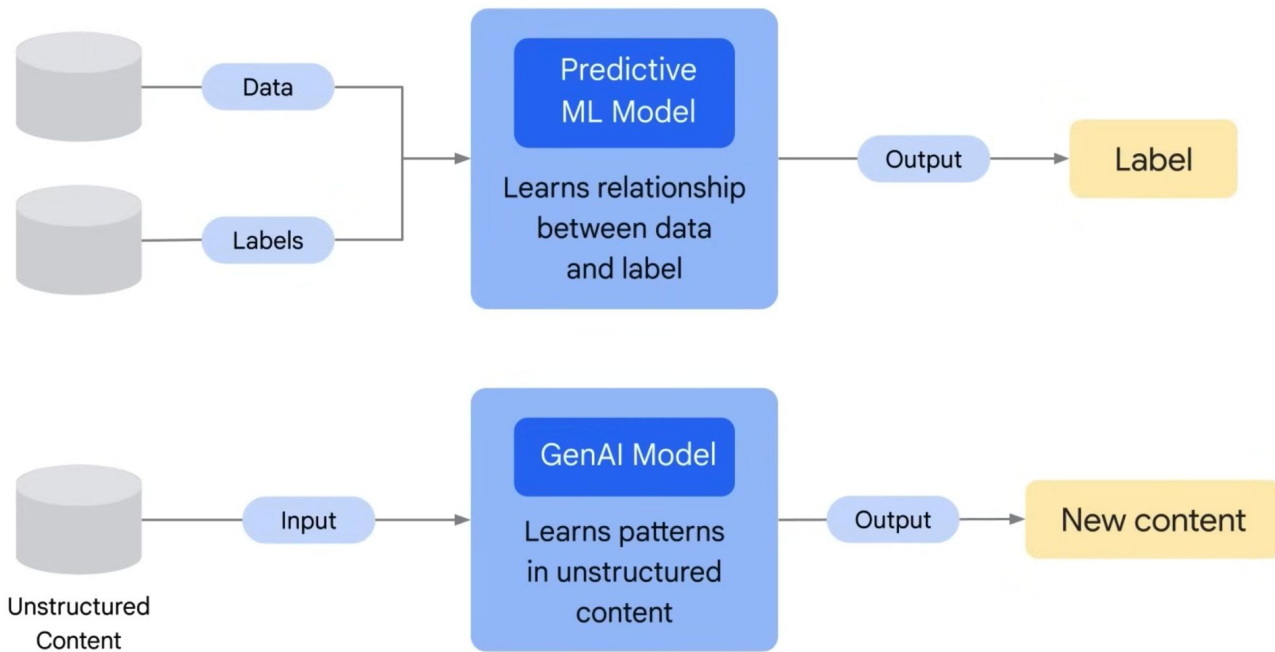
# Overview



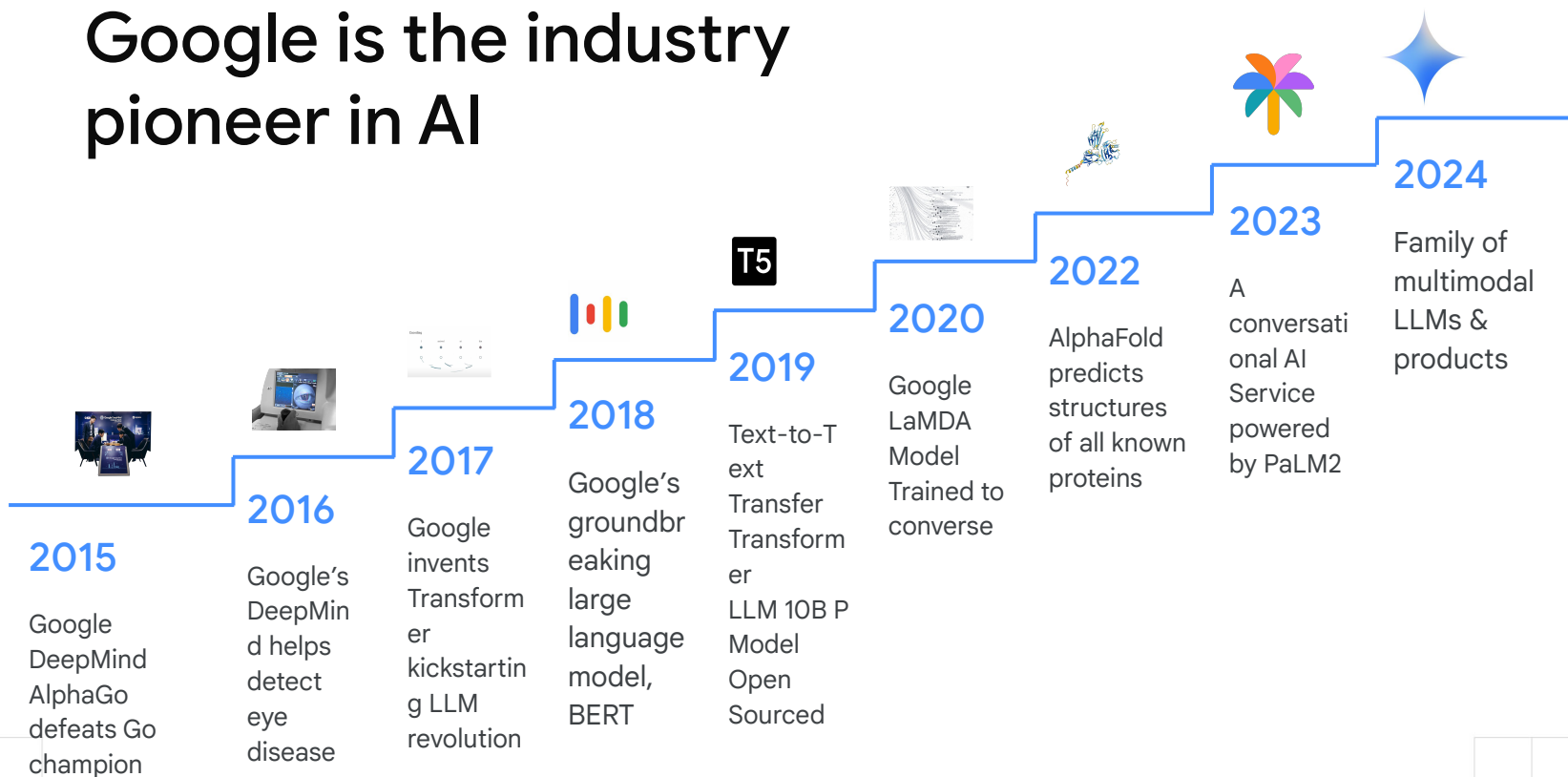
# Generative AI



# Generative AI



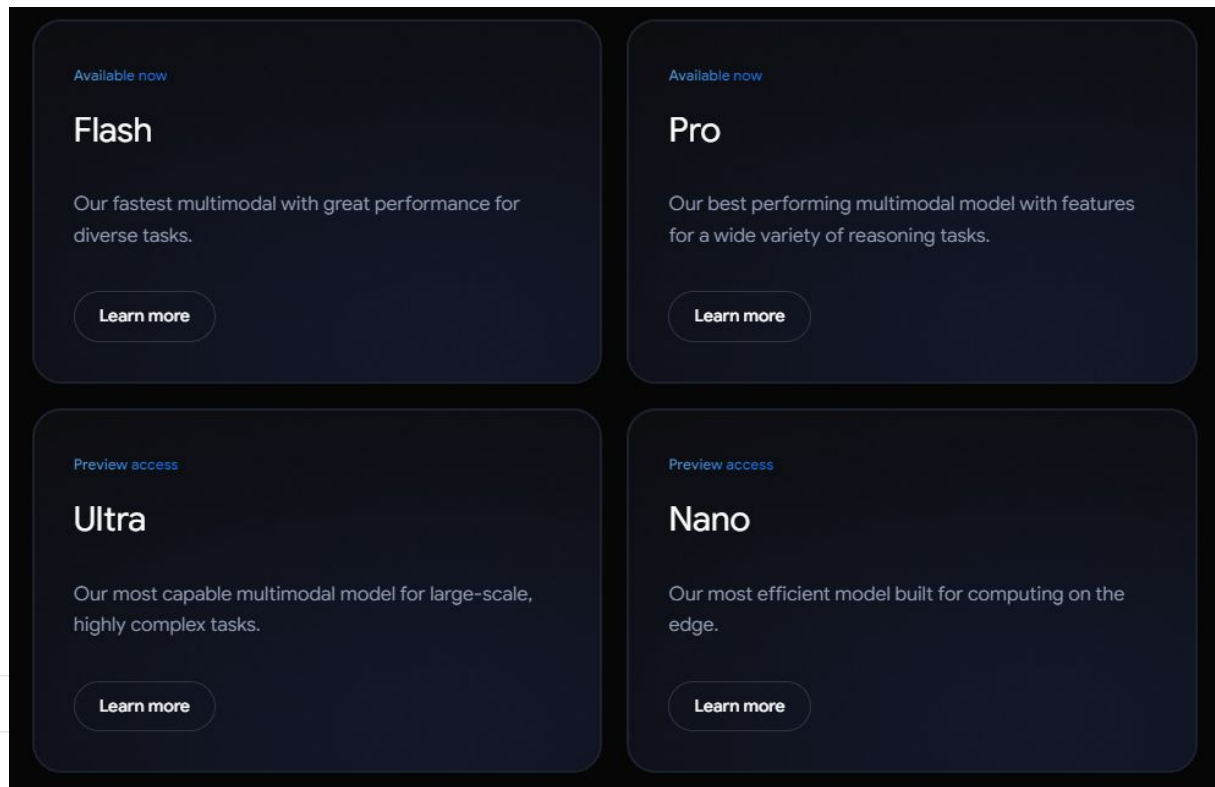
# Google is the industry pioneer in AI



“**Gemini** is Google’s latest  
multimodality LLM model -  
reasoning seamlessly across  
text, images, video, audio and  
code.”



# Gemini model sizes



Model Size	Availability	Description	Action
Flash	Available now	Our fastest multimodal with great performance for diverse tasks.	<a href="#">Learn more</a>
Pro	Available now	Our best performing multimodal model with features for a wide variety of reasoning tasks.	<a href="#">Learn more</a>
Ultra	Preview access	Our most capable multimodal model for large-scale, highly complex tasks.	<a href="#">Learn more</a>
Nano	Preview access	Our most efficient model built for computing on the edge.	<a href="#">Learn more</a>

**Any Open Source  
LLM Model?**



# Gemma



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

# Gemma Model Variants

## CodeGemma

Harnessing the foundation of our original pre-trained Gemma models, CodeGemma brings powerful code completion and generation capabilities in sizes fit for your local computer.

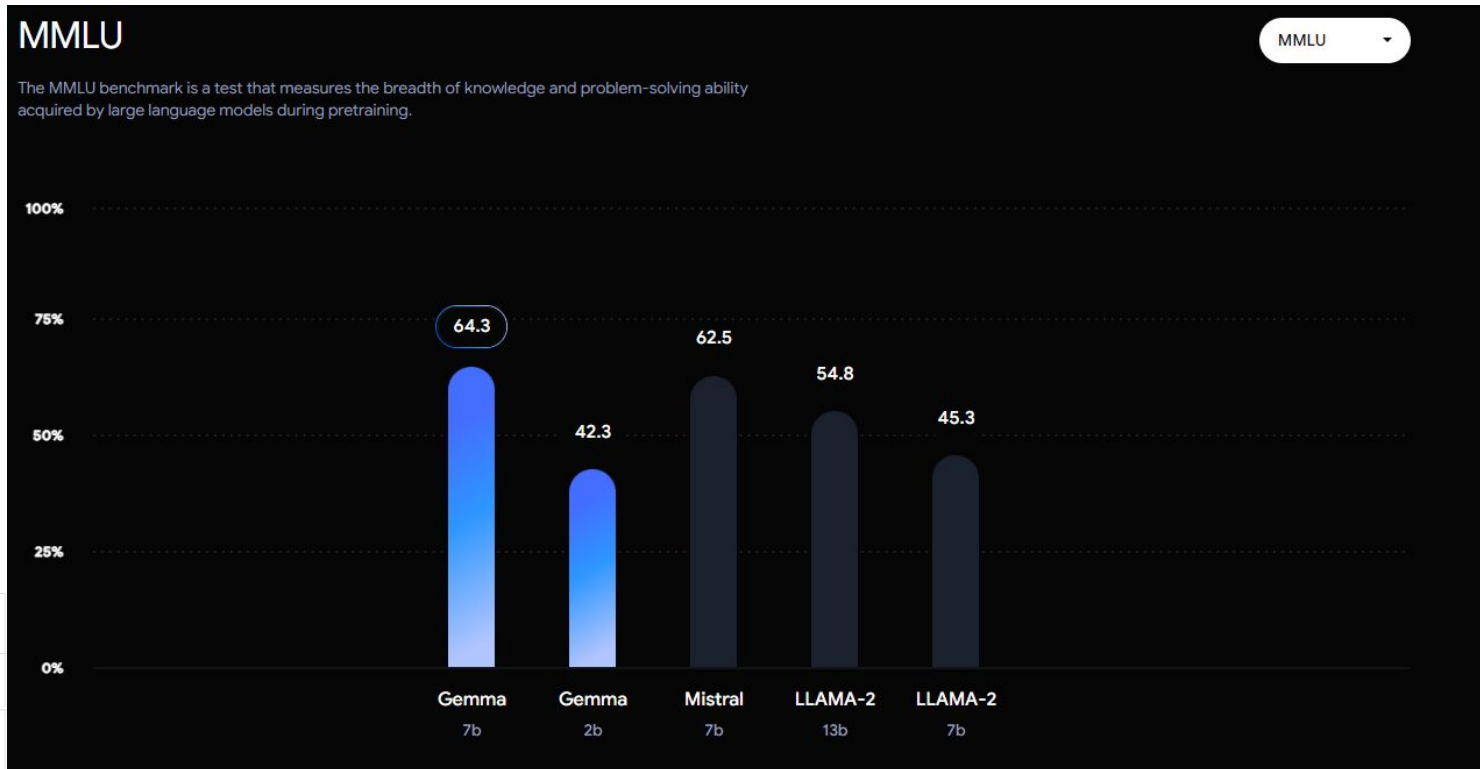
## PaliGemma

PaliGemma is an open vision-language model that is designed for class-leading fine-tune performance on a wide range of vision-language tasks.

## RecurrentGemma

RecurrentGemma is a technically distinct model that leverages recurrent neural networks and local attention to improve memory efficiency.

# Open LLM Models Benchmark



# Access Gemma Model



# Access Gemma Model



## Discover quickstarts on Kaggle

Visit the Kaggle Models page to find quickstarts, code examples, and discussions for Gemma.

[Open in Kaggle](#)



## Train and deploy on Google Cloud

Gemma works best on Google Cloud, with end-to-end TPU optimization for market-leading performance and total cost of ownership on Vertex.

[Open in Vertex AI](#)



## Try low-rank adaptation with JAX via Keras 3

Adapt Gemma models to your unique domain and data with the backend framework of your choice via Keras 3.

[Open in Colab](#)



## Hugging Face

Utilize Hugging Face Transformers and TRL for fine-tuning and inference tasks with Gemma models.

[View on Hugging Face](#)



## NVIDIA

Fine-tune Gemma models with NVIDIA NeMo Framework and export to TensorRT-LLM for production.

[View in Github](#)



## LangChain

This tutorial shows you how to get started with Gemma and LangChain, running in Google Cloud or in your Colab environment.

[Open in Colab](#)



## Anyscale

These docs show how to use Gemma via Anyscale Endpoint as fully managed API endpoints.

[View on Anyscale](#)



## MongoDB

This article presents how to leverage Gemma as the foundation model in a retrieval-augmented generation pipeline or system.

[View on MongoDB](#)



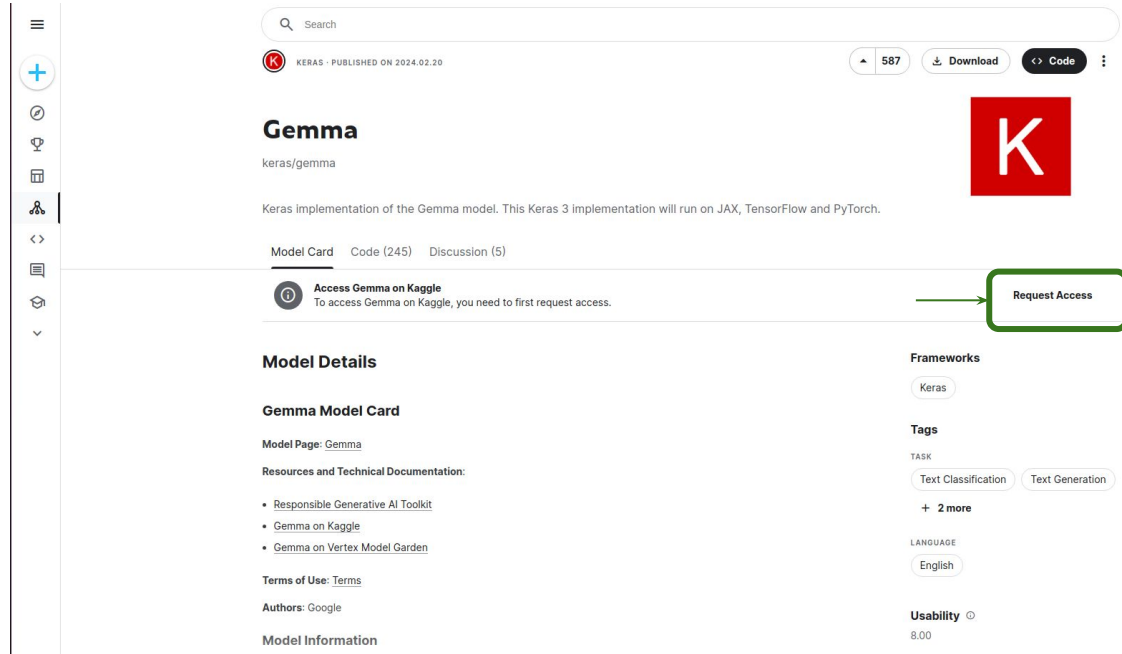
## Weights and Biases

Dive deep into W&B's Model Registry and Launch tools through a step-by-step example using Google's Gemma models.

[View on Weights and Biases](#)



# Access Gemma Model in Kaggle



The screenshot shows the Kaggle interface for the Gemma model. At the top, there's a search bar and navigation links for 'Download' and 'Code'. The main title 'Gemma' is followed by 'keras/gemma'. A red square with a white 'K' logo is prominent. Below the title, it states 'Keras implementation of the Gemma model. This Keras 3 implementation will run on JAX, TensorFlow and PyTorch.' There are tabs for 'Model Card', 'Code (245)', and 'Discussion (5)'. A green box highlights the 'Request Access' button, which is linked from the 'Access Gemma on Kaggle' section. The 'Model Details' section includes 'Gemma Model Card', 'Model Page: Gemma', and 'Resources and Technical Documentation' with links to 'Responsible Generative AI Toolkit', 'Gemma on Kaggle', and 'Gemma on Vertex Model Garden'. The 'Frameworks' section lists 'Keras'. The 'Tags' section shows 'Text Classification' and 'Text Generation'. The 'Usability' section shows a score of 8.00.

Search

587 Download Code

## Gemma

keras/gemma

Keras implementation of the Gemma model. This Keras 3 implementation will run on JAX, TensorFlow and PyTorch.

Model Card Code (245) Discussion (5)

**Access Gemma on Kaggle**  
To access Gemma on Kaggle, you need to first request access.

**Request Access**

### Model Details

#### Gemma Model Card

Model Page: [Gemma](#)

Resources and Technical Documentation:

- [Responsible Generative AI Toolkit](#)
- [Gemma on Kaggle](#)
- [Gemma on Vertex Model Garden](#)

Terms of Use: [Terms](#)

Authors: Google

Model Information

#### Frameworks

Keras

#### Tags

TASK

Text Classification Text Generation

+ 2 more


LANGUAGE

English

#### Usability

8.00

# Access Gemma Model in Kaggle




## Gemma

keras/gemma

Keras implementation of the Gemma model. This Keras 3 implementation will run on JAX, TensorFlow and PyTorch.

Model Card Code (245) Discussion (5)

 You've consented to the license for Gemma

[View License Consent](#)

### Model Details

#### Gemma Model Card

Model Page: [Gemma](#)

Resources and Technical Documentation:

- [Responsible Generative AI Toolkit](#)
- [Gemma on Kaggle](#)
- [Gemma on Vertex Model Garden](#)

Terms of Use: [Terms](#)

Authors: Google

#### Model Information

Summary description and brief definition of inputs and outputs.

#### Frameworks

Keras

#### Tags

TASK

Text Classification Text Generation


+ 2 more

LANGUAGE

English

#### Usability

8.00



You now have access to start building with Gemma

Explore our resources to get started with your favorite tools.

[Dismiss](#)

# Access Gemma Model in Kaggle

## Gemma

▲ 3469   Download   <> Code   ⋮

Model Card   Code (434)   Discussion (48)

### Model Variations

Keras   PyTorch   Transformers   Gemma C++   TensorRT-LLM   TensorFlow Lite   MaxText   Pax   Flax   GGUF

VARIATION

🔧 gemma\_instruct\_2b\_en (managed by  Keras) ▼

VERSIONS

Version 2

Download

+ New Notebook

This variation is maintained by the Keras organization. For more information about this variation, please refer to its dedicated model detail page.

#### Gemma

Keras

6 Variations · 245 Notebooks

Keras implementation of the Gemma mod...

▲ 587



### About Variation

FINE-TUNABLE

Yes

LICENSE


Gemma


BASE MODEL




[gemma · Keras · gemma\\_2b\\_en](#)







# Access Gemma Model in HF




The screenshot shows the Hugging Face interface for the `google/gemma-2b` model. At the top, the Hugging Face logo and a search bar are visible. A yellow banner promotes joining an organization. The model card for `google/gemma-2b` is displayed, showing it has 748 likes. Below the model name are tags for `Text Generation`, `Transformers`, `Safetensors`, `GGUF`, `gemma`, `Inference Endpoints`, and `text`. Navigation tabs for `Model card`, `Files and versions`, and `Community` (with 61 members) are present. A notice indicates that the user has been granted access to this gated model. The page title is `Gemma Model Card`, and the model page is identified as `Gemma`.


 **Hugging Face**


Hugging Face is way more fun with friends and colleagues!  [Join an organization](#)

 `google/gemma-2b`   like 748

 Text Generation  Transformers  Safetensors  GGUF `gemma`  Inference Endpoints  text

 **Model card**  Files and versions  Community **61**

 Edit model card

 **Gated model** You have been granted access to this model

**Gemma Model Card**

Model Page: [Gemma](#)

<https://huggingface.co/google/gemma-2b>

Google Developers

# Demo

Huggingface , Kaggle, Colab





Thank you.

**Manikandan Krishnamurthy**

Google Cloud Architect 📐 | 5x GCP Customer  
Delivery 🌐 | 2x GCP Certified 🏆 | 6x GCP He...



Sincere thank to “Leong Lai Fong” (GDE) for Getting Started Gemma guide.

```
function filterStudies([ studies, filterByOrg = false, filterByTopic = false ]){  
  return studies.filter(study => {  
    if (filterByOrg) {  
      return study.organization === 'Google';  
    }  
    if (filterByTopic) {  
      return study.topic === 'Gemma';  
    }  
    return true;  
  });  
}
```