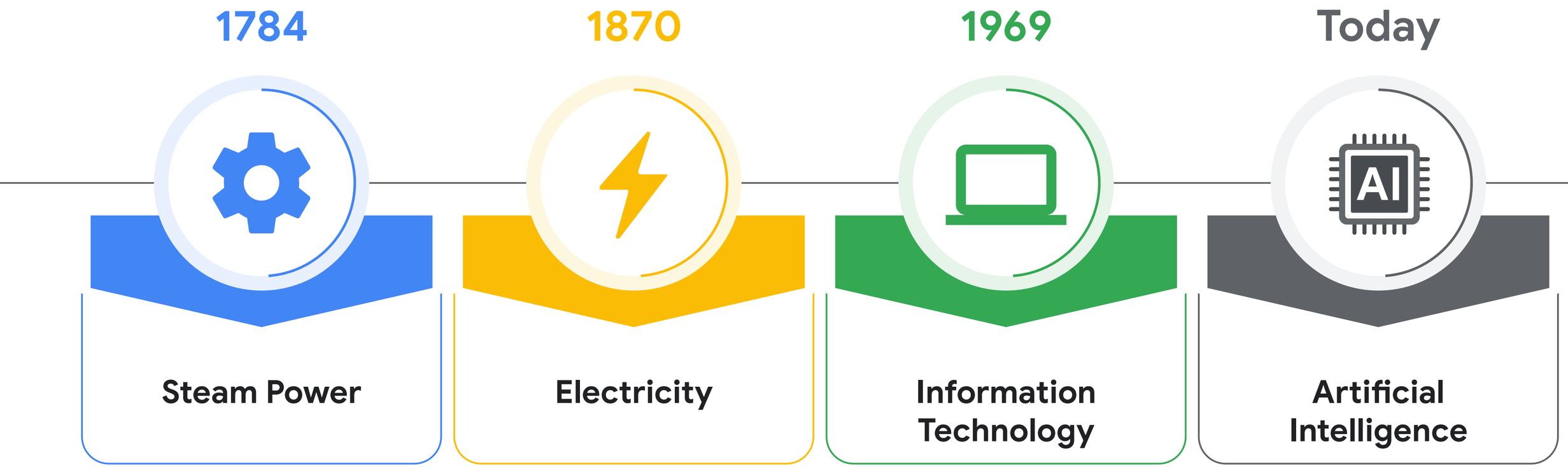
Welcome!

PRIMER ON LARGE LANGUAGE MODELS & GENERATIVE AI

Google Cloud

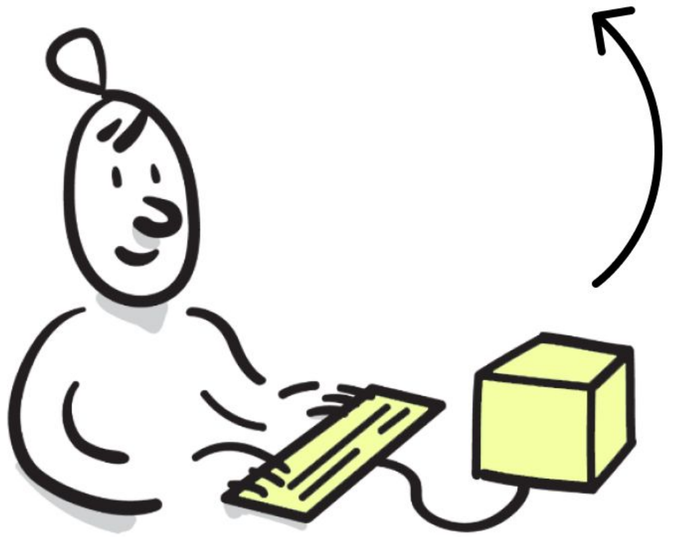
We’re in an AI-driven revolution



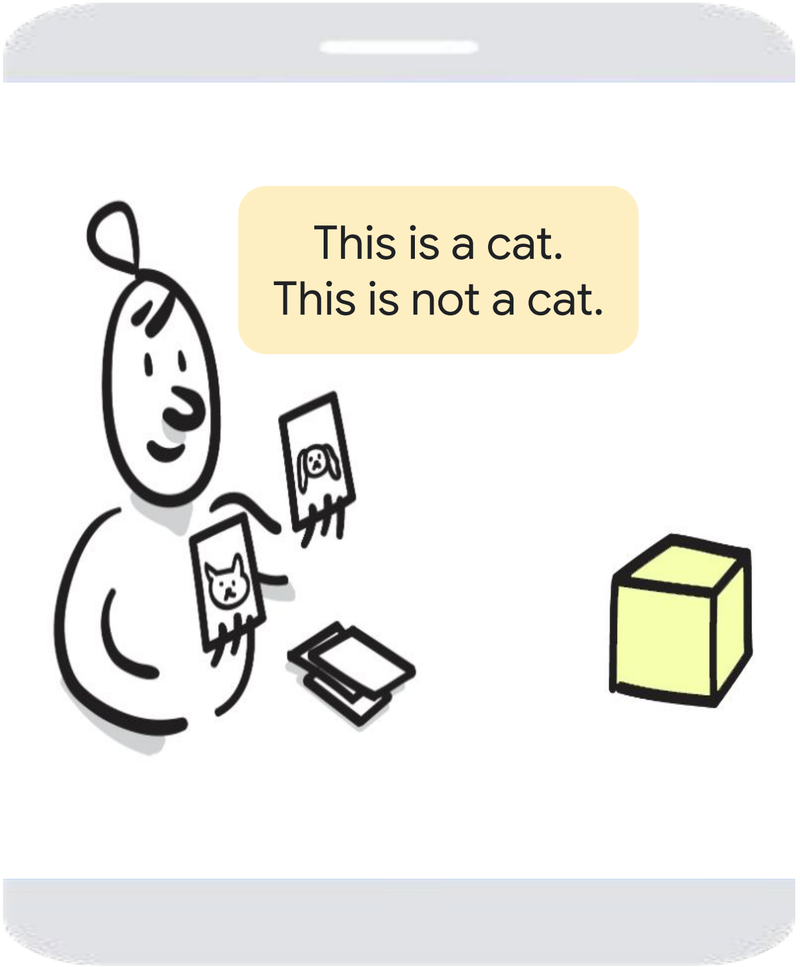
Source: AI: Recent Trends and Applications, Emerging Communication and Computing

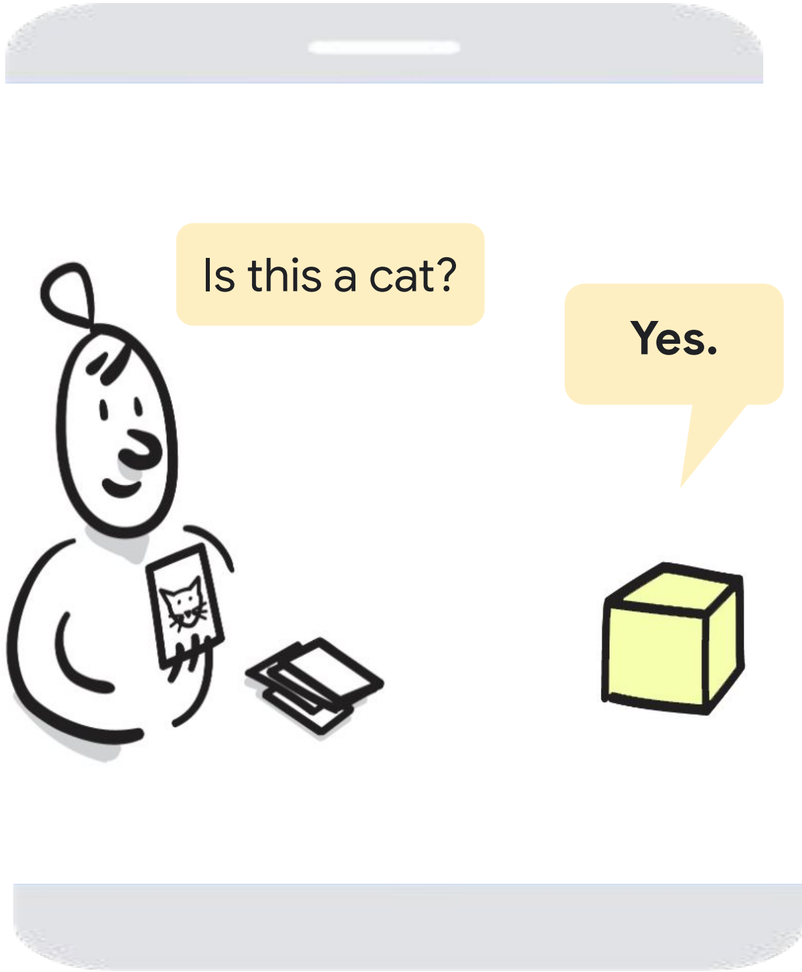
Traditional Programming

```
cat:
        type: animal
        legs: 4
        ears: 2
        fur: yes
        likes: yarn, catnip
```

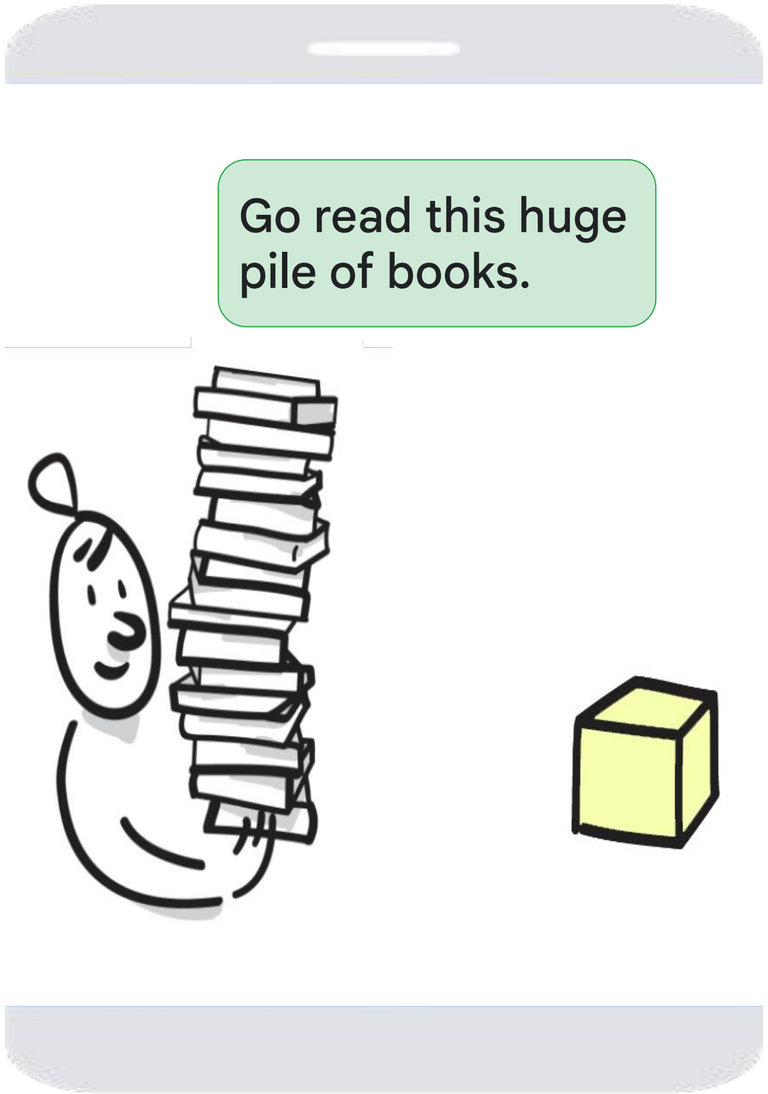


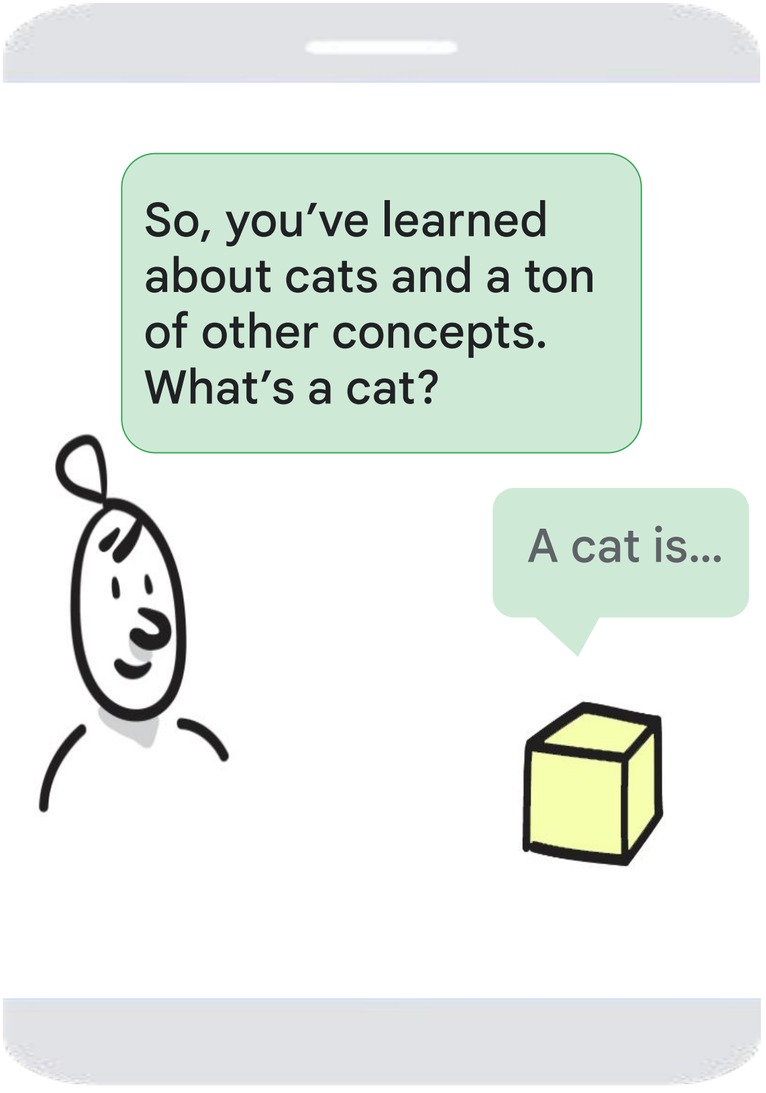
Wave of Neural Networks (~2012)





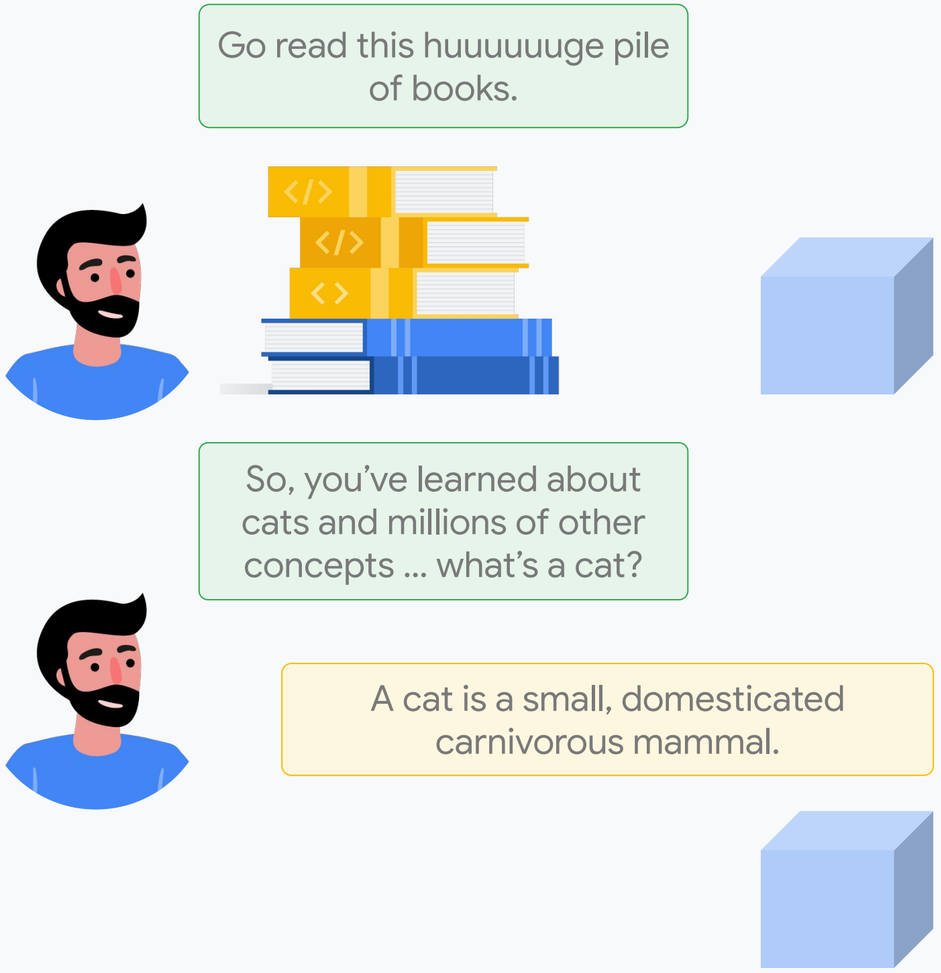
Generative Language Models (Now)





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)



Generative language models

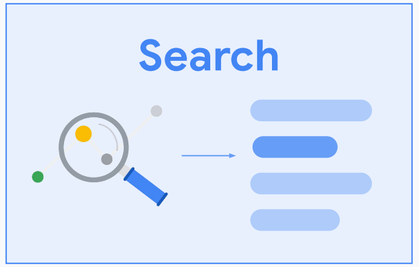
LaMDA, PaLM, GPT-3, etc.

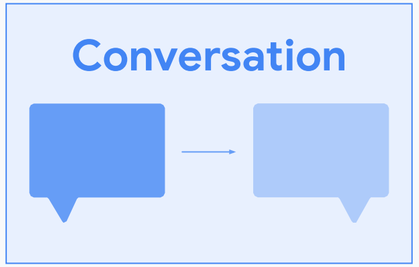
Why are large language models different?

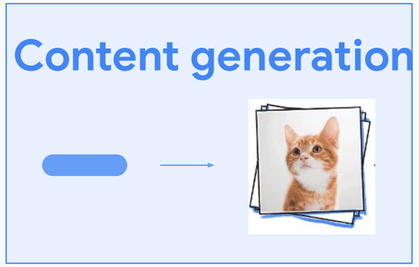
LLMs are characterized by **emergent abilities**, or the ability to perform tasks that were not included in their training examples.

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

Transformers

- Pathbreaking Neural Network Architecture
- Open Sourced by Google in 2017
- Started the revolution in Language Models

2018

BERT

(Bidirectional Encoder Representations from Transformers)

- World's first Language Model
- Open Sourced by Google in 2018
- SOTA on number of language benchmarks

2019

T5

(Text-to-Text Transfer Transformer)

- Large Language Encoder-Decoder Model
- 10-billion parameter model
- Open Sourced by Google in 2019

2020

LaMDA

(Language Model for Dialog Applications)

- Model trained on dialogue data
- Model could talk about virtually anything
- Published by Google in 2020

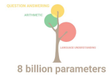
2021

Diffusion Models

- High Fidelity Image Generation Using Diffusion Models

PaLM

- (Pathways Language Model)
- Single model to generalize across domains



2022

CALM

(Confident Adaptive Language Modeling)

- Accelerating the text generation of LMs

2023

Bard

- A conversational AI service powered by LaMDA.

Vertex Gen AI

- Vertex AI: Gen AI Studio, Gen AI APIs, Model Garden , Foundation Model
- Generative AI App Builder: Conversation AI, Enterprise Search,

Prompt: The prompt is your text input that you pass to the model.

Prompt Design: The art and science of figuring out what text to feed your language model to get it to take on the behavior you want.

prompt design

= prompt engineering

= in context learning

= priming:

The art and science of figuring out what text to feed your language model to get it to take on the behavior you want.

Zero-shot prompt: The model is provided with **no example** when prompting for response.

One-shot prompt: The model is provided with **one example** to the LLM within the prompt to give some guidance on what type of response you want.

Few-shot prompt: Few-shot prompts are similar to one-shot prompts, but the model is given **multiple labeled examples** of the task.

WHAT IS POSSIBLE WITH GENERATIVE AI ON GOOGLE CLOUD?

Google Cloud

Consumers and enthusiasts

Draft a packing list for my weekend fishing and camping trip

Help me plan a neighborhood block party

Give me a list of idioms for “let’s circle back” that aren’t cringe

I want to write a novel. How do I get started?

Bard + MakerSuite

Enterprises

Allow data analysts to search and summarize market reports while **controlling our data**

Can we generate content while **controlling costs**?

Help my customers understand my financial products while being **safe, explainable, and regulatory compliant**

Vertex AI

Generative AI

Prototyping in the Generative AI Studio

Generative AI

Generative AI

Vertex AI

Access, train, deploy, and manage ML models

Model Garden

- Open Source models
- Task-specific AutoML & APIs
- Foundation models

Generative AI Studio

- Prompt design
- Prompt tuning
- Fine tuning

ML Platform

Data science toolkit
Notebooks + integration with data services

Fully-managed tools, workflows and infrastructure

End-to-end ML Ops

Generative AI

Vertex AI

Access, train, deploy, and manage ML models

Generative AI Studio

Prompt design

Prompt tuning

Generative AI

Generative AI Studio

Easy to use & low-code tooling for tuning and deploying foundation models

Generative AI GitHub Repository

Sample code and notebooks for GenAI on Google Cloud

goo.gle/gen-ai-github

Table of Contents

Learn more about **Generative AI**

at goo.gle/generativeai