Google Developer Group
London

Exploring Foundation Models with Model Garden and Vertex AI Studio

{ Build with AI }

Google Developer Group
London

Vinoth Arumugam

Principle Machine Learning Engineer

{ Build with AI }

> Artificial intelligence is the new electricity.

Andrew Ng

The AI Revolution: A Paradigm Shift

The Old Way: Building models from scratch for every single task. A long, resource-intensive, and data-hungry process.

The New Way: Using a single, massive "foundation model" as a base to build a wide range of applications.

Chapter One

Foundational Models

What Exactly are Foundation Models?

Answer - The Giants of AI

Large-scale, pre-trained AI models built on vast amount of data. They can perform a wide range of tasks, from image vision-language, audio, and video models.

Versatility

A single model for multiple tasks.

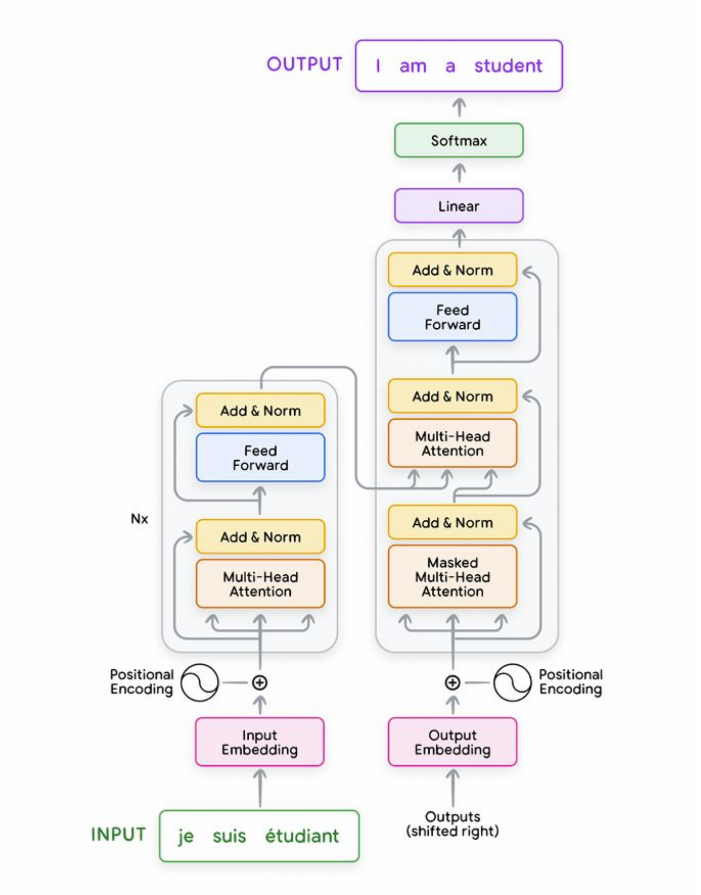
Adaptability

They can be fine-tuned for specific use cases with minimal data.

Emergent Abilities

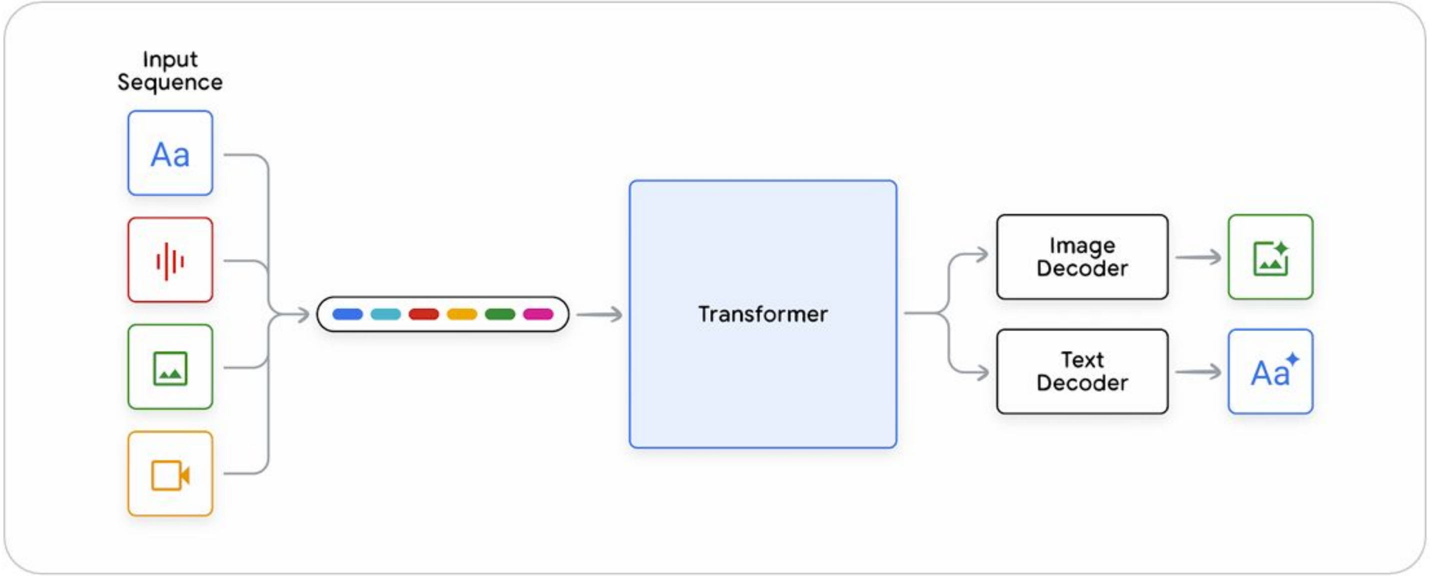
They exhibit capabilities not explicitly programmed, like zero-shot and few-shot learning.

Transformer Architecture



Gemini Overview

Gemini



Chapter Two

Meet **Foundational Model** (within Vertex AI)

The Two Pillars.

Model Garden

The hub for discovering and exploring models.

Vertex AI Studio

The workbench for building and deploying applications.

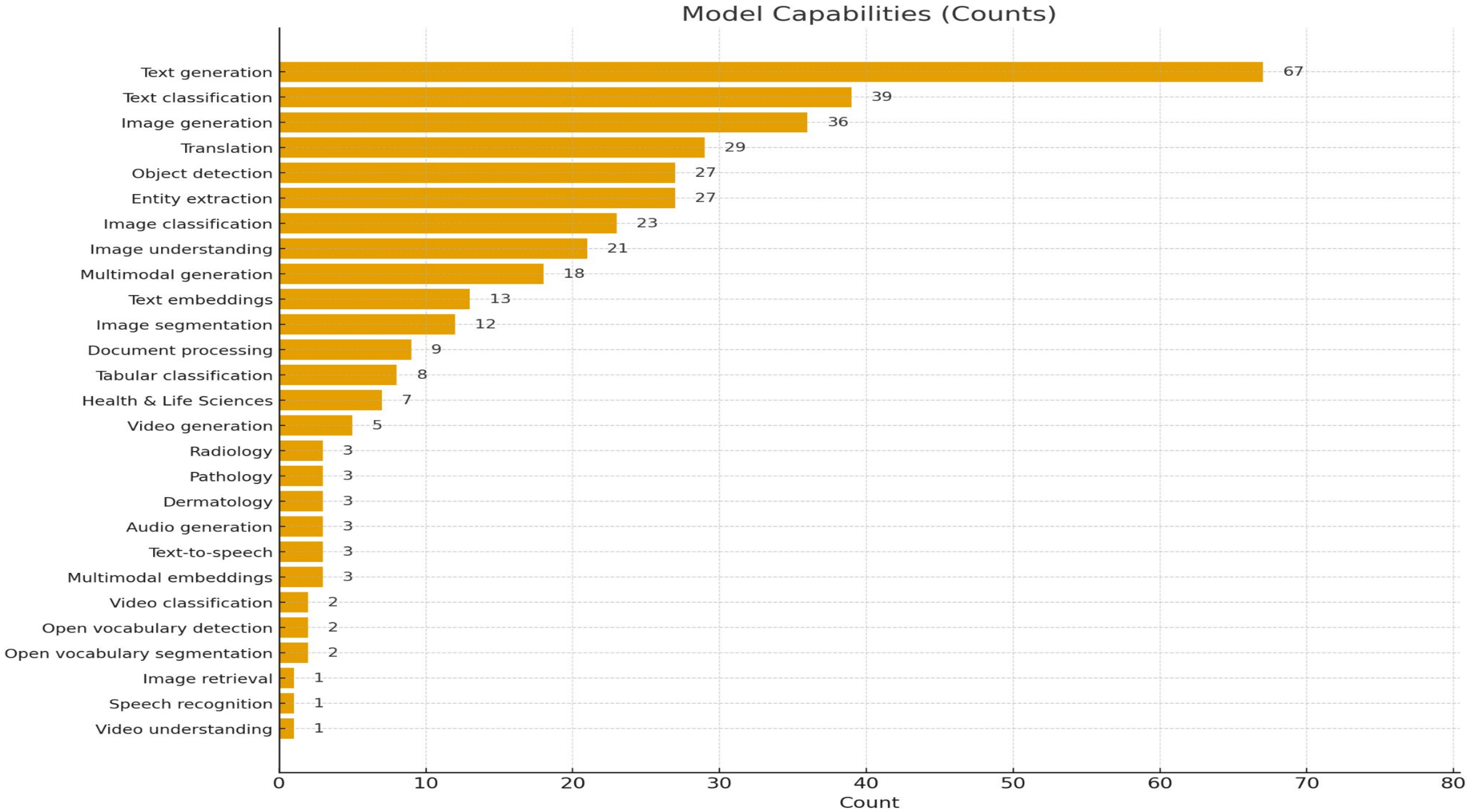
Model Garden

- **What it is:** A **curated library** to **discover, test, customize, and deploy** models from Google and partners.

Model Coverage

- **200+ foundation models** spanning Google, open-source, and third-party offerings from **9 partners.**

ANTHROP\C	Meta	Hugging Face	Mistral AI
Ai2	AI21labs	qodo	CAMB.AI
CSM			


Model Capabilities (Counts)
Text generation
67
Text classification
39
Image generation
36
Translation
29
Object detection
27
Entity extraction
27
Image classification
23
Image understanding
21
Multimodal generation
18
Text embeddings
13
Image segmentation
12
Document processing
9
Tabular classification
8
Health & Life Sciences
7
Video generation
5
Radiology
3
Pathology
3
Dermatology
3
Audio generation
3
Text-to-speech
3
Multimodal embeddings
3
Video classification
2
Open vocabulary detection
2
Open vocabulary segmentation
2
Image retrieval
1
Speech recognition
1
Video understanding
1
0
10
20
30
40
50
60
70
80
Count

Vertex AI Studio

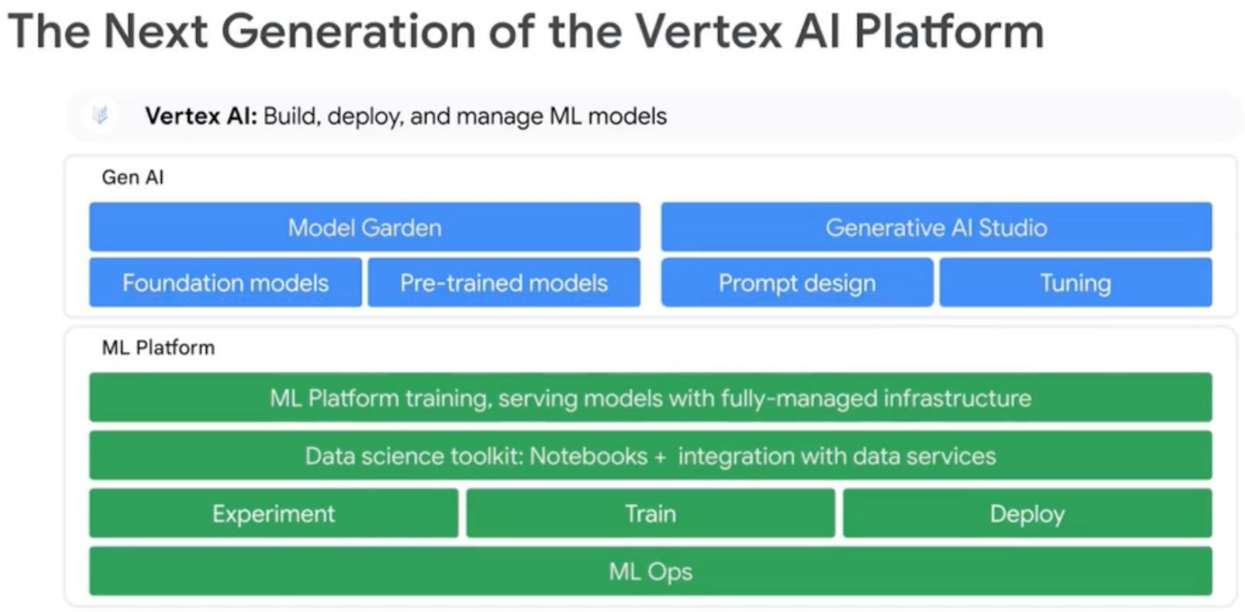
- **What it is:** A user-friendly, browser-based environment for interacting with foundation models.

Core Capabilities

- **Prompt playground:** Try prompts and tweak model settings without writing code.
- **Fine-tuning:** Adapt models with your own data to make them task-specific.
- **GCP integration:** Connect seamlessly with other Google Cloud services for end-to-end workflows.
- **Production handoff UI:** Purpose-built tools for prompt design, system instructions, grounding, and a one-click **Get code** export.

What’s New / Advanced Features

- **Prompt Optimizer (GA):** Improve and standardize prompt quality at scale.
- **Controlled generation:** Enforce output schemas (e.g., valid JSON).
- **Built-in safety:** Configurable filters and safety attributes surfaced directly in Studio.



The Next Generation of the Vertex AI Platform

Vertex AI: Build, deploy, and manage ML models

Gen AI

Model Garden

Foundation models

Pre-trained models

Generative AI Studio

Prompt design

Tuning

ML Platform

ML Platform training, serving models with fully-managed infrastructure

Data science toolkit: Notebooks + integration with data services

Experiment

Train

Deploy

ML Ops

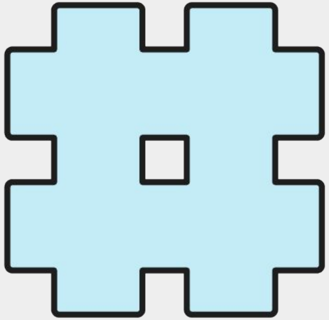
Partner Models as MaaS (model as a service)

Curated partner models offered as **managed APIs** via Vertex AI

Serverless: no provisioning or scaling to manage

You send requests to **Vertex AI endpoints**; partner served behind the scenes

Examples include **Claude** and **Mistral** families (availability varies)



Model Security Scanning (What’s Covered)

Google containers: Tested/benchmarked; **active vulnerability scanning**

Featured partners: Checkpoint authenticity scans

Hugging Face imports: Scans for **malware, pickle exec, Keras Lambda, secrets**

Policy: Unsafe → **blocked**; **suspicious/possible RCE** → **flagged** for your review

Responsible AI & Governance

Granular filters ⟶ Block hate speech, dangerous, explicit, or harassment content.

Configurable in Studio ⟶ No code needed, adjust directly in the playground.

Probability-based blocking ⟶ Harmful outputs stopped before reaching users.

Aligned with model settings ⟶ Safety sits alongside creativity controls (temperature, tokens, top-p).

Production ready ⟶ Same guardrails can be exported into code.

Chapter Three

UI walk through

Thank you!