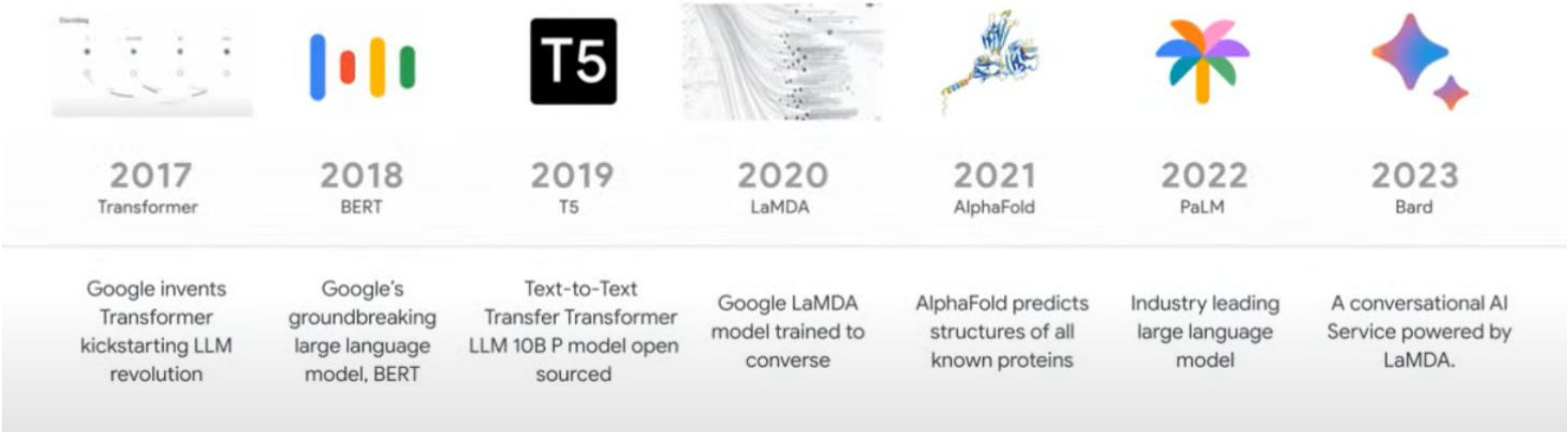
Generative AI with Google Cloud

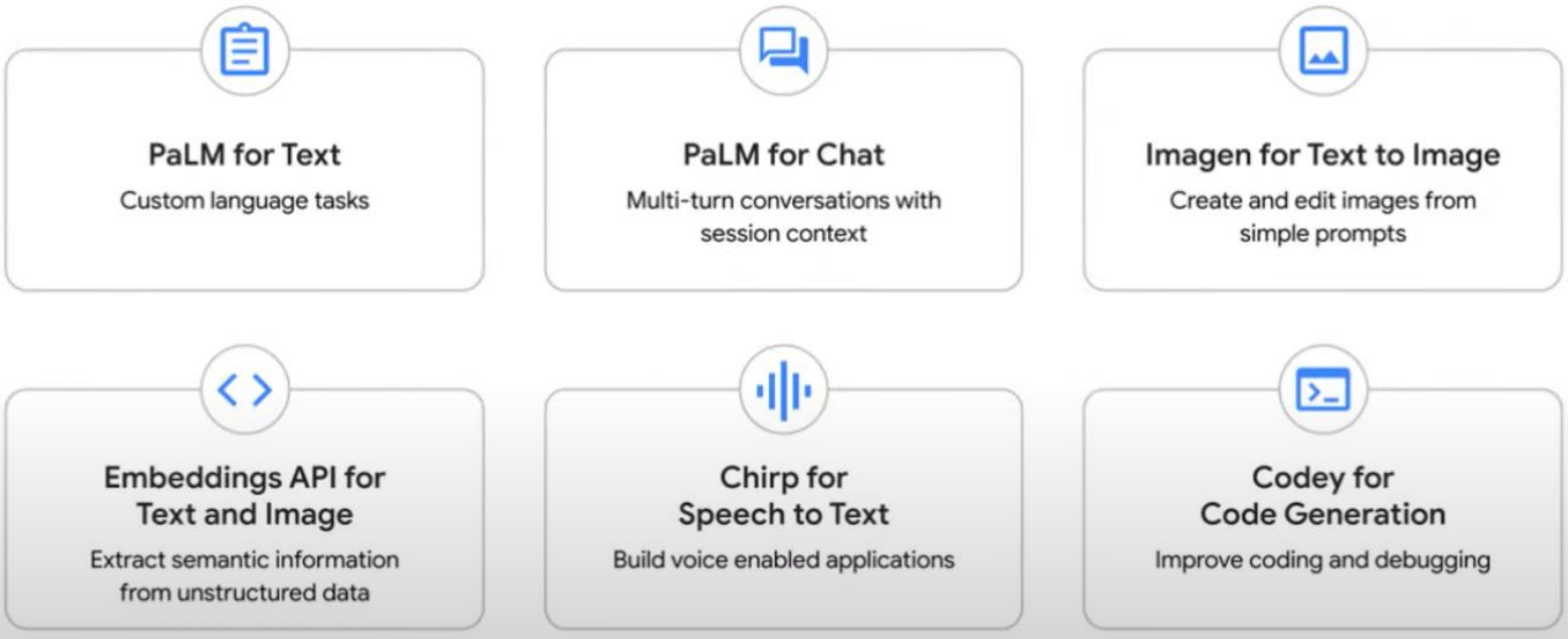
Vertex AI

Revolution at Google in AI



2017 Transformer	2018 BERT	2019 T5	2020 LaMDA	2021 AlphaFold	2022 PaLM	2023 Bard
Google invents Transformer kickstarting LLM revolution	Google's groundbreaking large language model, BERT	Text-to-Text Transfer Transformer LLM 10B P model open sourced	Google LaMDA model trained to converse	AlphaFold predicts structures of all known proteins	Industry leading large language model	A conversational AI Service powered by LaMDA.

First Party Foundation Models:



PaLM for Text

Custom language tasks

PaLM for Chat

Multi-turn conversations with session context

Imagen for Text to Image

Create and edit images from simple prompts

Embeddings API for Text and Image

Extract semantic information from unstructured data

Chirp for Speech to Text

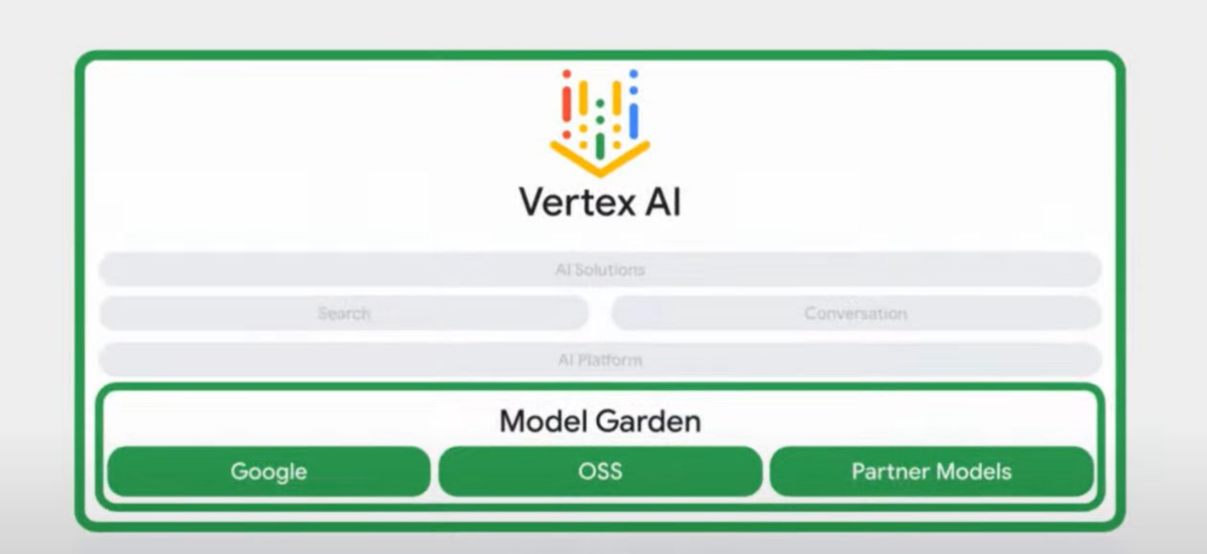
Build voice enabled applications

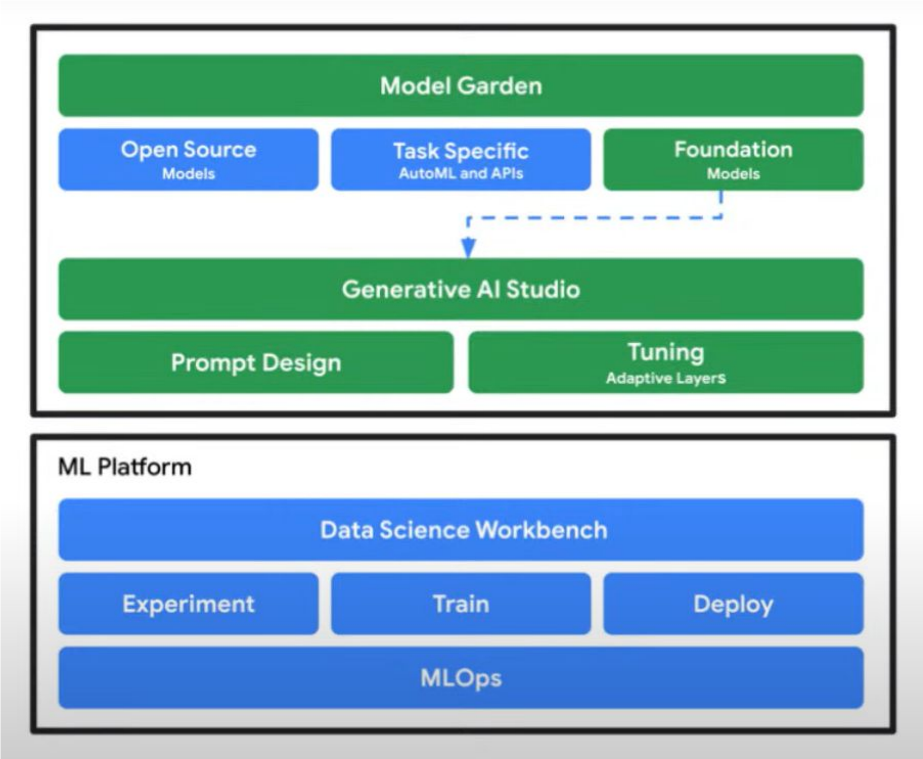
Codey for Code Generation

Improve coding and debugging

Model Garden:

Access, Customize, and Experiment with different foundation model and APIs.





Model Garden

Open Source
Models

Task Specific
AutoML and APIs

Foundation
Models

Generative AI Studio

Prompt Design

Tuning
Adaptive Layers

ML Platform

Data Science Workbench

Experiment

Train

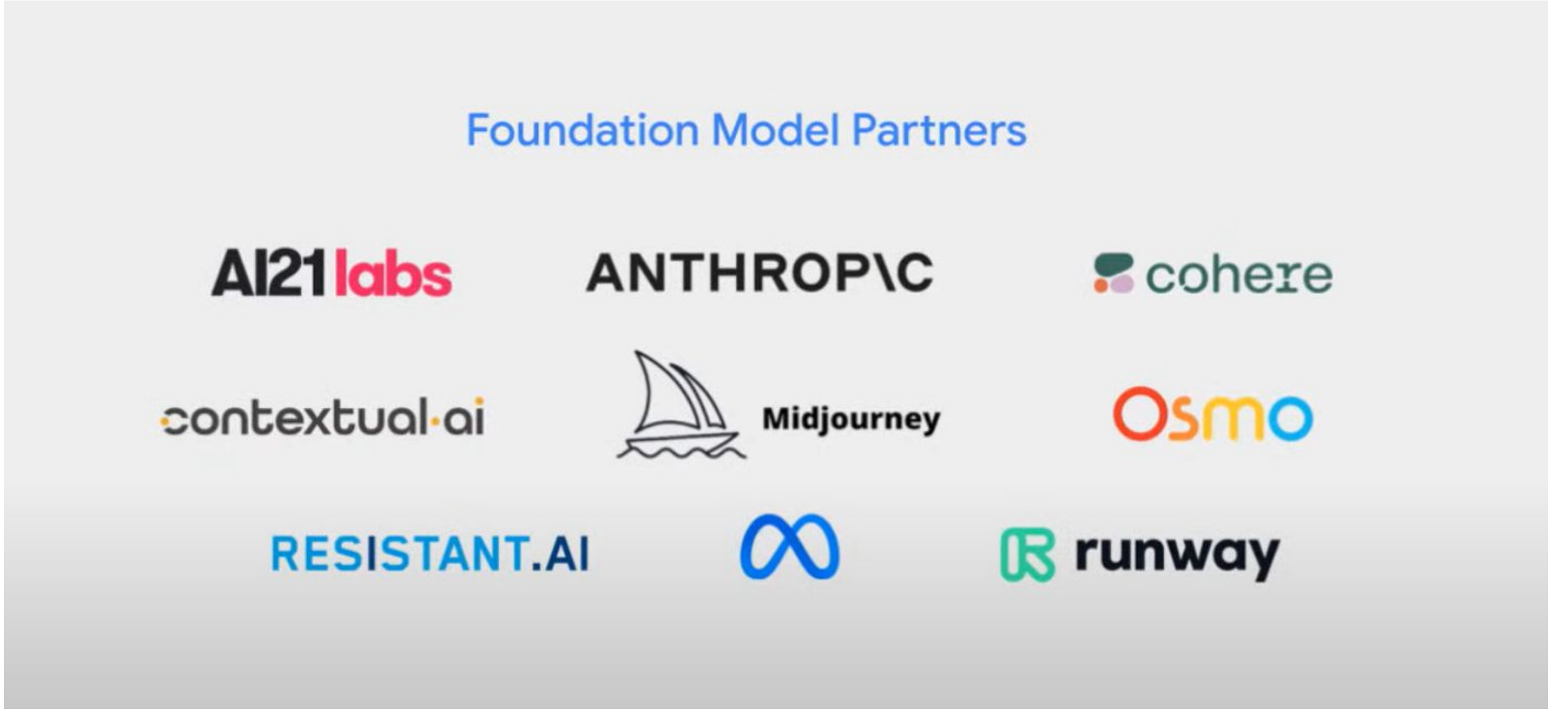
Deploy

MLOps

Vertex AI provides AI practitioners with tools to customize and build with Gen AI

vertex.ai

Open Model Ecosystem: 100+ Models



Foundation Model Partners

AI21labs

ANTHROP\C

cohere

contextual·ai

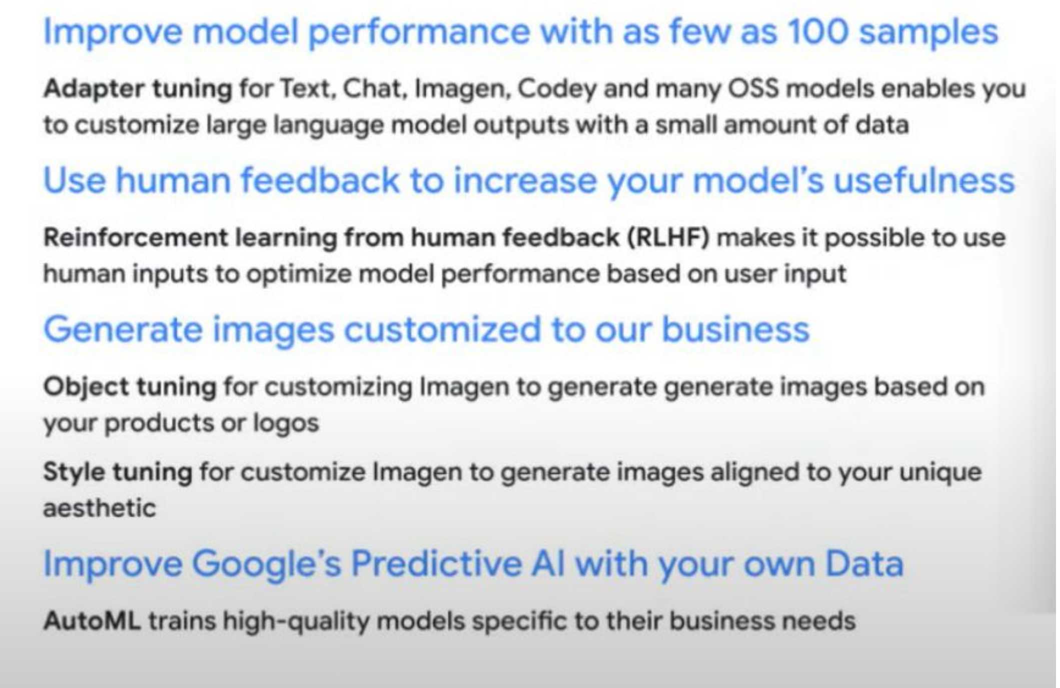
Midjourney

Osmo

RESISTANT.AI

runway

Tune Models to fit your use case:



Improve model performance with as few as 100 samples

Adapter tuning for Text, Chat, Imagen, Codey and many OSS models enables you to customize large language model outputs with a small amount of data

Use human feedback to increase your model's usefulness

Reinforcement learning from human feedback (RLHF) makes it possible to use human inputs to optimize model performance based on user input

Generate images customized to our business

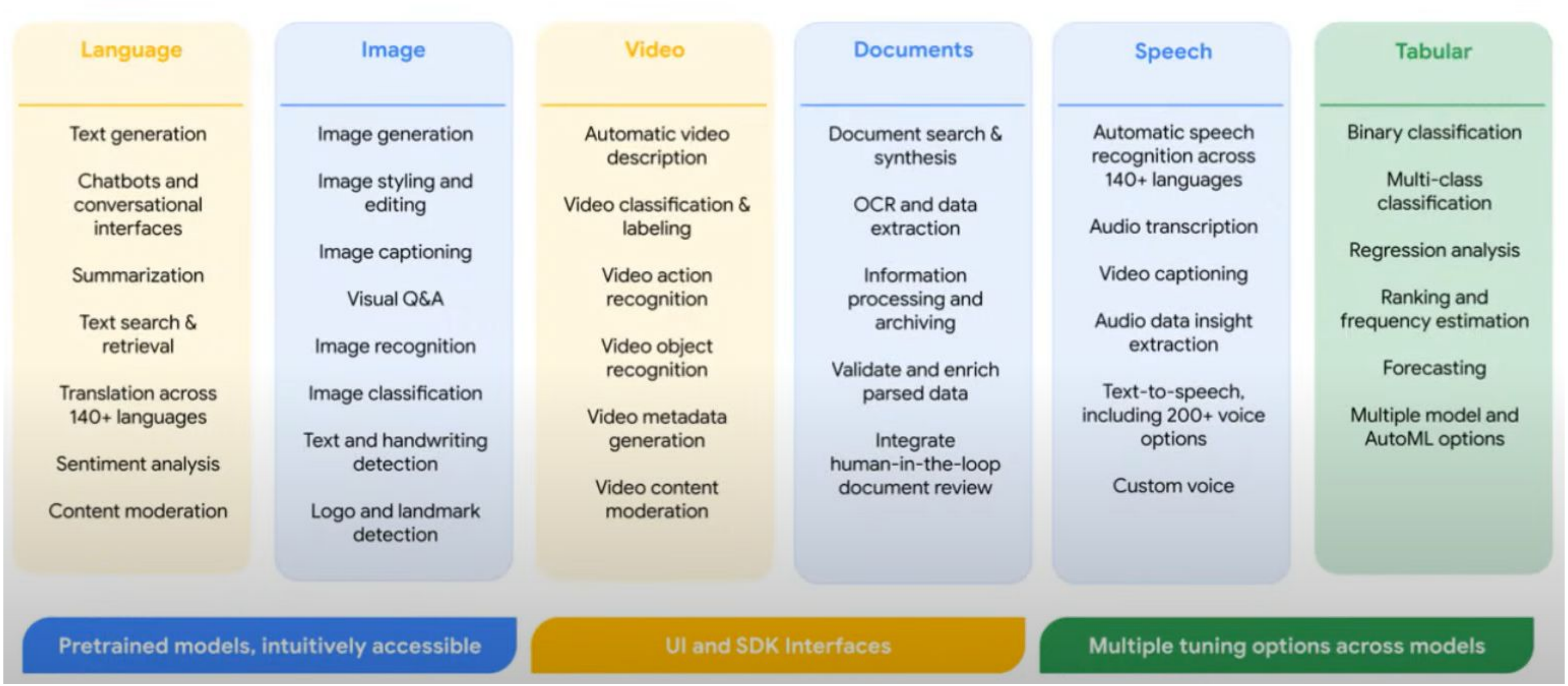
Object tuning for customizing Imagen to generate generate images based on your products or logos

Style tuning for customize Imagen to generate images aligned to your unique aesthetic

Improve Google's Predictive AI with your own Data

AutoML trains high-quality models specific to their business needs

All kinds of Tasks:



Language

- Text generation
- Chatbots and conversational interfaces
- Summarization
- Text search & retrieval
- Translation across 140+ languages
- Sentiment analysis
- Content moderation

Image

- Image generation
- Image styling and editing
- Image captioning
- Visual Q&A
- Image recognition
- Image classification
- Text and handwriting detection
- Logo and landmark detection

Video

- Automatic video description
- Video classification & labeling
- Video action recognition
- Video object recognition
- Video metadata generation
- Video content moderation

Documents

- Document search & synthesis
- OCR and data extraction
- Information processing and archiving
- Validate and enrich parsed data
- Integrate human-in-the-loop document review

Speech

- Automatic speech recognition across 140+ languages
- Audio transcription
- Video captioning
- Audio data insight extraction
- Text-to-speech, including 200+ voice options
- Custom voice

Tabular

- Binary classification
- Multi-class classification
- Regression analysis
- Ranking and frequency estimation
- Forecasting
- Multiple model and AutoML options

Pretrained models, intuitively accessible

UI and SDK Interfaces

Multiple tuning options across models

Vertex AI Model Garden is for GenAI Developers

Accelerate time to value by building on top of Google's world class infrastructure and services that help ensure security and reliability

State of the art

- Built on Google research and continuous innovation
- Best in class selection of 1P, OSS and 3P models

End-to-end governance

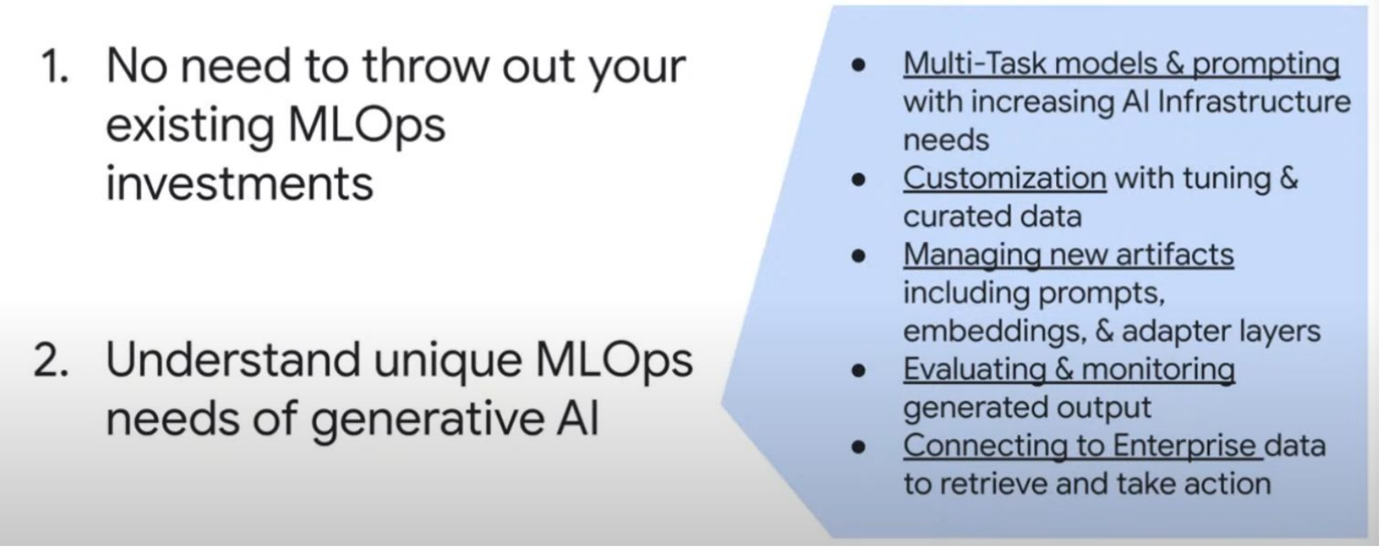
- **Prompting, Tuning & Distillation:** Customize LLMs for your domain and use case. Transfer and distill large scale learning to your models. Leverage Vertex AI for LLMOps managements
- **MLOps:** Leverage Vertex AI's capabilities for model evaluation, model management, prompt engineering, prompt management, and deployment

Enterprise readiness

- **Your Data, Your Terms:** Control and protect your data at every step of training and deployment
- **Responsible AI:** Tooling, enablement, and support to empower customers to build responsible Generative AI
- **Enterprise-ready, out of the box:** Accelerate time to value and developer efficiency with developer-friendly tooling built on enterprise security, reliability, and scalability

Manage Models with MLOPs & LLMOPs

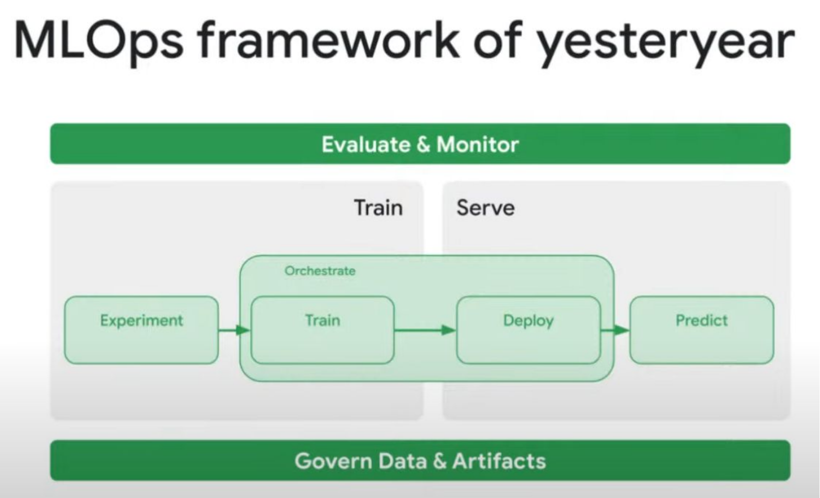
Its a set of standardized processes and capabilities for building, deploying,and operationalizing ML systems rapidly and reliably.



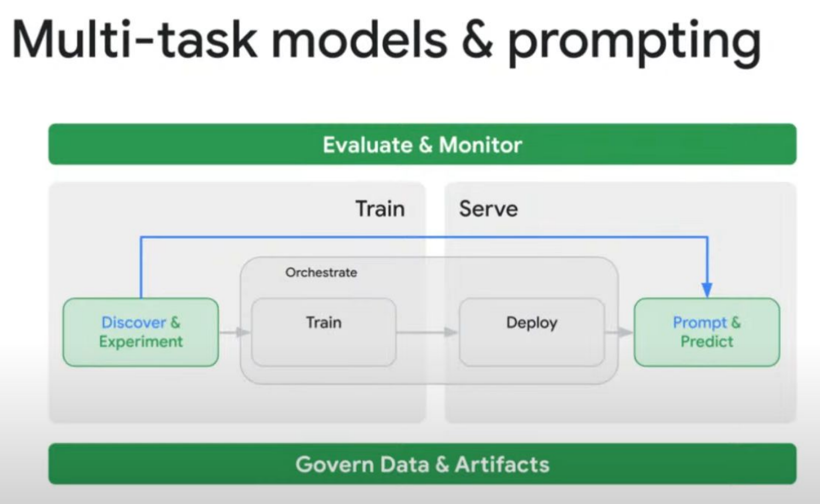
1. No need to throw out your existing MLOps investments
2. Understand unique MLOps needs of generative AI

- Multi-Task models & prompting with increasing AI Infrastructure needs
- Customization with tuning & curated data
- Managing new artifacts including prompts, embeddings, & adapter layers
- Evaluating & monitoring generated output
- Connecting to Enterprise data to retrieve and take action

MLOps on Vertex AI



MLOps framework of yesteryear

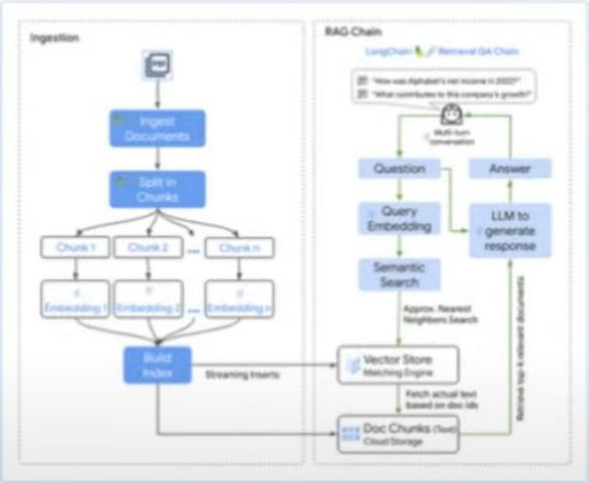


Multi-task models & prompting

Connecting to Enterprise Data

Embeddings & Vector Search

- Suite of Vertex embeddings APIs for Text & Image
- Fully managed vector database for high-scale low latency vector search
- Integrations with LangChain



https://cloud.google.com/blog/products/ai-machine-learning/generative-ai-applications-with-vertex-ai-palm-2-models-and-langchain

Grounding

Generate responses based on your data & provide citations to reduce hallucinations

Extensions

Build, access, & manage extensions to connect models to real-time data and real-world actions