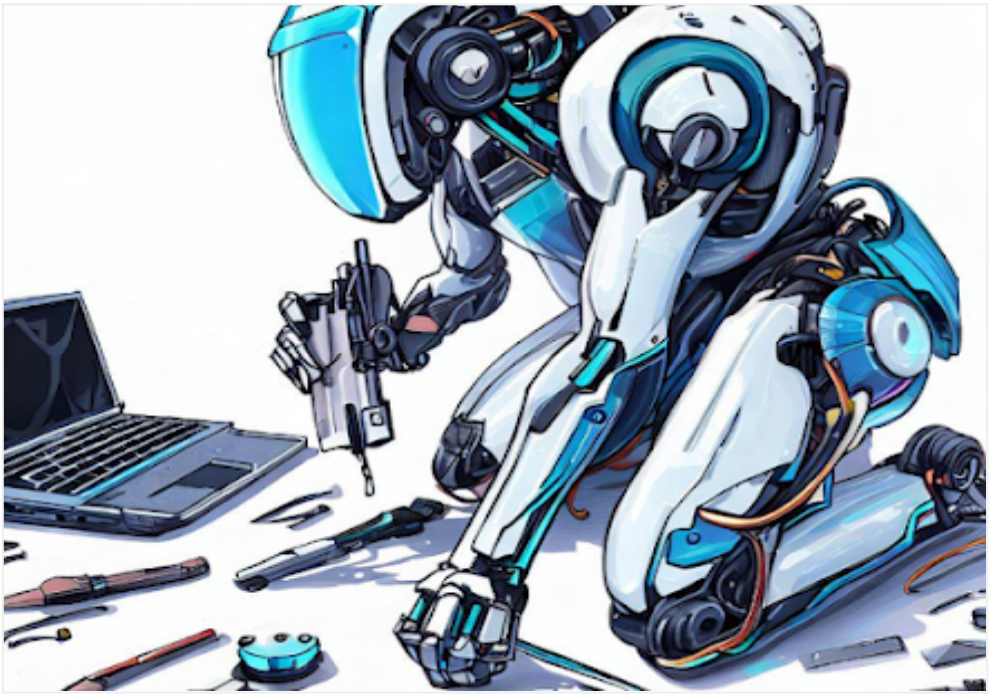
The Ultimate Guide to Generative AI Studio on Google Cloud's Vertex AI



Google is one of the leading companies in the field of artificial intelligence (AI), especially in the domain of generative AI, which is the ability to create new content or data from existing data.

Generative AI can be used for various tasks such as text generation, image synthesis, code completion, speech recognition, and more. Google has developed several foundation models, which are large and powerful neural networks that can learn from massive amounts of data and generate diverse outputs across different modalities. Some of these models are PaLM, Imagen, Codey, and Chirp, which can handle text, image, code, and speech generation respectively.

However, building applications with generative AI can be challenging for developers and data scientists who may not have the expertise or resources to interact with, tune, and deploy these foundation models. To

address this challenge, Google has recently launched Generative AI Studio, a managed environment in Vertex AI that makes it easy to use Google's foundation models as APIs and customize them with business data.

Vertex AI is a machine learning (ML) platform that lets users build, deploy, and scale ML models and AI applications. Vertex AI combines data engineering, data science, and ML engineering workflows, enabling teams to collaborate using a common toolset. Vertex AI provides various tools and services for different stages of the ML lifecycle, such as data labeling, model training, model tuning, model deployment, model monitoring, etc. Vertex AI also integrates with other Google Cloud products and services, such as BigQuery, Dataproc, Spark, etc.

What is Generative AI Studio?

Generative AI Studio is a low-code tool that provides a simple and intuitive user interface for prompt design, tuning, and deployment of generative AI models. It also integrates with Vertex AI's end-to-end ML platform and MLOps capabilities for scaling, managing, and governing generative AI models in production.

Few of the AI models are mentioned as below:

- **PaLM**: PaLM stands for Pre-trained Language Model. It is a multimodal foundation model that can generate natural language text for various purposes such as summarization, translation, question answering, chatbot dialogue, etc. PaLM can also generate text conditioned on other modalities such as images or tables. PaLM is available in two versions: PaLM 2 for Text and PaLM 2 for Chat. The former is optimized for generating long-form text such as articles or essays while the latter is optimized for generating short-form text such as conversational responses or captions.

- **Imagen**: Imagen is a foundation model that can generate realistic images from natural language text descriptions. Imagen can also generate images conditioned on other modalities such as sketches or emojis. Imagen can be used for various tasks such as image synthesis, image editing, image captioning, etc.
- **Codey**: Codey is a foundation model that can generate executable code from natural language text descriptions or pseudocode. Codey can also generate code conditioned on other modalities such as images or tables. Codey can be used for various tasks such as code completion, code synthesis, code documentation, etc.
- **Chirp**: Chirp is a foundation model that can generate natural language speech from text inputs. Chirp can also generate speech conditioned on other modalities such as images or emotions. Chirp can be used for various tasks such as speech synthesis, speech recognition, speech translation, etc.

Key Features of Generative AI Studio

Generative AI Studio offers several features that make it a unique and powerful tool for building generative AI applications. Some of these features are:

- Low-code generative AI: Users can access Google's multimodal foundation models as APIs with only a few lines of code and no ML background required. Vertex AI's managed endpoints make it easy to build generative capabilities into an application without worrying about the complexities of provisioning storage and compute resources or optimizing the model for inference.

- Intuitive user interface: Users can manually create text inputs or prompts that inform a foundation model with a familiar chat-like

experience that enables people without developer expertise to interact with a model. Users can also test sample prompts provided by Google or browse through a gallery of examples to get inspired by the possibilities of generative AI.

- Flexible tuning options: Generative AI Studio provides a wide range of capabilities for customizing foundation models with business data. Users can tune prompts with different parameters such as temperature, top-k, top-p, frequency penalty, presence penalty, etc. Users can also fine-tune model weights with their own datasets using Vertex AI's data labeling and training services. Users can also compare different versions of tuned models and evaluate their performance using metrics such as perplexity, accuracy, diversity, etc.

- Integration with end-to-end ML tools: Once deployed, foundation models can be scaled, managed, and governed in production using Vertex AI's end-to-end MLOps capabilities and fully managed AI infrastructure. Users can monitor model performance, track model lineage, audit model usage, enforce model governance policies, and more.

Capabilities/Use Cases of Generative AI Studio

Generative AI Studio enables users to interact with and tune a variety of foundation models that can handle different types of generative tasks across different domains. Some of these models are PaLM, Imagen, Codey, Chirp.

Some of the real-world examples of use cases that can be built with Generative AI Studio are:

- Restaurant chatbot: A chatbot that can interact with customers and provide information about the menu, prices, availability, etc. using PaLM 2 for Chat.
- Logo generator: A logo generator that can create logos from text descriptions or sketches using Imagen.
- Web app builder: A web app builder that can create web pages from text descriptions or pseudocode using Codey.
- Podcast creator: A podcast creator that can create audio content from text inputs or images using Chirp.

How to Access and Use Generative AI Studio

Generative AI Studio is a Google Cloud console tool that can be accessed from the Vertex AI section. Users need to have a Google Cloud account and a project with billing enabled to use Generative AI Studio. Users can also sign up for a free trial to get $300 credit to spend on Google Cloud services.

To use Generative AI Studio, users need to follow these steps:

1. **Select a foundation model**: Users can choose from a list of available foundation models such as PaLM, Imagen, Codey, or Chirp. Users can also see the model details such as description, input and output formats, supported languages, etc.
2. **Create a prompt**: Users can manually create a text input or prompt that informs the foundation model what to generate. Users can also use sample prompts provided by Google or browse through a gallery of examples to get inspired.
3. **Generate an output**: Users can click on the Generate button to see the output generated by the foundation model based on the prompt. Users can also see the model parameters such as temperature, top-k, top-p, etc. that control the output quality and diversity.
4. **Tune the prompt**: Users can tune the prompt with different parameters such as temperature, top-k, top-p, frequency penalty, presence penalty, etc. to get different outputs from the same

prompt. Users can also fine-tune the model weights with their own datasets using Vertex AI's data labeling and training services.

5. **Deploy the model**: Users can deploy the tuned model to a Vertex AI endpoint with a few clicks. Users can also specify the endpoint name, region, machine type, etc. Users can then use the endpoint URL and API key to integrate the model into their applications.

Generative AI Studio is now generally available to all Google Cloud users. The pricing for Generative AI Studio is aligned to how Vertex AI ML workloads are priced. Specific pricing will be available soon.

Limitations of Generative AI Studio

Generative AI Studio is a powerful and easy-to-use tool for building generative AI applications, but it also has some limitations that users should be aware of. Some of these limitations are:

- Data quality and quantity: The quality and quantity of data used to fine-tune the foundation models can affect the output quality and diversity. Users should ensure that their data is relevant, clean, and sufficient for their tasks.

- Model bias and ethics: The foundation models may have inherent biases or ethical issues due to the data they are trained on or the way they are designed. Users should be careful about how they use and interpret the outputs generated by the models and ensure that they do not harm or offend anyone.

- Model security and privacy: The foundation models may expose sensitive or confidential information from the data they are trained on or the prompts they are given. Users should ensure that they protect their data and models from unauthorized access or misuse.

Conclusion

Generative AI Studio is a new tool from Google Cloud that enables users to build generative AI applications with Google's foundation models in a low-code and intuitive way. Generative AI Studio is a promising tool for developers and data scientists who want to leverage generative AI for their applications without spending too much time and effort on ML engineering.

However, users should also be aware of the limitations and challenges of generative AI such as data quality and quantity, model bias and ethics, model security and privacy, etc.

Users should also follow best practices and guidelines for using generative AI responsibly and ethically.

source

https://cloud.google.com/vertex-ai/docs/generative-ai/learn/generative-ai-studio

https://cloud.google.com/ai/generative-ai

https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform