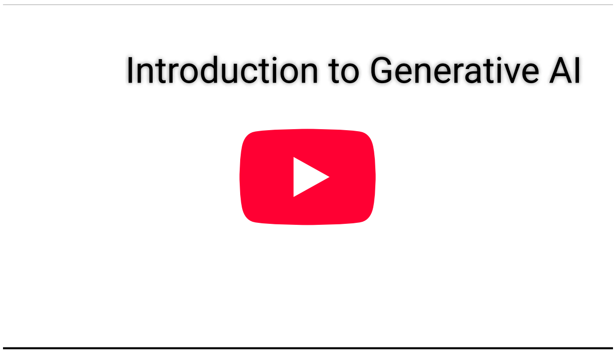
Overview of Vertex AI

Vertex AI is a unified, open platform for building, deploying, and scaling generative AI and machine learning (ML) models and AI applications. It provides access to the Model Garden, featuring a curated catalog of over 200 models—including Google's foundation models (such as Gemini) and a comprehensive selection of partner and open models—along with the underlying TPU/GPU infrastructure. Vertex AI supports cutting-edge GenAI workflows as well as AI inference workflows for MLOps. It offers end-to-end MLOps tools and enterprise-grade controls for governance, security, and compliance.

Key capabilities of Vertex AI

Vertex AI includes tools and services that support generative AI as well as AI inference and machine learning workflows.

Generative AI capabilities



Vertex AI brings together a comprehensive toolset with Google's advanced foundation models tools that you can use to build production-ready generative AI agents and applications, as follows:

- **Prompting**: Start with prompt design (/vertex-ai/generative-ai/docs/learn/prompts/introduction-prompt-design) in Vertex AI Studio (/vertex-ai/generative-ai/docs/start/quickstarts/quickstart). Vertex AI Studio includes tools for prompt design and model management that you can use to prototype, build, and deploy generative AI applications.
- **Models**: Vertex AI Model Garden (/vertex-ai/generative-ai/docs/model-garden/explore-models) is a centralized hub containing over 200 enterprise-ready models from Google, leading third-party partners (such as Anthropic's Claude), and popular open-source options (such as Llama).

  This selection of models includes the following:

  - **Google's foundational generative AI models** (/vertex-ai/generative-ai/docs/models):

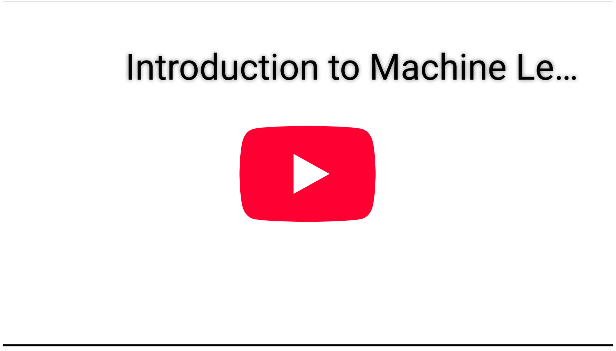
    - **Gemini** (/vertex-ai/generative-ai/docs/models/gemini): Multimodal capabilities for text, images, video, and audio; and thinking capabilities for models, such as Gemini 3 Flash and Gemini 3 Pro (with Nano Banana).
    - **Imagen on Vertex AI** (/vertex-ai/generative-ai/docs/models/imagen): Generate and edit images.
    - **Veo on Vertex AI** (/vertex-ai/generative-ai/docs/models/veo): Generate videos from text and images.
  - **Partner and open source models**: Access a curated selection of leading models such as Anthropic's Claude, Mistral AI models, and Llama with superior price-performance. These models are available as fully managed *model as a service (MaaS)* APIs.
- **Model customization**: Tailor models to your business to create unique AI assets. This ranges from Grounding with your enterprise data or Google Search to reduce hallucinations, to using Vertex AI Training for Supervised Fine-Tuning (SFT) or Parameter-Efficient Fine-Tuning (PEFT) of models like Gemini. For more information about model customization, see Introduction to tuning (/vertex-ai/generative-ai/docs/models/tune-models).
- **Generative AI Evaluations**: Objectively assess and compare model and agent performance with the Gen AI evaluation service (/vertex-ai/generative-ai/docs/models/evaluation-overview). Ensure safety and compliance by deploying runtime defense features like Model Armor (/security-command-center/docs/model-armor-overview) to proactively inspect and protect against emergent threats, such as prompt injection and data exfiltration.
- **Agent builders**: Vertex AI Agent Builder is a full-stack agentic transformation system that helps you create, manage, and deploy AI agents. Use the open-source Agent Development Kit (ADK) (/agent-builder/agent-development-kit/overview) to build and orchestrate agents, and then deploy them to the managed, serverless Vertex AI Agent Engine (/agent-builder/agent-engine/overview) for use at scale in production. Each agent is assigned an Agent Identity (Identity and Access Management Principal) for security and a clear audit trail.
- **Access External Information**: Enhance model responses by connecting to reliable sources with Grounding (/vertex-ai/generative-ai/docs/grounding/overview), interacting with external APIs using Function Calling (/vertex-ai/generative-ai/docs/multimodal/function-calling), and retrieving information from knowledge bases with RAG.

- **Responsible AI and Safety**: Use built-in safety features (/vertex-ai/generative-ai/docs/learn/responsible-ai) to block harmful content and ensure responsible AI usage.

For more information about Generative AI on Vertex AI, see the Generative AI on Vertex AI documentation (/vertex-ai/generative-ai/docs/overview).

AI inference capabilities

Vertex AI provides tools and services that map to each stage of the ML workflow:

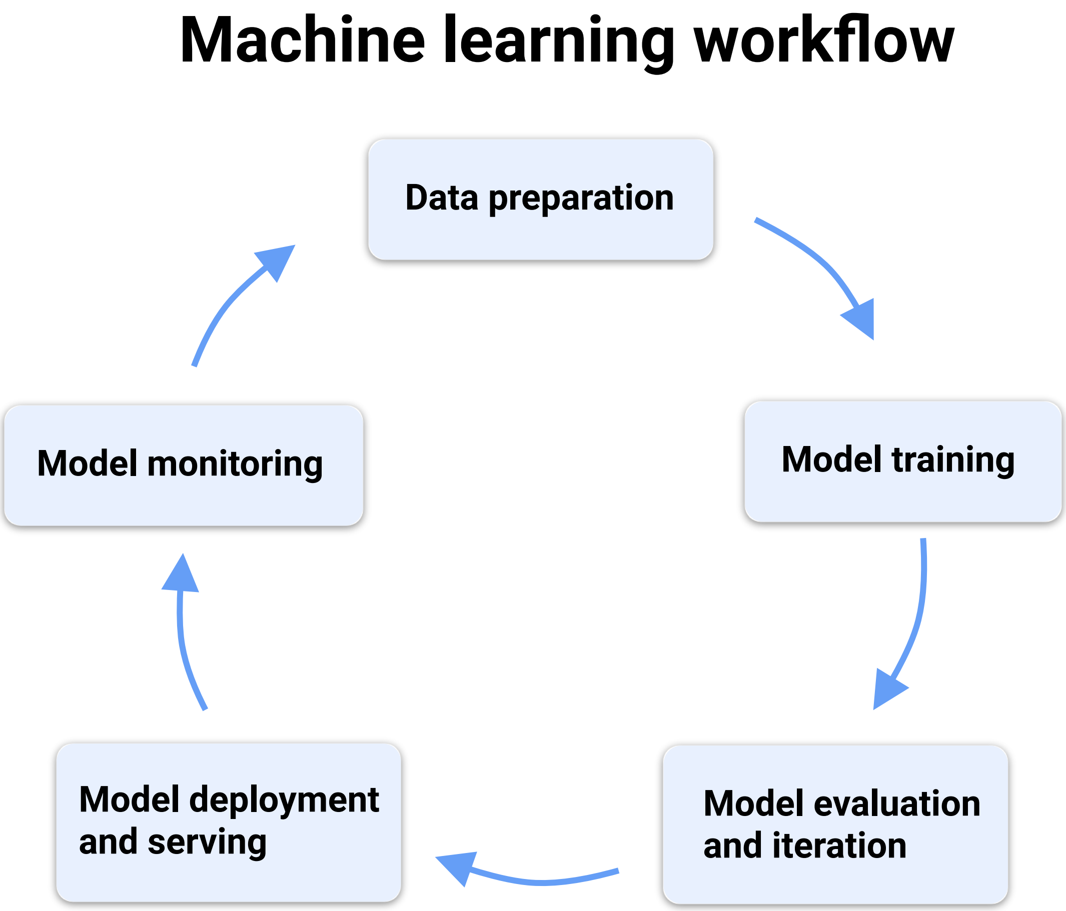


1. **Data preparation**: Collect, clean, and transform your data.
    - Use Vertex AI Workbench notebooks to perform exploratory data analysis (EDA) (/vertex-ai/docs/glossary#exploratory_data_analysis).
    - Integrate with Cloud Storage and BigQuery for data access.
    - Use Dataproc Serverless Spark (/dataproc-serverless/docs/overview) for large-scale data processing.
2. **Model training**: Train your ML model.
    - Choose between AutoML (/vertex-ai/docs/training-overview#automl) for code-free training or Custom training (/vertex-ai/docs/training/overview) for full control.
    - Manage and compare training runs using Vertex AI Experiments (/vertex-ai/docs/experiments/intro-vertex-ai-experiments).
    - Register trained models in the Vertex AI Model Registry (/vertex-ai/docs/model-registry/introduction).
    - Vertex AI Training (/vertex-ai/docs/training-overview) offers both serverless training and training clusters.
        - Use Vertex AI serverless training to run your custom training code on-demand in a fully managed environment. See the [Vertex AI serverless training overview][serverless].
        - Use Vertex AI training clusters for large jobs that need assured capacity on dedicated, reserved accelerator clusters. See Vertex AI training clusters

overview (/vertex-ai/docs/training/managed-training/overview).

- Use Ray on Vertex AI to scale Python and ML workloads with the open-source Ray framework on a managed, interactive cluster. See Ray on Vertex AI overview (/vertex-ai/docs/open-source/ray-on-vertex-ai/overview).
- Use Vertex AI Vizier (/vertex-ai/docs/vizier/overview) to adjust model hyperparameters in complex ML models.

3. **Model evaluation and iteration**: Assess and improve model performance.
   - Use model evaluation (/vertex-ai/docs/evaluation/introduction) metrics to compare models.
   - Integrate evaluations within Vertex AI Pipelines (/vertex-ai/docs/pipelines/introduction) workflows.
4. **Model serving**: Deploy and get inferences from your model.
   - Deploy for online inferences (/vertex-ai/docs/predictions/overview#online-inference) with prebuilt or custom containers.
   - Perform batch inferences (/vertex-ai/docs/predictions/overview#batch-inference) for large datasets.
   - Use Optimized TensorFlow runtime (/vertex-ai/docs/predictions/optimized-tensorflow-runtime) for efficient TensorFlow serving.
   - Understand model inferences with Vertex Explainable AI (/vertex-ai/docs/explainable-ai/overview).
   - Serve features from Vertex AI Feature Store (/vertex-ai/docs/featurestore/overview).
   - Deploy models trained with BigQuery ML (/vertex-ai/docs/beginner/bqml).
5. **Model monitoring**: Track deployed model performance over time.
   - Use Vertex AI Model Monitoring (/vertex-ai/docs/model-monitoring/overview) to detect training-serving skew and inference drift.


Machine learning workflow
Data preparation
Model training
Model evaluation and iteration
Model deployment and serving
Model monitoring

MLOps Tools

Automate, manage, and monitor your ML projects:

- **Vertex AI Pipelines** (/vertex-ai/docs/pipelines/introduction)**:** Orchestrate and automate ML workflows as reusable pipelines.
- **Vertex AI Model Registry** (/vertex-ai/docs/model-registry/introduction)**:** Manage the lifecycle of your ML models, including versioning and deployment.
- **Vertex AI serverless training** (/vertex-ai/docs/training/overview): Run your custom training code on-demand in a fully managed environment
- **Vertex AI Model Monitoring** (/vertex-ai/docs/model-monitoring/overview)**:** Monitor deployed models for data skew and drift to maintain performance.
- **Vertex AI Experiments** (/vertex-ai/docs/experiments/intro-vertex-ai-experiments)**:** Track and analyze different model architectures and hyperparameters.
- **Vertex AI Feature Store** (/vertex-ai/docs/featurestore/latest/overview): Manage and serve feature data for training models or making real-time predictions.

- **Vertex ML Metadata** (/vertex-ai/docs/ml-metadata/introduction)**:** Track and manage metadata for ML artifacts.
- **Vertex AI training clusters** (/vertex-ai/docs/training/managed-training/overview): Train large-scale jobs that require assured capacity on a dedicated, reserved cluster of accelerators.
- **Ray on Vertex AI** (/vertex-ai/docs/open-source/ray-on-vertex-ai/overview): Scale Python and ML workloads using the open-source Ray framework on a managed, interactive cluster.

What's next

- Dive into Generative AI on Vertex AI (/vertex-ai/generative-ai/docs/learn/overview).
- Learn about Vertex AI's MLOps features (/vertex-ai/docs/start/introduction-mlops).
- Explore interfaces that you can use to interact with Vertex AI (/vertex-ai/docs/start/introduction-interfaces).

Last updated 2026-01-22 UTC.