Google Cloud

Gen AI: Navigate the Landscape

Congratulations on completing the third course of the Gen AI Leader learning path. This course summary is your review guide. Print it for a handy reference as you continue your gen AI learning journey.

Before starting your gen AI project consider:

Needs:

- **Scale**: How many users will there be?
- **Customization**: How specialized is this AI?
- **User interaction**: How will users engage?
- **Privacy**: How sensitive is the data?
- **Latency**: What response time can you have?
- **Connectivity**: What is your connectivity?

Resources:

- **People**: Do you have access to AI expertise?
- **Money**: What's your budget?
- **Time**: What are your project timelines?

Platform: The foundation for building and scaling AI initiatives.

Vertex AI: Google Cloud's unified machine learning (ML) platform designed to streamline the entire ML workflow. It provides the infrastructure, tools, and pre-trained models you need to build, deploy, and manage your ML and generative AI solutions.

Infrastructure: The foundation upon which everything else rests. It provides the core computing resources needed for generative AI. This includes the physical hardware (like servers, GPUs, and TPUs), along with the essential software needed to train, store, and run AI models.

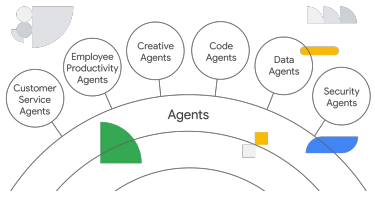
AI on the edge: You can run AI solutions on infrastructure (devices or servers) closer to where the action is happening.

Google provides tools like **Lite Runtime (LiteRT)** to help developers deploy AI models on edge devices.

Gemini Nano is Google's most efficient and compact AI model, specifically designed to run on devices.

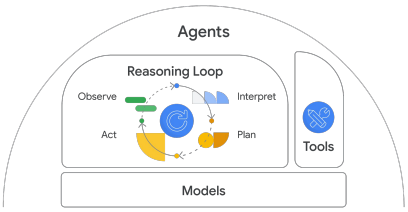
Gen AI landscape

- Gen-AI-powered applications
- Agent
- Platform
- Model
- Infrastructure



Agent: A gen AI agent is an application that tries to achieve a **goal** by **observing** the world and **acting** upon it using the tools it has at its disposal.

It does this using:



- **Reasoning loop:** An iterative process where the agent observes, interprets, reasons, and acts, often using prompt engineering
- **Tools**: Functionalities that allow the agent to interact with its environment, such as accessing and processing data or interacting with software or hardware.
- **Model:** The brains of the AI system, which consist of various algorithms that learn patterns from data and can make predictions or generate new content.

Vertex AI gives you options for how to handle AI models for your project.

- **Model Garden:** Pick from existing Google, third-party, or open-source models.
- **Model Builder:** Train and use your own models. Go fully custom and create and train models at scale using an ML framework. Or use AutoML to create and train models with minimal technical knowledge and effort.