Google Cloud

Generative AI Leader

Certification exam study guide

Table of contents

Introduction: Generative AI Leader

The Google Cloud Generative AI Leader training and exam are for professionals who need to understand how generative AI (gen AI) can be applied within a business context, particularly when using Google Cloud tools and services. This individual can identify use cases for gen AI, engage in discussions with technical and non-technical teams, and influence gen AI initiatives.

Exam focus

The exam assesses your knowledge in four key areas:

- Fundamentals of generative AI (~30% of the exam)
- Google Cloud's generative AI offerings (~35% of the exam)
- Techniques to improve gen AI model output (~20% of the exam)
- Business strategies for a successful gen AI solution (~15% of the exam)

This study guide is meant to be a starting point, rather than an exhaustive list of resources, as you prepare for the Generative AI Leader certification exam.

Google Cloud
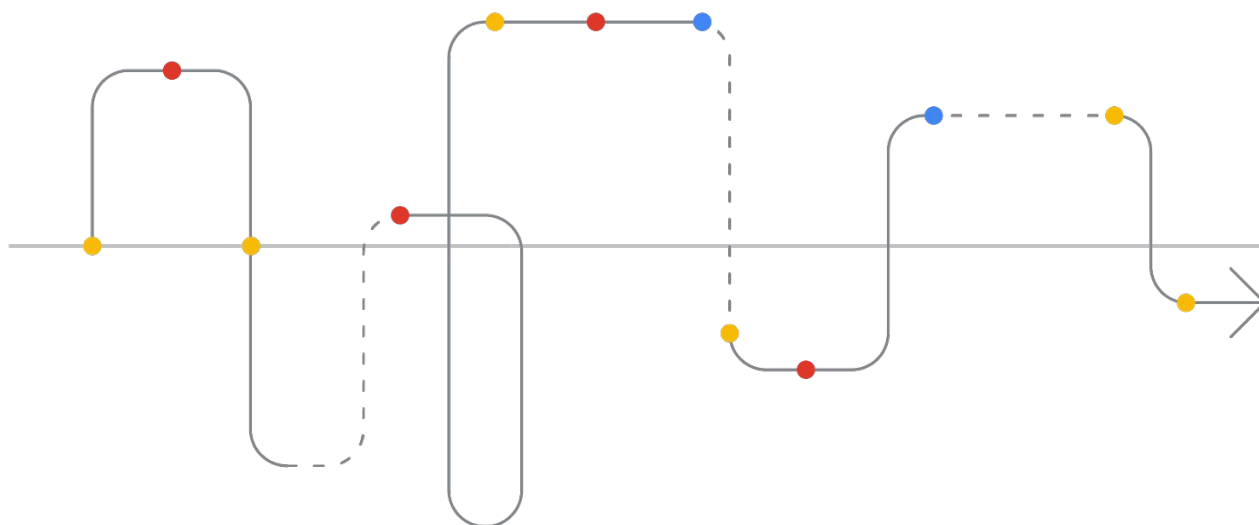
Fundamentals of generative AI

Artificial intelligence (AI): Building machines that can perform tasks that typically require human intelligence, such as learning, problem-solving, and decision-making.

Machine learning (ML): A subfield of AI where machines learn from data to perform specific tasks.

Generative AI: An application of ML that focuses on creating new content.

Deep learning: A subset of ML that uses artificial neural networks with many layers to extract complex patterns from data.

Foundation models: Powerful ML models trained on massive amounts of unlabeled data, allowing them to develop a broad understanding of the world.

Large language models (LLMs): A type of foundation model that is designed to understand and generate human language.

AI
ML
Gen AI
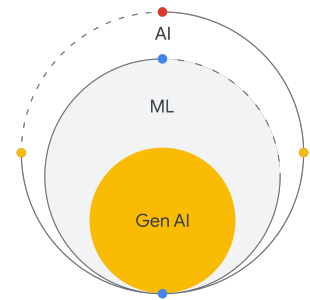
Generative AI is a type of AI that can create new content and ideas. Gen AI applications can be multimodal, enabling them to process and generate different types of data like text, images, and code simultaneously. Gen AI can:

Create
Generate new content

Summarize
Condense information into concise summaries

Discover
Find information at the right time

Automate
Automate previously manual tasks

Foundation models: Large AI models trained on massive datasets, allowing them to be adapted to many tasks. They are the basis of gen AI.

Key features of foundation models:
- **Trained** on diverse data.
- **Flexible** to a wide range of use cases.
- **Adaptable** to specialized domains through additional, targeted training.

Prompting: Prompting is the method of interacting with foundation models and guiding them. It involves providing them with instructions or inputs to generate desired outputs.

> Ask Gemini ▷

Prompt engineering: The art and science of creating effective inputs, known as prompts, for generative AI models to maximize their value and tailor responses to specific needs.

Labeled data: Data that has associated tags, such as a name, type, or number.

Unlabeled data: Raw, unprocessed information that hasn't been tagged and lacks meaning by itself such as unorganized photos or streams of audio recordings.

ML has **three primary learning approaches**:

Supervised learning trains models on labeled data to predict outputs for new inputs.

Unsupervised learning uses unlabeled data to find natural groupings and patterns.

Reinforcement learning learns through interaction and feedback to maximize rewards and minimize penalties.

Fundamentals of generative AI

Data is information that can come in many forms: numbers, dates, text descriptions, and even images or sounds.

- ◆ **Structured data:** Data that is organized and easy to search, often stored in relational databases.

- ◆ **Unstructured data:** Data that lacks a predefined structure and requires sophisticated analysis techniques.

- ◆ **Quality data:** Data that is accurate, complete, consistent, and relevant.

- ◆ **Accessible data:** Data for model training needs to be readily available, usable, and in the proper format.

Gen AI landscape

- **Gen-AI-powered application:** The user-facing part of generative AI. This is the layer that allows users to interact with and leverage the capabilities of AI.

- **Agent:** A piece of software that learns how to best achieve a goal based on inputs and tools available to it.

- **Platform:** This layer offers APIs, data management capabilities, and model deployment tools. It bridges the gap between models and agents while simplifying the complexities of infrastructure management.

- **Model:** A complex algorithm trained on vast amounts of data. It learns patterns and relationships in the data, allowing it to generate new content, translate languages, answer questions, and much more.

- **Infrastructure**: This layer provides the core computing resources needed for generative AI. This includes the physical hardware (like servers, GPUs, and TPUs) and software needed to store and run AI models and training data.

ML lifecycle

- **Data ingestion and preparation:** The process of collecting, cleaning, and transforming raw data into a usable format for analysis or model training.

- **Model training:** The process of creating your ML model using data.

- **Model deployment:** The process of making a trained model available for use.

- **Model management:** The process of managing and maintaining your models over time.
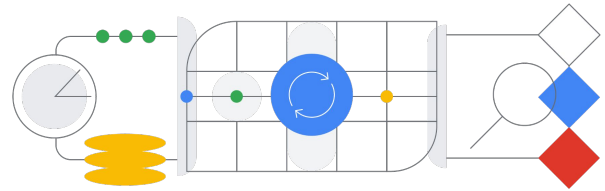
- **Gemini:** It supports multimodal understanding, advanced conversational AI, content creation, and question answering.

- **Gemma:** It offers developers user-friendly, customizable solution for local deployments and specialized AI applications

- **Imagen:** It's a text-to-image diffusion model that generates high-quality images from textual descriptions.

- **Veo:** It generates video content based on text descriptions or still images.

Resources to learn more

[A generative AI primer for the busy executives](#)

[Large language models (LLMs) with Google AI](#)

[Real-world gen AI use cases from the world's leading organizations](#)

Google Cloud's generative AI offerings

Google is an AI-first company

- Gen AI tools are integrated across Google's ecosystem.
- Google ensures you stay updated with the latest AI advancements.
- Google provides an ecosystem that puts security and ethics at the forefront.
- Google Cloud provides an enterprise-grade foundation you can build on.
- Google's open approach gives you flexibility and choice in your AI solutions.

Google's tooling for personal productivity

Gemini is a Google gen AI model that powers many different solutions.

The **Gemini app** is Google's generative AI chatbot, which provides assistance with tasks like writing, summarizing, translating, and creating images. With **Gemini Advanced**, companies can access extra features and enterprise-grade protections.

Gemini for Google Workspace integrates gen AI into familiar Workspace apps, allowing you to do things like compose emails in Gmail, generate images in Slides, and summarize notes in Meet.

Gemini for Google Cloud is your AI assistant for Google Cloud. It can help you write and debug code, manage and optimize cloud applications, analyze data in BigQuery, and strengthen your security posture.

Beyond Gemini, tools like NotebookLM address specific user needs, like document understanding.

NotebookLM allows you to upload your files and then acts as a research assistant, summarizing key points, answering questions, and generating ideas, all while staying grounded in your source material.

Vertex AI is Google Cloud's unified ML platform. It empowers you to build, train, and deploy ML applications. Vertex AI gives you access to generative AI models (such as Gemini) and lets you tune them to meet your needs, and then deploy them.

Vertex AI Search: Search and recommendation solutions for your business.

Use it with the Gemini API with **Google AI Studio** or **Vertex AI Studio**.

- Google AI Studio is available free of charge and is meant for quick AI prototyping.
- Vertex AI Studio is for building and deploying production-ready AI applications at scale.

Customer Engagement Suite: Tools to support your company in engaging with customers effectively. They can be built on top of Google's [Contact Center as a Service (CCaaS)](#), an enterprise-grade contact center solution that is native to the cloud.

- [Conversational Agents](#): Act as effective chatbots to your customers.
- [Agent Assist](#): Support your live human contact center agents.
- [Conversational Insights](#): Gain insights into all your communications with customers.

Google Agentspace: Integrate customized search and conversation agents that can access and understand data from various internal sources into your organization's internal websites or dashboards.

Tooling

- **Extensions:** Connect to external services (via APIs).
- **Functions:** Define specific actions or tasks.
- **Data stores:** Provide access to information.
- **Plugins:** Add new skills and integrations

Google Cloud's generative AI offerings

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.

Agent components:

Agent | Reasoning Loop | Tools | Model
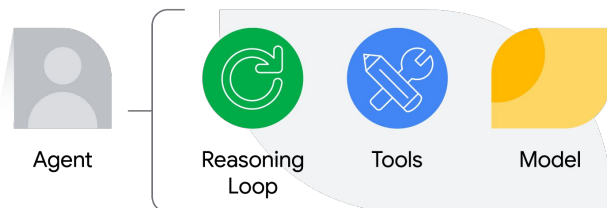
Reasoning loop: An iterative process where the agent observes, interprets, reasons, and acts, often using prompt engineering.
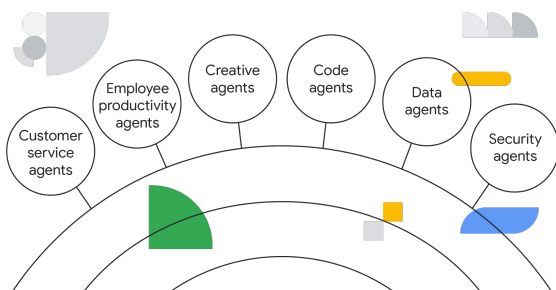
- **Tools**: Functionalities that allow the agent to interact with its environment, such as accessing and processing data or interacting with hardware.

- **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.

Types of agents

- **Deterministic (traditional):** Agents that are built with predefined paths and actions.

- **Generative:** Agents that are defined with natural language using LLMs to give a real conversational feel to your chatbot.

- **Hybrid Agents:** These agents combine both deterministic and generative capabilities, and this combination makes them very powerful.

Agents can serve several different functions.

Examples:

Customer service agents | Employee productivity agents | Creative agents | Code agents | Data agents | Security agents

Platform: The foundation for building and scaling AI initiatives.

As Google Cloud's unified machine learning (ML) platform, Vertex AI is 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.

With Vertex AI MLOps tools, AI teams can better collaborate to monitor and improve their models.

Model: 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.

Infrastructure: 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.

Google Cloud

Google Cloud's generative AI offerings: APIs

Speech-to-Text API

- The API converts speech into text.
- It also transcribes audio and video content.

Text-to-Speech API

- It converts text to natural-sounding speech.
- The API also creates voice user interfaces and personalized communication.

Translation API

- The Translation API translates text, documents, websites, audio and video files.

Document Translation API

- It translates formatted documents while keeping the original layout.

Document AI API

- The Document AI API extracts data from varied formats.
- It automates data capture and document processing.
- The API can also summarize documents.

Cloud Vision API

- The API analyzes image content, tagging images based on detected objects and text.
- It can also identify faces and landmarks.
- The API supports use cases like content moderation and visual search.

Cloud Video Intelligence API

- Allows developers to analyze video content and extract meaningful info
- Content recommendation, video search, and media analysis

Natural Language API

- Helps derive insights from unstructured text
- Understand the sentiment of text, classify content, and extract important entities

Building applications from your agents

You can access the Gemini API via tools like:

- Google Cloud developer tools like Cloud Run functions and Cloud Run
- Low code and no code tooling like Apps Script and AppSheet

Resources to learn more

Bringing AI agents to enterprises with Google Agentspace

Search from Vertex AI

What is AI Applications?

Approachable AI: Get started with Apps Script & Gemini

Techniques to improve generative AI model output

Prompting techniques

Zero-shot: Asking the model to complete a task with no prior examples.

One-shot: Providing the model with one example to learn from.

Few-shot: Giving the model multiple examples to learn from.

Role: Assigning a persona to the model to influence its style, tone, and focus.

Prompt chaining: Engaging in a back and forth conversation with the AI.

Reasoning loop: Prompt engineering techniques

- **ReAct (reason and act):** Allow the LLM to reason and take action on a user query.
- **CoT (chain-of-thought):** Guide an LLM through a problem-solving process by providing examples with intermediate reasoning steps.
- **Metaprompting:** Use prompting to guide the AI model to generate, modify, or interpret other prompts.

## Streamlining prompting workflows

- **Reusing prompts**: Saving prompts as templates for repeated use.
- Leveraging **prompt chaining**: Continuing conversations within the same chatbot to maintain context.
- Using **saved info** in Gemini: Storing specific information for the model to use consistently.
- **Gems** are personalized AI assistants within Gemini. They provide personalized responses tailored to specific instructions. They also streamline workflows like templates, prompts, and guided interactions.

Model guidance and refinement

Grounding: Connecting the AI's output to verifiable sources of information.

RAG: Retrieval-augmented generation

1. **Retrieval:** The LLM retrieves relevant information from external sources using tooling.
2. **Augmentation:** The retrieved information is incorporated into the prompt to the LLM.
3. **Generation:** The LLM processes the prompt and generates a response.
4. **Iteration (optional):** The LLM can iterate on the retrieval process as necessary.
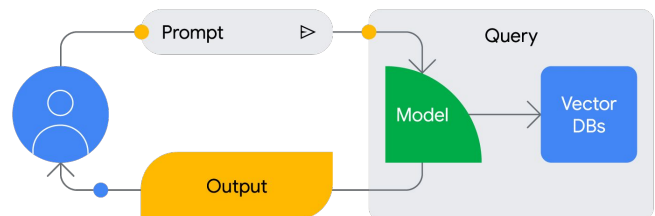
Sampling parameters

Settings that influence the AI model's behavior, allowing for customized results.

- **Token count:** This represents meaningful chunks of text (like words and punctuation).
- **Temperature:** This parameter controls the "creativity" or randomness of the model's word choices during text generation.
- **Top-p (nucleus sampling):** The cumulative probability of the most likely tokens considered during text generation. This is another way to control the randomness of the model's output.
- **Safety settings:** These settings allow you to filter out potentially harmful or inappropriate content from the model's output.
- **Output length:** This determines the maximum length of the generated text.

Techniques to improve generative AI model output

Foundation model limitations

Data dependency: Foundation model performance relies heavily on data. Biased or incomplete data will affect their outputs.

Knowledge cutoff: AI models are trained up to a specific knowledge cutoff date, meaning they might lack information about events after that point.

Bias: LLMs learn from large datasets which may contain biases. Even subtle biases can be magnified in the model's outputs.

Fairness: Assessing the fairness of generative AI models is a key aspect of responsible development.

Hallucinations: When AI models produce outputs that aren't accurate or based on real information.

Edge cases: Rare and atypical scenarios can expose a model's weaknesses, leading to unexpected results.
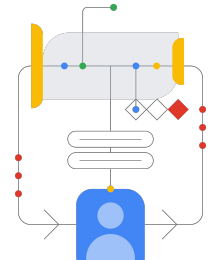
Managing your model

Google Cloud offers tools for managing the entire lifecycle of ML models. This includes the following:

- **Versioning**: Keep track of different versions of the model with Vertex AI Model Registry.
- **Performance tracking**: Review the model metrics to check the model's performance.
- **Drift monitoring**: Watch for changes in the model's accuracy over time with Vertex AI Model Monitoring.
- **Data management**: Use Vertex AI Feature Store to manage the data features the model uses.
- **Storage**: Use Vertex AI Model Garden to store and organize the models in one place.
- **Automate:** Use Vertex AI Pipelines to automate your machine learning tasks.

Humans in the loop (HITL): A process where human input and feedback are directly integrated into ML workflows.

- **Content moderation:** HITL ensures user-generated content is moderated contextually, catching harmful material algorithms might overlook.
- **Sensitive applications:** HITL provides critical oversight in fields like healthcare and finance, ensuring accuracy and reducing risks from automated systems.
- **High risk decision-making:** For high-stakes decisions, HITL can help safeguard accuracy and accountability through human review of ML model outputs.
- **Pre-generation review:** Human experts review and validate ML outputs before deployment, catching errors or biases before user impact.
- **Post-generation review:** Continuous human review and feedback after deployment help ML models improve and adapt to changing contexts and user needs.

Additional terms

- **Context window:** The amount of text the model can consider.
- **Fine-tuning:** A technique used to enhance a pre-trained or foundation models' performance for specific tasks or domains.

Resources to learn more

What is retrieval-augmented generation (RAG)?

Prompt engineering for AI guide

What is human-in-the-Loop (HITL) in AI & ML?

Business strategies for a successful gen AI solution

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?

Gen AI strategy: Combine a top-down approach (leadership setting the vision and strategy) with a bottom-up approach (employees identifying practical applications and providing feedback).

- **Strategic focus:** Prioritize focused gen AI implementations with clear business value.
- **Exploration:** Encourage experimentation and collaboration to discover valuable gen AI applications.
- **Responsible AI:** Establish ethical guidelines and ensure secure and responsible AI development.
- **Resourcing:** Invest in data strategy, leverage existing resources, and develop AI talent.
- **Impact:** Measure gen AI's impact on business goals and demonstrate tangible benefits.
- **Continuous improvement:** Continuously refine gen AI solutions based on feedback and data.

Responsible AI: Ensuring your AI applications don't cause harm and are used in an ethical manner.

Responsible AI needs to be considered throughout the entire AI lifecycle, from data preparation and model training to deployment and ongoing monitoring.

Secure AI: Protecting your AI applications from harm.

The **Secure AI Framework (SAIF)** helps organizations manage AI/ML model risks and ensure security.

Google Cloud's secure-by-design infrastructure helps support security across the AI/ML lifecycle. Various tools help protect data, models, and applications.

- Identity and Access Management (IAM) for controlling resource access.
- Security Command Center for security posture visibility.
- Workload monitoring tools to help build and maintain secure AI systems.

Factors when choosing a model for your use case

- **Modality:** Choose a generative AI model whose input and output data types (modality) align with your application's specific needs, whether it's text, images, audio, or video.
- **Context window:** You may need to balance an AI model's ability to generate coherent and relevant responses against the increased computational costs.
- **Performance:** A model's accuracy, speed, and efficiency are critical factors. Consider the trade-offs between performance and cost.
- **Availability and reliability:** Choose a model that is consistently available and performs reliably under load. Consider factors like uptime guarantees, redundancy, and disaster recovery mechanisms.

To **plan for your gen AI strategy**, establish a clear vision, prioritize use cases, invest in capabilities, manage change, measure value, and champion responsible AI.

Resources to learn more

[AI Principles](#)

[Introduction to Vertex Explainable AI](#)

[Google's Secure AI Framework (SAIF)](#)

Creating your own study guide with gen AI

Complete the following steps to practice your newly minted gen AI skills as you prepare for your exam.

Step 1: Open NotebookLM

1. Navigate to the NotebookLM website: notebooklm.google.com.
2. If you haven't used it before, you'll likely need to sign in with your Google account.

Step 2: Create a new notebook and upload a source

1. Once you're in NotebookLM, click **Create new**. Note: It might look like a plus sign or a button labeled "New."
2. You'll be prompted to add source documents. Select **Website**.
3. Copy and paste the link of this PDF study guide into NotebookLM, and select **Insert**.

Note: After you've pasted the link and selected upload, NotebookLM will process the document. This might take a few moments, but you'll see it appear in your list of sources once it's ready.

Step 3: Ask targeted questions to build your study guide sections

Start extracting key information by asking NotebookLM specific questions. Think about the common sections you'd want in a study guide:

- **Key concepts:** What are the fundamental ideas and principles?
- **Services/products:** What are the specific Google Cloud services covered? What do they do?
- **Use cases:** When and why would you use these services?
- **Best practices:** What are the recommended ways to approach certain tasks?
- **Important terms:** What are the key vocabulary and definitions?

Step 4: Refine and organize the information

Don't stop at the initial answers. Use follow-up questions to clarify concepts or delve deeper. As NotebookLM provides answers, you can do this:

1. **Copy and paste:** Select the relevant information from the chat window and paste it into a separate document (like a Google Doc, a plain text file, or even within NotebookLM itself if you want to create a summary there).
2. **Rephrase and summarize:** NotebookLM's answers are a great starting point. Rephrase the information in your own words to ensure better understanding and retention. Summarize longer explanations into concise points for quick review.

Step 5: Iterate and expand your study guide

Your first pass won't be perfect, and that's okay.

1. **Review:** Go through your initial study guide and identify areas that need more detail or clarification.
2. **Incorporate new information:** As you learn more or find other relevant resources, integrate that information into your study guide.
3. **Make a podcast:** Listen to NotebookLM generate an audio summary or explanation of key sections of this study guide.
4. **Use the Q&A:** Directly ask NotebookLM specific questions about concepts or services within your uploaded PDF study materials, and receive concise, contextually relevant answers to clarify your understanding.

Congratulations. You've now started building your own personalized study guide using NotebookLM.