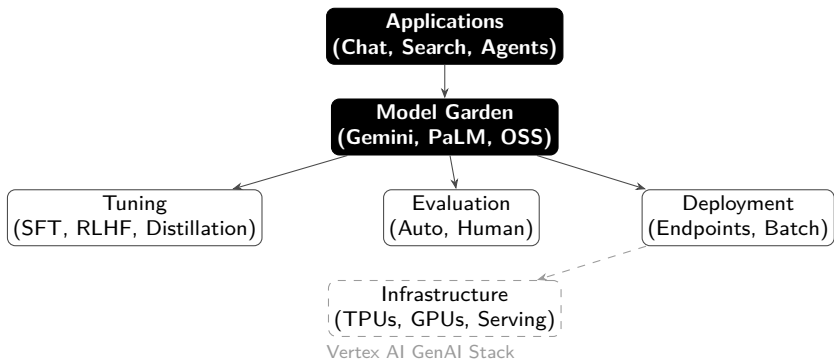




Vertex AI for Generative AI

GCP Generative AI | Model Garden, Gemini, Tuning, RAG & Responsible AI

Vertex AI GenAI Platform



Model Garden	Catalog of foundation models: Gemini (Google), open-source (Llama, Mistral, Gemma), and task-specific models. Deploy, tune, or use via API.
Gemini	Google's flagship multimodal model. Supports text, images, audio, video, code. Versions: Gemini Ultra, Pro, Flash (fastest/cheapest).
Gemini API	generateContent for single-turn, chat for multi-turn. Supports system instructions, function calling, structured output, grounding.
Embeddings API	Generate vector embeddings for text and multimodal content. Use for semantic search, clustering, classification, RAG.

Model Selection

Gemini for multimodal (text+image+video). PaLM 2 for text. Codey for code. Imagen for image generation. Choose based on modality needs.

Gemini Multimodal (text+image+video)	Imagen Image generation
Codey Code generation	Chirp Speech-to-text

- Temperature** — Controls randomness (0=deterministic, 1=creative)
- Top-K** — Sample from top K most likely tokens
- Top-P** — Sample from tokens summing to probability P
- Max tokens** — Maximum response length

Prompt Engineering

System instructions	Set model behavior, persona, output format. Persistent context across turns. Most influential control.
Few-shot prompting	Provide input-output examples in the prompt. 3-5 examples typically sufficient. Improves consistency.
Chain-of-thought	Ask model to reason step-by-step before answering. Improves accuracy on complex reasoning tasks.
Temperature	0.0 = deterministic (factual tasks). 1.0 = creative (brainstorming). 0.2-0.4 for most production use.
Top-K / Top-P	Top-K: sample from K most likely tokens. Top-P (nucleus): sample from tokens summing to P probability. Control diversity.
Output format	Request JSON, YAML, markdown, or structured output. Use response_mime_type for guaranteed JSON. Schema enforcement.

Technique	When to Use	Example
Zero-shot	Simple tasks, model already capable	Classify this text as positive/negative
Few-shot	Need specific format or behavior	3 examples of Q&A pairs then the question
Chain-of-thought	Math, logic, multi-step reasoning	Think step by step before answering
ReAct	Tool use, external data needed	Thought > Action > Observation loop

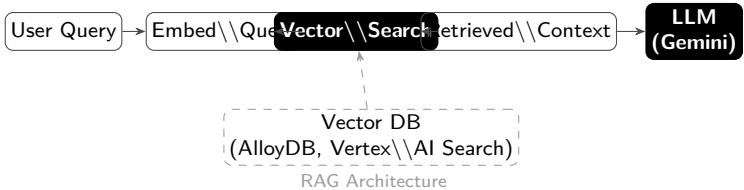
```
Vertex AI GenAI SDK
from vertexai.generative_models import GenerativeModel
model = GenerativeModel('gemini-1.5-pro')
response = model.generate_content('Explain ML0ps')
print(response.text)
```

Model Tuning & Customization

Supervised Fine-Tuning (SFT)	Train on labeled examples (input-output pairs). Adapts model to specific domain/task. Requires 100-10,000+ examples.
RLHF (Reinforcement Learning from Human Feedback)	Train reward model from human preferences. Then optimize generation policy. Best for alignment and safety.
Distillation	Transfer knowledge from larger model to smaller one. Student model learns from teacher outputs. Reduces inference cost.
LoRA / Adapter tuning	Lightweight fine-tuning that adds small trainable layers. Faster, cheaper than full fine-tuning. Good for domain adaptation.
Prompt tuning	Learn a soft prompt (virtual tokens) prepended to input. Keeps base model frozen. Cheapest customization method.

Method	Data Needed	Cost	Best For
Prompt engineering	0	None	Quick iteration, simple tasks
Prompt tuning	100+ examples	Low	Task-specific optimization
LoRA/Adapter	500+ examples	Medium	Domain adaptation
SFT	1,000+ examples	High	New behaviors, formats
RLHF	10,000+ comparisons	Very high	Alignment, safety, quality

RAG (Retrieval-Augmented Generation)



Vertex AI Search	Managed RAG service. Ingest documents (PDF, HTML, Cloud Storage, BigQuery). Auto-chunking, embedding, indexing. Search + summarize.
Vector Search (Matching Engine)	Managed approximate nearest neighbor (ANN) service. ScaNN algorithm. Low-latency similarity search on embeddings.
Grounding	Connect Gemini to external sources (Google Search, Vertex AI Search, custom data). Reduces hallucination. Provides citations.
Chunking strategies	Fixed-size (500-1000 tokens), semantic (by topic), recursive (split by structure). Overlap 10-20% between chunks.
AlloyDB + pgvector	PostgreSQL vector search. Store embeddings alongside relational data. Combines vector similarity with SQL filters.

- RAG reduces hallucination by grounding generation in retrieved facts
- Evaluate RAG quality: retrieval precision/recall, answer faithfulness, relevance
- Hybrid search: combine vector similarity with keyword matching (BM25) for best results
- Re-ranking: use cross-encoder to re-order retrieved chunks before passing to LLM



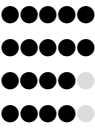
RAG Pipeline Stack

Responsible AI for Generative Models

Safety filters	Vertex AI provides configurable safety filters for: hate speech, dangerous content, sexually explicit, harassment. Adjustable thresholds.
Grounding	Reduces hallucination by anchoring responses to verified sources. Provide citations for generated content.
Model cards	Documentation of model capabilities, limitations, intended use, evaluation metrics, ethical considerations.
Evaluation	Vertex AI Evaluation: automated metrics (BLEU, ROUGE, faithfulness) and human evaluation (relevance, safety, quality).
PII handling	Use DLP API to detect/redact PII before sending to models. Filter PII from model outputs.

- Google AI Principles: socially beneficial, avoid unfair bias, built for safety, accountable, privacy-preserving
- Test for bias across demographics before production deployment
- Implement content moderation pipelines for user-facing generative AI applications
- Watermarking: SynthID for identifying AI-generated content (images, text)

Gemini Pro accuracy
Gemini Flash speed
RAG hallucination reduction
Safety filter coverage



1M+	100+	4
Gemini Context Window	Languages Supported	Safety Filter Categories

Official resources

⚠️ Check the current exam guide

Google Cloud is updating its exams to the products announced at Google Cloud Next '26 (Gemini Enterprise Agent Platform). New in 2025-2026: **Generative AI Leader** (foundational) and **Professional Agentic Architect** (beta closed; general-availability registration opens November 2, 2026, per its page checked 2026-10-05). Vertex AI documentation now redirects to Gemini Enterprise Agent Platform; these sheets keep the Vertex AI names.

STUDY GUIDE 1 link

Generative AI Leader study guide

Google Cloud · services.google.com/fh/files/misc/generative_ai_leader_study_guide_english.pdf

FREE

EXAM GUIDE 4 links

Generative AI Leader exam guide

Google Cloud · services.google.com/fh/files/misc/generative_ai_leader_exam_guide_english.pdf

FREE

Generative AI Leader certification page

Google Cloud · cloud.google.com/learn/certification/generative-ai-leader

FREE

Professional Agentic Architect exam guide

Google Cloud · services.google.com/fh/files/misc/profession...ic_architect_exam_guide_english.pdf
beta closed, GA registration from 2026-11-02

FREE

Professional Agentic Architect certification page

Google Cloud · cloud.google.com/learn/certification/agentic-architect

FREE

VIDEO 1 link

Generative AI Leader learning path

Google Skills · skills.google/paths/1951

FREE

DOCS 1 link

Models on Gemini Enterprise Agent Platform (ex Vertex AI)

Google Cloud · docs.cloud.google.com/gemini-enterprise-agent-platform/models
Gemini models, grounding, agents

FREE

CHEATSHEET 1 link

gcloud CLI cheat sheet

Google Cloud · cloud.google.com/sdk/docs/cheatsheet

FREE

8 official links · 8 free, 0 paid · checked 2026-10-05