

Model families serve different workloads

Generative AI applications may use several model families.

Families

- Language models
- Multimodal models
- Embedding models
- Image-generation models
- Reasoning models

Key message

The application's required input, output, and task determine the appropriate model.

Language and multimodal models

Model family	Typical input	Typical output	Common uses
Language model	Text or code	Text or code	Question answering, summarization, classification, extraction, code generation
Multimodal model	Text plus images, audio, or video	Text or other supported content	Image understanding, document analysis, spoken interaction, mixed-media assistants

Language model

“Summarize a typed support ticket.”

Multimodal model

“Inspect a customer’s photo and explain the visible equipment problem.”

Exam focus

When a scenario requires understanding both an image and accompanying instructions, choose a model that supports those modalities.

Embedding models represent meaning as numbers

An embedding model converts content into a numerical vector that represents semantic meaning.

Example • “A search for “reset my login credentials” can match documentation titled “Changing a forgotten password.””

What embeddings enable

- Applications compare vectors to find content with similar meaning.
- Keyword search looks for matching terms.
- Vector search finds conceptually related content.

Common uses

- Semantic search
- Document retrieval
- Recommendation and clustering
- Retrieval-augmented generation

Exam trap

Choose an embedding model when the requirement centers on semantic similarity, vector search, or retrieval.

Image-generation models

An image-generation model creates or modifies images from instructions and, when supported, reference images.

Example prompt · “Create a product-support illustration showing a technician inspecting a network rack.”

Typical tasks

- Text-to-image generation
- Image editing
- Style or composition changes
- Creating visual variations

Key distinction

- Image understanding: Interprets an existing image
- Image generation: Produces new visual content

Remember

Some multimodal models understand images without generating them.

Reasoning models

A reasoning model allocates more computation to working through complex problems before producing an answer.

Example · “Compare three support-resolution plans, apply five policy constraints, and justify the safest valid choice.”

Strong use cases

- Multi-step analysis
- Mathematics and logic
- Complex coding tasks
- Planning under multiple constraints

Tradeoffs

- Higher latency
- Potentially higher cost
- Unnecessary for many simple requests

Exam focus

Choose a reasoning model when task complexity justifies additional analysis, not merely because the prompt contains a question.

Match capabilities to requirements

Application requirement	Best capability match
Summarize a customer email	Language model
Interpret a photo and written question	Multimodal model
Retrieve semantically similar support articles	Embedding model
Create a marketing illustration	Image-generation model
Solve a complex planning problem with constraints	Reasoning model
Retrieve documentation and compose an answer	Embedding model plus language or reasoning model

Selection questions

1. What content will the model receive?
2. What output must it produce?
3. How complex is the task?
4. What accuracy, latency, and cost constraints apply?
5. Does the application require current or private data?

Start with the task

A single application can route different tasks to different models.