GOOGLE

Gemini Models

Which Model to Use and When

A Practical Decision Guide · 2025

The Gemini Model Family

Four tiers each built for a different scale of task

Most Capable

Gemini Ultra

Frontier reasoning & research

- Complex multi-step reasoning
- Scientific & advanced coding
- High-fidelity multimodal tasks
- Long-horizon research queries

Balanced

Gemini Pro

Everyday enterprise workloads

- Summarization & analysis
- Code generation & review
- Document Q&A
- Customer-facing AI features

Fast & Efficient

Gemini Flash

High-volume, low-latency tasks

- Real-time chat & agents
- API calls at scale
- Classification & extraction
- Mobile & edge use cases

On-Device

Gemini Nano

Privacy-first on-device AI

- Offline inference
- Smart reply & autocomplete
- Low-resource environments
- Data privacy critical apps

Which Model Should You Use?

Match your task to the right tier optimize for capability, speed, or cost

USE CASE / SCENARIO	MODEL	WHY IT FITS
Research / frontier science	Ultra	Requires deep multi-step reasoning across specialized domains
Advanced code generation & debugging	Ultra	Complex logic, security review, and architecture-level decisions
Enterprise Q&A & document analysis	Pro	Balanced capability with cost-efficient token usage
Content creation & summarization	Pro	Strong language quality without frontier-level compute
Real-time chat & conversational agents	Flash	Low latency responses under 1 second at scale
High-volume classification & tagging	Flash	Optimized throughput, fraction of the cost of Pro
On-device / offline & privacy-critical	Nano	Runs fully locally no network, no data sent to cloud

💡 Pro tip: Start with Flash for prototyping scale up to Pro or Ultra only if quality falls short. Use Nano when privacy or offline access is required.