ALIBABA AI · QWEN

Alibaba AI Models

Language · Vision · Coding · Audio The Qwen Model Family

Qwen2.5-Max | Qwen2.5-72B | Qwen2.5-Coder | Qwen2.5-VL | Qwen-Audio | Qwen2.5-Math

A Practical Decision Guide · 2025 | Open Source & Commercial API via Alibaba Cloud

The Alibaba Qwen Model Family

Six specialized models across language, vision, coding, math, and audio open-source and enterprise-ready

LANGUAGE, REASONING & MATH MODELS

Flagship Model

Qwen2.5-Max

Alibaba's most capable frontier model

- Beats GPT-4o on major benchmarks
- 1M token context window
- Strong multilingual (29+ langs)
- Available via Alibaba Cloud API

Best Open Weight

Qwen2.5-72B

Top open-source 72B model Apache 2.0

- Outperforms Llama 3 70B & Mixtral
- 128K context window
- Instruction + chat tuned
- Self-hostable, free for commercial use

Math & Science

Qwen2.5-Math

Specialized model for math & STEM reasoning

- Top score on MATH & GSM8K benchmarks
- Step-by-step equation solving
- Science & physics reasoning
- 7B / 72B sizes available

SPECIALIZED MODALITY MODELS

Vision+Language

Qwen2.5-VL

Multimodal vision-language understanding

- Image, chart & doc Q&A
- Video understanding support
- OCR & document parsing

Best for Code

Qwen2.5-Coder

Purpose-built coding model 32B params

- 92 programming languages
- Repo-level code understanding
- Beats GPT-4o on code benchmarks

Audio+Language

Qwen-Audio

Audio understanding & speech tasks

- Speech recognition (ASR)
- Audio question answering
- Sound event understanding

Which Alibaba Qwen Model Should You Use?

Match your task across language, math, code, vision, and audio open-source & Alibaba Cloud API

USE CASE / SCENARIO	MODEL	WHY IT FITS
General language tasks, Q&A & reasoning	Qwen2.5-Max	Flagship frontier model beats GPT-4o with 1M token context
Self-hosted open-weight deployments	Qwen2.5-72B	Best open-source 72B model Apache 2.0, outperforms Llama 3 70B
Math, science & STEM problem solving	Qwen2.5-Math	Purpose-built math model top scores on MATH & GSM8K benchmarks
Code generation, review & completion	Coder 32B	Specialized 32B coder 92 languages, beats GPT-4o on code evals
Image, chart & document understanding	Qwen2.5-VL	Multimodal vision model OCR, video & chart Q&A in one model
Speech recognition & audio understanding	Qwen-Audio	End-to-end audio model ASR, sound Q&A & multilingual speech
Cost-efficient API at scale	Qwen2.5-72B	Open-weight + self-host means zero API cost optimal at volume

💡 Pro tip: Use Qwen2.5-Max for frontier quality via API, Qwen2.5-72B for free self-hosting, Coder 32B for code, Qwen2.5-Math for STEM, and Qwen2.5-VL for all vision tasks.