Contents Menu Expand Light mode Dark mode Auto light/dark, in light mode Auto light/dark, in dark mode

Hide navigation sidebar

Hide table of contents sidebar

Skip to content

Toggle site navigation sidebar

—

Qwen

Toggle Light / Dark / Auto color theme

Toggle table of contents sidebar

—

Qwen

Getting Started

* [Quickstart](getting_started/quickstart.html)
* [Key Concepts](getting_started/concepts.html)

Inference

* [Hugging Face transformers](inference/chat.html)

Run Locally

* [Ollama](run_locally/ollama.html)
* [MLX-LM](run_locally/mlx-lm.html)
* [llama.cpp](run_locally/llama.cpp.html)

Web UI

* [Text Generation Web UI](web_ui/text_generation_webui.html)

Quantization

* [AWQ](quantization/awq.html)
* [GPTQ](quantization/gptq.html)
* [llama.cpp](quantization/llama.cpp.html)

Deployment

* [vLLM](deployment/vllm.html)
* [TGI](deployment/tgi.html)
* [SkyPilot](deployment/skypilot.html)
* [OpenLLM](deployment/openllm.html)

Training

* [SFT](training/SFT/index.html)

Toggle navigation of SFT

—

* [LLaMA-Factory](training/SFT/llama_factory.html)

Framework

* [Function Calling](framework/function_call.html)
* [Qwen-Agent](framework/qwen_agent.html)
* [LlamaIndex](framework/LlamaIndex.html)
* [Langchain](framework/Langchain.html)

Benchmark

* [Performance of Quantized Models](benchmark/quantization_benchmark.html)
* [Qwen2.5 Speed Benchmark](benchmark/speed_benchmark.html)

Back to top

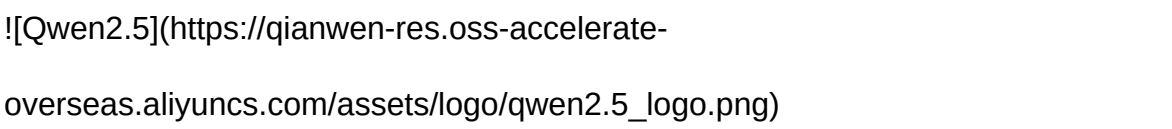
[View this page](_sources/index.rst.txt "View this page")

Toggle Light / Dark / Auto color theme

Toggle table of contents sidebar

—

Welcome to Qwen!¶



Qwen is the large language model and large multimodal model series of the Qwen Team, Alibaba Group. Now the large language models have been upgraded to Qwen2.5. Both language models and multimodal models are pretrained on large-scale multilingual and multimodal data and post-trained on quality data for aligning to human preferences. Qwen is capable of natural language understanding, text generation, vision understanding, audio understanding, tool use, role play, playing as AI agent, etc.

The latest version, Qwen2.5, has the following features:

* Dense, easy-to-use, decoder-only language models, available in **0.5B** , **1.5B** , **3B** , **7B** , **14B** , **32B** , and **72B** sizes, and base and instruct variants.

* Pretrained on our latest large-scale dataset, encompassing up to **18T** tokens.

* Significant improvements in instruction following, generating long texts (over 8K tokens),

understanding structured data (e.g, tables), and generating structured outputs especially JSON.

* More resilient to the diversity of system prompts, enhancing role-play implementation and condition-setting for chatbots.

* Context length support up to **128K** tokens and can generate up to **8K** tokens.

* Multilingual support for over **29** languages, including Chinese, English, French, Spanish, Portuguese, German, Italian, Russian, Japanese, Korean, Vietnamese, Thai, Arabic, and more.

For more information, please visit our:

* [Blog](https://qwenlm.github.io/)

* [GitHub](https://github.com/QwenLM)

* [Hugging Face](https://huggingface.co/Qwen)

* [ModelScope](https://modelscope.cn/organization/qwen)

* [Qwen2.5 Collection](https://huggingface.co/collections/Qwen/qwen25-66e81a666513e518adb90d9e)

Join our community by joining our [Discord](https://discord.gg/yPEP2vHTu4) and [WeChat](https://github.com/QwenLM/Qwen/blob/main/assets/wechat.png) group. We are looking forward to seeing you there!

[Next Quickstart](getting_started/quickstart.html)

Made with [Sphinx](https://www.sphinx-doc.org/) and

[@pradyunsg](https://pradyunsg.me)'s [Furo](https://github.com/pradyunsg/furo)