Title: Qwen

URL: https://en.wikipedia.org/wiki/Qwen

PageID: 78475283

Categories: Category:2023 in artificial intelligence, Category:2023 software, Category:Alibaba Group, Category:Generative pre-trained transformers, Category:Large language models, Category:Open-source artificial intelligence, Category:Software using the Apache license

-----

Qwen (also called Tongyi Qianwen , Chinese : ■■■■) is a family of large language models developed by Chinese company Alibaba Cloud . In July 2024, it was ranked as the top Chinese language model in some benchmarks and third globally behind the top models of Anthropic and OpenAI . [1]

Models

Alibaba first launched a beta of Qwen in April 2023 under the name Tongyi Qianwen. [2] The model's architecture was based on the Llama architecture developed by Meta AI . [3] [4] It was publicly released in September 2023 after receiving approval from the Chinese government. [5] In December 2023 it released its 72B and 1.8B models for download, while Qwen 7B weights were released in August. [6] [7] Their models are sometimes described as open source , but the training code has not been released nor has the training data been documented, and they do not meet the terms of either the Open Source AI Definition or the Model Openness Framework from the Linux Foundation .

In June 2024 Alibaba launched Qwen2 and in September it released some of its models with open weights, while keeping its most advanced models proprietary. [8] [9] Qwen2 contains both dense and sparse models. [10]

In November 2024, QwQ-32B-Preview, a model focusing on reasoning similar to OpenAI's o1 , was released under the Apache 2.0 License , although only the weights were released, not the dataset or training method. [11] [12] QwQ has a 32K token context length and performs better than o1 on some benchmarks. [13]

The Qwen-VL series is a line of visual language models that combines a vision transformer with a LLM. [3] [14] Alibaba released Qwen2-VL with variants of 2 billion and 7 billion parameters. [15] [16] [17]

In January 2025, Qwen2.5-VL was released with variants of 3, 7, 32, and 72 billion parameters. [18] All models except the 72B variant are licensed under the Apache 2.0 license. [19] Qwen-VL-Max is Alibaba's flagship vision model as of 2024, and is sold by Alibaba Cloud at a cost of US$0.00041 per thousand input tokens. [20]

Alibaba has released several other model types such as Qwen-Audio and Qwen2-Math. [21] In total, it has released more than 100 open weight models, with its models having been downloaded more than 40 million times. [9] [22] Fine-tuned versions of Qwen have been developed by enthusiasts, such as "Liberated Qwen", developed by San Francisco-based Abacus AI, which is a version that responds to any user request without content restrictions. [23]

On January 29, 2025, Alibaba launched Qwen2.5-Max. According to a blog post from Alibaba, Qwen2.5-Max outperforms other foundation models such as GPT-4o , DeepSeek-V3 , and Llama -3.1-405B in key benchmarks. [24] [25] In February 2025, Alibaba announced on their official X account that the 2.5-Max model would be opened up, however it has not been released. [26]

On March 24, 2025, Alibaba launched Qwen2.5-VL-32B-Instruct as a successor to the Qwen2.5-VL model. It was released under the Apache 2.0 license. [27] [28]

On March 26, 2025, Qwen2.5-Omni-7B was released under the Apache 2.0 license and made available through chat.qwen.ai, as well as platforms like Hugging Face , GitHub , and ModelScope.

[29] The Qwen2.5-Omni model accepts text, images, videos, and audio as input and can generate both text and audio as output, allowing it to be used for real-time voice chatting, similar to OpenAI's GPT-4o. [29]

On April 28, 2025, the Qwen3 model family was released, [30] with all models licensed under the Apache 2.0 license. The Qwen3 model family includes both dense (0.6B, 1.7B, 4B, 8B, 14B, and 32B parameters) and sparse models (30B with 3B activated parameters, 235B with 22B activated parameters). They were trained on 36 trillion tokens in 119 languages and dialects. [31] All models except the 0.6B, 1.7B, and 4B variants have a 128K token context window . Like OpenAI's o1 and QwQ 32B, the Qwen3 models support reasoning , which can be enabled or disabled through the tokenizer. The Qwen3 models are available through chat.qwen.ai and can be downloaded via Hugging Face and ModelScope. [32]

On September 5, 2025, Alibaba launched Qwen3-Max. [33] According to Alibaba's official X account, it outperforms other foundation non-reasoning models such as Qwen3-235B-A22B-Instruct-2507, Kimi K2, Claude 4 Opus Non-thinking , and DeepSeek V3.1 . [34] There is no dedicated thinking mode for Qwen3-Max as of yet. [35]

On September 10, 2025, Qwen3-Next was released under the Apache 2.0 license and made available through chat.qwen.ai, as well as platforms like Hugging Face and Model Scope. Qwen3-Next includes two post-trained Instruct and Thinking models. Qwen3-Next was created with a new model-architecture called Qwen3-Next, in the belief that Context Length Scaling and Total Parameter Scaling are two major trends in the future of large models. Qwen3-Next introduces several key improvements over the Qwen3 architecture: a hybrid attention mechanism, a highly sparse Mixture-of-Experts (MoE) structure, training-stability-friendly optimizations, and a multi-token prediction mechanism for faster inference. Based on the Qwen3-Next architecture, a model with 80B total parameters and 3B active parameters was created. The Qwen3-Next model performs comparable to, or in some cases better than, Qwen3-32b while using less than 10% of its training cost (in GPU hours). In inference, especially with contexts greater than 32K tokens, it reaches greater than 10x higher throughput. Qwen3.5 will use a refined version of the Qwen3-Next architecture. [36]

See also

List of large language models

References

External links

Free and open-source software portal

China portal

Official website

Qwen on GitHub

Qwen on Hugging Face

v

t

e

Alibaba Cloud

AliExpress

AliGenie

AliMusic

AliOS

Alipay

Qwen

Taobao

Tmall Tmall Genie

Tmall Genie

Xuexi Qiangguo

Alibaba Health

Alibaba Pictures

Amblin Partners Amblin Entertainment Amblin Television DreamWorks Pictures

Amblin Entertainment

Amblin Television

DreamWorks Pictures

Ant Group Tianhong Asset Management Zhima Credit

Tianhong Asset Management

Zhima Credit

AutoNavi

Cainiao

Ele.me

Heyi Pictures

Lazada Group

Shenma

South China Morning Post

Tudou

UCWeb UC Browser

UC Browser

World Electronic Sports Games

Youku

Zhejiang Hupan Entrepreneurship Research Center

Jack Ma

Daniel Zhang

J. Michael Evans

Peng Lei

Jonathan Lu

Joseph Tsai

Wang Jian

Maggie Wu

Commons

Category

v

t

e

LMArena

List of chatbots

List of LLMs

character.ai

ChatGPT

Claude

Command

Copilot

DeepSeek

Ernie

Gemini

GLM

Grok

Hunyuan

Kimi

Llama

Mistral

Perplexity

Poe

Qwen

You.com

Category

v

t

e

Autoencoder

Deep learning

Fine-tuning

Foundation model

Generative adversarial network

Generative pre-trained transformer

Large language model

Model Context Protocol

Neural network

Prompt engineering

Reinforcement learning from human feedback

Retrieval-augmented generation

Self-supervised learning

Stochastic parrot

Synthetic data

Top-p sampling

Transformer

Variational autoencoder

Vibe coding

Vision transformer

Waluigi effect

Word embedding

Character.ai

ChatGPT

DeepSeek

Ernie

Gemini

Grok

Copilot

Claude

Gemini

Gemma

GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5

1

2

3

J

4

4o

4.5

4.1

OSS

5

Llama

o1

o3

o4-mini

Qwen

Base44

Claude Code

Cursor

Devstral

GitHub Copilot

Kimi-Dev

Qwen3-Coder

Replit

Xcode

Aurora

Firefly

Flux

GPT Image 1

Ideogram

Imagen

Midjourney

Qwen-Image

Recraft

Seedream

Stable Diffusion

Dream Machine

Hailuo AI

Kling

Midjourney Video

Runway Gen

Seedance

Sora

Veo

Wan

15.ai

Eleven

MiniMax Speech 2.5

WaveNet

Eleven Music

Endel

Lyria

Riffusion

Suno AI

Udio

Agentforce

AutoGLM

AutoGPT

ChatGPT Agent

Devin AI

Manus

OpenAI Codex

Operator

Replit Agent

01.AI

Aleph Alpha

Anthropic

Baichuan

Canva

Cognition AI

Cohere

Contextual AI

DeepSeek

ElevenLabs

Google DeepMind

HeyGen

Hugging Face

Inflection AI

Krikey AI

Kuaishou

Luma Labs

Meta AI

MiniMax

Mistral AI

Moonshot AI

OpenAI

Perplexity AI

Runway

Safe Superintelligence

Salesforce

Scale AI

SoundHound

Stability AI

Synthesia

Thinking Machines Lab

Upstage

xAI

Z.ai

Category

v

t

e

History timeline

timeline

Companies

Projects

Parameter Hyperparameter

Hyperparameter

Loss functions

Regression Bias–variance tradeoff Double descent Overfitting

Bias–variance tradeoff

Double descent

Overfitting

Clustering

Gradient descent SGD Quasi-Newton method Conjugate gradient method

SGD

Quasi-Newton method

Conjugate gradient method

Backpropagation

Attention

Convolution

Normalization Batchnorm

Batchnorm

Activation Softmax Sigmoid Rectifier

Softmax

Sigmoid

Rectifier

Gating

Weight initialization

Regularization

Datasets Augmentation

Augmentation

Prompt engineering

Reinforcement learning Q-learning SARSA Imitation Policy gradient

Q-learning

SARSA

Imitation

Policy gradient

Diffusion

Latent diffusion model

Autoregression

Adversary

RAG

Uncanny valley

RLHF

Self-supervised learning

Reflection

Recursive self-improvement

Hallucination

Word embedding

Vibe coding

Machine learning In-context learning

In-context learning

Artificial neural network Deep learning

Deep learning

Language model Large language model NMT

Large language model

NMT

Reasoning language model

Model Context Protocol

Intelligent agent

Artificial human companion

Humanity's Last Exam

Artificial general intelligence (AGI)

AlexNet

WaveNet

Human image synthesis

HWR

OCR

Computer vision

Speech synthesis 15.ai ElevenLabs

15.ai

ElevenLabs

Speech recognition Whisper

Whisper

Facial recognition

AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)

Gemma

Grok

LaMDA

BLOOM

DBRX

Project Debater

IBM Watson

IBM Watsonx

Granite

PanGu-Σ

DeepSeek

Qwen

AlphaGo

AlphaZero

OpenAI Five

Self-driving car

MuZero

Action selection AutoGPT

AutoGPT

Robot control

Alan Turing

Warren Sturgis McCulloch

Walter Pitts

John von Neumann

Claude Shannon

Shun'ichi Amari

Kunihiko Fukushima

Takeo Kanade

Marvin Minsky

John McCarthy

Nathaniel Rochester

Allen Newell

Cliff Shaw

Herbert A. Simon

Oliver Selfridge

Frank Rosenblatt

Bernard Widrow

Joseph Weizenbaum

Seymour Papert

Seppo Linnainmaa

Paul Werbos

Geoffrey Hinton

John Hopfield

Jürgen Schmidhuber

Yann LeCun

Yoshua Bengio

Lotfi A. Zadeh

Stephen Grossberg

Alex Graves

James Goodnight

Andrew Ng

Fei-Fei Li

Alex Krizhevsky

Ilya Sutskever

Oriol Vinyals

Quoc V. Le

Ian Goodfellow

Demis Hassabis

David Silver

Andrej Karpathy

Ashish Vaswani

Noam Shazeer

Aidan Gomez

John Schulman

Mustafa Suleyman

Jan Leike

Daniel Kokotajlo

François Chollet

Neural Turing machine

Differentiable neural computer

Transformer Vision transformer (ViT)

Vision transformer (ViT)

Recurrent neural network (RNN)

Long short-term memory (LSTM)

Gated recurrent unit (GRU)

Echo state network

Multilayer perceptron (MLP)

Convolutional neural network (CNN)

Residual neural network (RNN)

Highway network

Mamba

Autoencoder

Variational autoencoder (VAE)

Generative adversarial network (GAN)

Graph neural network (GNN)

Category