Qwen3 Technical Report

Qwen Team

https://huggingface.co/Qwen
https://modelscope.cn/organization/qwen
https://github.com/QwenLM/Qwen3

Abstract

In this work, we present Qwen3, the latest version of the Qwen model family. Qwen3 comprises a series of large language models (LLMs) designed to advance performance, efficiency, and multilingual capabilities. The Qwen3 series includes models of both dense and Mixture-of-Expert (MoE) architectures, with parameter scales ranging from 0.6 to 235 billion. A key innovation in Qwen3 is the integration of thinking mode (for complex, multi-step reasoning) and non-thinking mode (for rapid, context-driven responses) into a unified framework. This eliminates the need to switch between different models——such as chat-optimized models (e.g., GPT-4o) and dedicated reasoning models (e.g., QwQ-32B)——and enables dynamic mode switching based on user queries or chat templates. Meanwhile, Qwen3 introduces a thinking budget mechanism, allowing users to allocate computational resources adaptively during inference, thereby balancing latency and performance based on task complexity. Moreover, by leveraging the knowledge from the flagship models, we significantly reduce the computational resources required to build smaller-scale models, while ensuring their highly competitive performance. Empirical evaluations demonstrate that Qwen3 achieves state-of-the-art results across diverse benchmarks, including tasks in code generation, mathematical reasoning, agent tasks, etc., competitive against larger MoE models and proprietary models. Compared to its predecessor Qwen2.5, Qwen3 expands multilingual support from 29 to 119 languages and dialects, enhancing global accessibility through improved cross-lingual understanding and generation capabilities. To facilitate reproducibility and community-driven research and development, all Qwen3 models are publicly accessible under Apache 2.0.

1 Introduction

The pursuit of artificial general intelligence (AGI) or artificial super intelligence (ASI) has long been a goal for humanity. Recent advancements in large foundation models, e.g., GPT-4o (OpenAI, 2024), Claude 3.7 (Anthropic, 2025), Gemini 2.5 (DeepMind, 2025), DeepSeek-V3 (Liu et al., 2024a), Llama-4 (Meta-AI, 2025), and Qwen2.5 (Yang et al., 2024b), have demonstrated significant progress toward this objective. These models are trained on vast datasets spanning trillions of tokens across diverse domains and tasks, effectively distilling human knowledge and capabilities into their parameters. Furthermore, recent developments in reasoning models, optimized through reinforcement learning, highlight the potential for foundation models to enhance inference-time scaling and achieve higher levels of intelligence, e.g., o3 (OpenAI, 2025), DeepSeek-R1 (Guo et al., 2025). While most state-of-the-art models remain proprietary, the rapid growth of open-source communities has substantially reduced the performance gap between open-weight and closed-source models. Notably, an increasing number of top-tier models (Meta-AI, 2025; Liu et al., 2024a; Guo et al., 2025; Yang et al., 2024b) are now being released as open-source, fostering broader research and innovation in artificial intelligence.

In this work, we introduce Qwen3, the latest series in our foundation model family, Qwen. Qwen3 is a collection of open-weight large language models (LLMs) that achieve state-of-the-art performance across a wide variety of tasks and domains. We release both dense and Mixture-of-Experts (MoE) models, with the number of parameters ranging from 0.6 billion to 235 billion, to meet the needs of different downstream applications. Notably, the flagship model, Qwen3-235B-A22B, is an MoE model with a total of 235 billion parameters and 22 billion activated ones per token. This design ensures both high performance and efficient inference.

Qwen3 introduces several key advancements to enhance its functionality and usability. First, it integrates two distinct operating modes, thinking mode and non-thinking mode, into a single model. This allows users to switch between these modes without alternating between different models, e.g., switching from Qwen2.5 to QwQ (Qwen Team, 2024). This flexibility ensures that developers and users can adapt the model's behavior to suit specific tasks efficiently. Additionally, Qwen3 incorporates thinking budgets, providing users with fine-grained control over the level of reasoning effort applied by the model during task execution. This capability is crucial to the optimization of computational resources and performance, tailoring the model's thinking behavior to meet varying complexity in real-world applications. Furthermore, Qwen3 has been pre-trained on 36 trillion tokens covering up to 119 languages and dialects, effectively enhancing its multilingual capabilities. This broadened language support amplifies its potential for deployment in global use cases and international applications. These advancements together establish Qwen3 as a cutting-edge open-source large language model family, capable of effectively addressing complex tasks across various domains and languages.

The pre-training process for Qwen3 utilizes a large-scale dataset consisting of approximately 36 trillion tokens, curated to ensure linguistic and domain diversity. To efficiently expand the training data, we employ a multi-modal approach: Qwen2.5-VL (Bai et al., 2025) is finetuned to extract text from extensive PDF documents. We also generate synthetic data using domain-specific models: Qwen2.5-Math (Yang et al., 2024c) for mathematical content and Qwen2.5-Coder (Hui et al., 2024) for code-related data. The pre-training process follows a three-stage strategy. In the first stage, the model is trained on about 30 trillion tokens to build a strong foundation of general knowledge. In the second stage, it is further trained on knowledge-intensive data to enhance reasoning abilities in areas like science, technology, engineering, and mathematics (STEM) and coding. Finally, in the third stage, the model is trained on long-context data to increase its maximum context length from 4,096 to 32,768 tokens.

To better align foundation models with human preferences and downstream applications, we employ a multi-stage post-training approach that empowers both thinking (reasoning) and non-thinking modes. In the first two stages, we focus on developing strong reasoning abilities through long chain-of-thought (CoT) cold-start finetuning and reinforcement learning focusing on mathematics and coding tasks. In the final two stages, we combine data with and without reasoning paths into a unified dataset for further fine-tuning, enabling the model to handle both types of input effectively, and we then apply general-domain reinforcement learning to improve performance across a wide range of downstream tasks. For smaller models, we use strong-to-weak distillation, leveraging both off-policy and on-policy knowledge transfer from larger models to enhance their capabilities. Distillation from advanced teacher models significantly outperforms reinforcement learning in performance and training efficiency.

We evaluate both pre-trained and post-trained versions of our models across a comprehensive set of benchmarks spanning multiple tasks and domains. Experimental results show that our base pre-trained models achieve state-of-the-art performance. The post-trained models, whether in thinking or non-thinking mode, perform competitively against leading proprietary models and large mixture-of-experts (MoE) models such as o1, o3-mini, and DeepSeek-V3. Notably, our models excel in coding, mathematics, and agent-related tasks. For example, the flagship model Qwen3-235B-A22B achieves 85.7 on AIME'24

and 81.5 on AIME'25 (AIME, 2025), 70.7 on LiveCodeBench v5 (Jain et al., 2024), 2,056 on CodeForces, and 70.8 on BFCL v3 (Yan et al., 2024). In addition, other models in the Qwen3 series also show strong performance relative to their size. Furthermore, we observe that increasing the thinking budget for thinking tokens leads to a consistent improvement in the model's performance across various tasks.

In the following sections, we describe the design of the model architecture, provide details on its training procedures, present the experimental results of pre-trained and post-trained models, and finally, conclude this technical report by summarizing the key findings and outlining potential directions for future research.

2 Architecture

The Qwen3 series includes 6 dense models, namely Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Qwen3-14B, and Qwen3-32B, and 2 MoE models, Qwen3-30B-A3B and Qwen3-235B-A22B. The flagship model, Qwen3-235B-A22B, has a total of 235B parameters with 22B activated ones. Below, we elaborate on the architecture of the Qwen3 models.

The architecture of the Qwen3 dense models is similar to Qwen2.5 (Yang et al., 2024b), including using Grouped Query Attention (GQA, Ainslie et al., 2023), SwiGLU (Dauphin et al., 2017), Rotary Positional Embeddings (RoPE, Su et al., 2024), and RMSNorm (Jiang et al., 2023) with pre-normalization. Besides, we remove QKV-bias used in Qwen2 (Yang et al., 2024a) and introduce QK-Norm (Dehghani et al., 2023) to the attention mechanism to ensure stable training for Qwen3. Key information on model architecture is provided in Table 1.

The Qwen3 MoE models share the same fundamental architecture as the Qwen3 dense models. Key information on model architecture is provided in Table 2. We follow Qwen2.5-MoE (Yang et al., 2024b) and implement fine-grained expert segmentation (Dai et al., 2024). The Qwen3 MoE models have 128 total experts with 8 activated experts per token. Unlike Qwen2.5-MoE, the Qwen3-MoE design excludes shared experts. Furthermore, we adopt the global-batch load balancing loss (Qiu et al., 2025) to encourage expert specialization. These architectural and training innovations have yielded substantial improvements in model performance across downstream tasks.

Qwen3 models utilize Qwen's tokenizer (Bai et al., 2023), which implements byte-level byte-pair encoding (BBPE, Brown et al., 2020; Wang et al., 2020; Sennrich et al., 2016) with a vocabulary size of 151,669.

Table 1: Model architecture of Qwen3 dense models.

Models	Layers	Heads (Q / KV)	Tie Embedding	Context Length
Qwen3-0.6B	28	16 / 8	Yes	32K
Qwen3-1.7B	28	16 / 8	Yes	32K
Qwen3-4B	36	32 / 8	Yes	128K
Qwen3-8B	36	32 / 8	No	128K
Qwen3-14B	40	40 / 8	No	128K
Qwen3-32B	64	64 / 8	No	128K

Table 2: Model architecture of Qwen3 MoE models.

Models	Layers	Heads (Q / KV)	# Experts (Total / Activated)	Context Length
Qwen3-30B-A3B	48	32 / 4	128 / 8	128K
Qwen3-235B-A22B	94	64 / 4	128 / 8	128K

3 Pre-training

In this section, we describe the construction of our pretraining data, the details of our pretraining approach, and present experimental results from evaluating the base models on standard benchmarks.

3.1 Pre-training Data

Compared with Qwen2.5 (Yang et al., 2024b), we have significantly expanded the scale and diversity of our training data. Specifically, we collected twice as many pre-training tokens—covering three times more languages. All Qwen3 models are trained on a large and diverse dataset consisting of **119 languages and dialects**, with a total of **36 trillion tokens**. This dataset includes high-quality content in various

domains such as coding, STEM (Science, Technology, Engineering, and Mathematics), reasoning tasks, books, multilingual texts, and synthetic data.

To further expand the pre-training data corpus, we first employ the Qwen2.5-VL model (Bai et al., 2025) to perform text recognition on a large volume of PDF-like documents. The recognized text is then refined using the Qwen2.5 model (Yang et al., 2024b), which helps improve its quality. Through this two-step process, we are able to obtain an additional set of high-quality text tokens, amounting to trillions in total. Besides, we employ Qwen2.5 (Yang et al., 2024b), Qwen2.5-Math (Yang et al., 2024c), and Qwen2.5-Coder (Hui et al., 2024) models to synthesize trillions of text tokens in different formats, including textbooks, question-answering, instructions, and code snippets, covering dozens of domains. Finally, we further expand the pre-training corpus by incorporating additional multilingual data and introducing more languages. Compared to the pre-training data used in Qwen2.5, the number of supported languages has been significantly increased from 29 to 119, enhancing the model's linguistic coverage and cross-lingual capabilities.

We have developed a multilingual data annotation system designed to enhance both the quality and diversity of training data. This system has been applied to our large-scale pre-training datasets, annotating over 30 trillion tokens across multiple dimensions such as educational value, fields, domains, and safety. These detailed annotations support more effective data filtering and combination. Unlike previous studies (Xie et al., 2023; Fan et al., 2023; Liu et al., 2024b) that optimize the data mixture at the data source or domain level, our method optimizes the data mixture at the instance-level through extensive ablation experiments on small proxy models with the fine-grained data labels.

3.2 Pre-training Stage

The Qwen3 models are pre-trained through a three-stage process:

(1) **General Stage (S1)**: At the first pre-training stage, all Qwen3 models are trained on over 30 trillion tokens using a sequence length of 4,096 tokens. At this stage, the models have been fully pre-trained on language proficiency and general world knowledge, with training data covering 119 languages and dialects.

(2) **Reasoning Stage (S2)**: To further improve the reasoning ability, we optimize the pre-training corpus of this stage by increasing the proportion of STEM, coding, reasoning, and synthetic data. The models are further pre-trained with about 5T higher-quality tokens at a sequence length of 4,096 tokens. We also accelerate the learning rate decay during this stage.

(3) **Long Context Stage**: In the final pre-training stage, we collect high-quality long context corpora to extend the context length of Qwen3 models. All models are pre-trained on hundreds of billions of tokens with a sequence length of 32,768 tokens. The long context corpus includes 75% of text between 16,384 to 32,768 tokens in length, and 25% of text between 4,096 to 16,384 in length. Following Qwen2.5 (Yang et al., 2024b), we increase the base frequency of RoPE from 10,000 to 1,000,000 using the ABF technique (Xiong et al., 2023). Meanwhile, we introduce YARN (Peng et al., 2023) and Dual Chunk Attention (DCA, An et al., 2024) to achieve a four-fold increase in sequence length capacity during inference.

Similar to Qwen2.5 (Yang et al., 2024b), we develop scaling laws for optimal hyper-parameters (e.g., learning rate scheduler, and batch size) predictions based on three pre-training stages mentioned above. Through extensive experiments, we systematically study the relationship between model architecture, training data, training stage, and optimal training hyper-parameters. Finally, we set the predicted optimal learning rate and batch size strategy for each dense or MoE model.

3.3 Pre-training Evaluation

We conduct comprehensive evaluations of the base language models of the Qwen3 series. The evaluation of base models mainly focuses on their performance in general knowledge, reasoning, mathematics, scientific knowledge, coding, and multilingual capabilities. The evaluation datasets for pre-trained base models include 15 benchmarks:

- **General Tasks**: MMLU (Hendrycks et al., 2021a) (5-shot), MMLU-Pro (Wang et al., 2024) (5-shot, CoT), MMLU-redux (Gema et al., 2024) (5-shot), BBH (Suzgun et al., 2023) (3-shot, CoT), SuperGPQA (Du et al., 2025)(5-shot, CoT).
- **Math & STEM Tasks**: GPQA (Rein et al., 2023) (5-shot, CoT), GSM8K (Cobbe et al., 2021) (4-shot, CoT), MATH (Hendrycks et al., 2021b) (4-shot, CoT).

- **Coding Tasks**: EvalPlus (Liu et al., 2023a) (0-shot) (Average of HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), Humaneval+, MBPP+) (Liu et al., 2023a), MultiPL-E (Cassano et al., 2023) (0-shot) (Python, C++, JAVA, PHP, TypeScript, C#, Bash, JavaScript), MBPP-3shot (Austin et al., 2021), CRUX-O of CRUXEval (1-shot) (Gu et al., 2024).
- **Multilingual Tasks**: MGSM (Shi et al., 2023) (8-shot, CoT), MMMLU (OpenAI, 2024) (5-shot), INCLUDE (Romanou et al., 2024) (5-shot).

For the base model baselines, we compare the Qwen3 series base models with the Qwen2.5 base models (Yang et al., 2024b) and other leading open-source base models, including DeepSeek-V3 Base (Liu et al., 2024a), Gemma-3 (Team et al., 2025), Llama-3 (Dubey et al., 2024), and Llama-4 (Meta-AI, 2025) series base models, in terms of scale of parameters. All models are evaluated using the same evaluation pipeline and the widely-used evaluation settings to ensure fair comparison.

Summary of Evaluation Results Based on the overall evaluation results, we highlight some key conclusions of Qwen3 base models.

(1) Compared with the previously open-source SOTA dense and MoE base models (such as DeepSeek-V3 Base, Llama-4-Maverick Base, and Qwen2.5-72B-Base), Qwen3-235B-A22B-Base outperforms these models in most tasks with significantly fewer total parameters or activated parameters.

(2) For the Qwen3 MoE base models, our experimental results indicate that: (a) Using the same pre-training data, Qwen3 MoE base models can achieve similar performance to Qwen3 dense base models with only **1/5** activated parameters. (b) Due to the improvements of the Qwen3 MoE architecture, the scale-up of the training tokens, and more advanced training strategies, the Qwen3 MoE base models can outperform the Qwen2.5 MoE base models with less than **1/2** activated parameters and fewer total parameters. (c) Even with **1/10** of the activated parameters of the Qwen2.5 dense base model, the Qwen3 MoE base model can achieve comparable performance, which brings us significant advantages in inference and training costs.

(3) The overall performance of the Qwen3 dense base models is comparable to the Qwen2.5 base models at higher parameter scales. For example, Qwen3-1.7B/4B/8B/14B/32B-Base achieve comparable performance to Qwen2.5-3B/7B/14B/32B/72B-Base, respectively. Especially in STEM, coding, and reasoning benchmarks, the performance of Qwen3 dense base models even surpasses Qwen2.5 base models at higher parameter scales.

The detailed results are as follows.

Qwen3-235B-A22B-Base We compare Qwen3-235B-A22B-Base to our previous similar-sized MoE Qwen2.5-Plus-Base (Yang et al., 2024b) and other leading open-source base models: Llama-4-Maverick (Meta-AI, 2025), Qwen2.5-72B-Base (Yang et al., 2024b), DeepSeek-V3 Base (Liu et al., 2024a). From the results in Table 3, the Qwen3-235B-A22B-Base model attains the highest performance scores across most of the evaluated benchmarks. We further compare Qwen3-235B-A22B-Base with other baselines separately for the detailed analysis.

(1) Compared with the recently open-source model Llama-4-Maverick-Base, which has about **twice** the number of parameters, Qwen3-235B-A22B-Base still performs better on most benchmarks.

(2) Compared with the previously state-of-the-art open-source model DeepSeek-V3-Base, Qwen3-235B-A22B-Base outperforms DeepSeek-V3-Base on 14 out of 15 evaluation benchmarks with only about **1/3** the total number of parameters and **2/3** activated parameters, demonstrating the powerful and cost-effectiveness of our models.

(3) Compared with our previous MoE Qwen2.5-Plus of similar size, Qwen3-235B-A22B-Base significantly outperforms it with fewer parameters and activated parameters, which shows the remarkable advantages of Qwen3 in pre-training data, training strategy, and model architecture.

(4) Compared with our previous flagship open-source dense model Qwen2.5-72B-Base, Qwen3-235B-A22B-Base surpasses the latter in all benchmarks and uses fewer than **1/3** of the activated parameters. Meanwhile, due to the advantage of the model architecture, the inference costs and training costs on each trillion tokens of Qwen3-235B-A22B-Base are much cheaper than those of Qwen2.5-72B-Base.

Qwen3-32B-Base Qwen3-32B-Base is our largest dense model among the Qwen3 series. We compare it to the baselines of similar sizes, including Gemma-3-27B (Team et al., 2025) and Qwen2.5-32B (Yang et al., 2024b). In addition, we introduce two strong baselines: the recently open-source MoE model Llama-4-Scout, which has three times the parameters of Qwen3-32B-Base but half the activated parameters;