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