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