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).