The post-training pipeline of Qwen3 is strategically designed with two core objectives:

(1) **Thinking Control**: This involves the integration of two distinct modes, namely the "non-thinking" and "thinking" modes, providing users with the flexibility to choose whether the model should engage in reasoning or not, and to control the depth of thinking by specifying a token budget for the thinking process.

(2) **Strong-to-Weak Distillation**: This aims to streamline and optimize the post-training process for lightweight models. By leveraging the knowledge from large-scale models, we substantially reduce both the computational costs and the development efforts required for building smaller-scale models.

As illustrated in Figure 1, the flagship models in the Qwen3 series follow a sophisticated four-stage training process. The first two stages focus on developing the models' "thinking" abilities. The next two stages aim to integrate strong "non-thinking" functionalities into the models.

Preliminary experiments suggest that directly distilling the output logits from teacher models into lightweight student models can effectively enhance their performance while maintaining fine-grained control over their reasoning processes. This approach eliminates the necessity of performing an exhaustive four-stage training process individually for every small-scale model. It leads to better immediate performance, as indicated by higher Pass@1 scores, and also improves the model's ability of exploration, as reflected in improved Pass@64 results. In addition, it achieves these gains with much greater training efficiency, requiring only 1/10 of the GPU hours compared to the four-stage training method.

In the following sections, we present the four-stage training process and provide a detailed explanation of the Strong-to-Weak Distillation approach.

4.1 Long-CoT Cold Start

We begin by curating a comprehensive dataset that spans a wide range of categories, including math, code, logical reasoning, and general STEM problems. Each problem in the dataset is paired with verified reference answers or code-based test cases. This dataset serves as the foundation for the "cold start" phase of long Chain-of-Thought (long-CoT) training.

The dataset construction involves a rigorous two-phase filtering process: query filtering and response filtering. In the query filtering phase, we use Qwen2.5-72B-Instruct to identify and remove queries that are not easily verifiable. This includes queries containing multiple sub-questions or those asking for general text generation. Furthermore, we exclude queries that Qwen2.5-72B-Instruct can answer correctly without using CoT reasoning. This helps prevent the model from relying on superficial guessing and ensures that only complex problems requiring deeper reasoning are included. Additionally, we annotate each query's domain using Qwen2.5-72B-Instruct to maintain balanced domain representation across the dataset.

After reserving a validation query set, we generate N candidate responses for each remaining query using QwQ-32B (Qwen Team, 2025). When QwQ-32B consistently fails to generate correct solutions, human annotators manually assess the accuracy of the responses. For queries with positive Pass@N, further stringent filtering criteria are applied to remove responses that (1) yield incorrect final answers, (2) contain substantial repetition, (3) clearly indicate guesswork without adequate reasoning, (4) exhibit inconsistencies between the thinking and summary contents, (5) involve inappropriate language mixing or stylistic shifts, or (6) are suspected of being overly similar to potential validation set items. Subsequently, a carefully selected subset of the refined dataset is used for the initial cold-start training of the reasoning patterns. The objective at this stage is to instill foundational reasoning patterns in the model without overly emphasizing immediate reasoning performance. This approach ensures that the model's potential is not limited, allowing for greater flexibility and improvement during the subsequent reinforcement learning (RL) phase. To achieve this objective effectively, it is preferable to minimize both the number of training samples and the training steps during this preparatory phase.

4.2 Reasoning RL

The query-verifier pairs used in the Reasoning RL stage must satisfy the following four criteria: (1) They were not used during the cold-start phase. (2) They are learnable for the cold-start model. (3) They are as challenging as possible. (4) They cover a broad range of sub-domains. We ultimately collect a total of 3,995 query-verifier pairs, and employed GRPO (Shao et al., 2024) to update the model parameters. We observe that using a large batch size and a high number of rollouts per query, along with off-policy training to improve sample efficiency, is beneficial to the training process. We have also addressed how to balance exploration and exploitation by controlling the model's entropy to increase steadily or remain