and our previous flagship open-source dense model Qwen2.5-72B-Base, which has more than twice the number of parameters compared to Qwen3-32B-Base. The results are shown in Table 4, which support three key conclusions:

(1) Compared with the similar-sized models, Qwen3-32B-Base outperforms Qwen2.5-32B-Base and Gemma-3-27B Base on most benchmarks. Notably, Qwen3-32B-Base achieves 65.54 on MMLU-Pro and 39.78 on SuperGPQA, significantly outperforming its predecessor Qwen2.5-32B-Base. In addition, Qwen3-32B-Base achieves significantly higher encoding benchmark scores than all baseline models.

(2) Surprisingly, we find that Qwen3-32B-Base achieves competitive results compared to Qwen2.5-72B-Base. Although Qwen3-32B-Base has less than half the number of parameters of Qwen2.5-72B-Base, it outperforms Qwen2.5-72B-Base in 10 of the 15 evaluation benchmarks. On coding, mathematics, and reasoning benchmarks, Qwen3-32B-Base has remarkable advantages.

(3) Compared to Llama-4-Scout-Base, Qwen3-32B-Base significantly outperforms it on all 15 benchmarks, with only one-third of the number of parameters of Llama-4-Scout-Base, but twice the number of activated parameters.

Qwen3-14B-Base & Qwen3-30B-A3B-Base The evaluation of the Qwen3-14B-Base and Qwen3-30B-A3B-Base is compared against baselines of similar sizes, including Gemma-3-12B Base, Qwen2.5-14B Base. Similarly, we also introduce two strong baselines: (1) Qwen2.5-Turbo (Yang et al., 2024b), which has 42B parameters and 6B activated parameters. Note that its activated parameters are twice those of Qwen3-30B-A3B-Base. (2) Qwen2.5-32B-Base, which has 11 times the activated parameters of Qwen3-30B-A3B and more than twice that of Qwen3-14B. The results are shown in Table 5, where we can draw the following conclusions.

(1) Compared with the similar-sized models, Qwen3-14B-Base significantly performs better than Qwen2.5-14B-Base and Gemma-3-12B-Base on all 15 benchmarks.

(2) Similarly, Qwen3-14B-Base also achieves very competitive results compared to Qwen2.5-32B-Base with less than half of the parameters.

(3) With only 1/5 activated non-embedding parameters, Qwen3-30B-A3B significantly outperforms Qwen2.5-14B-Base on all tasks, and achieves comparable performance to Qwen3-14B-Base and Qwen2.5-32B-Base, which brings us significant advantages in inference and training costs.

Qwen3-8B / 4B / 1.7B / 0.6B-Base For edge-side models, we take similar-sized Qwen2.5, Llama-3, and Gemma-3 base models as the baselines. The results can be seen in Table 6, Table 7, and Table 8. All Qwen3 8B / 4B / 1.7B / 0.6B-Base models continue to maintain strong performance across nearly all benchmarks. Notably, Qwen3-8B / 4B / 1.7B-Base models even outperform larger size Qwen2.5-14B / 7B / 3B Base models on over half of the benchmarks, especially on STEM-related and coding benchmarks, reflecting the significant improvement of the Qwen3 models.

4 Post-training

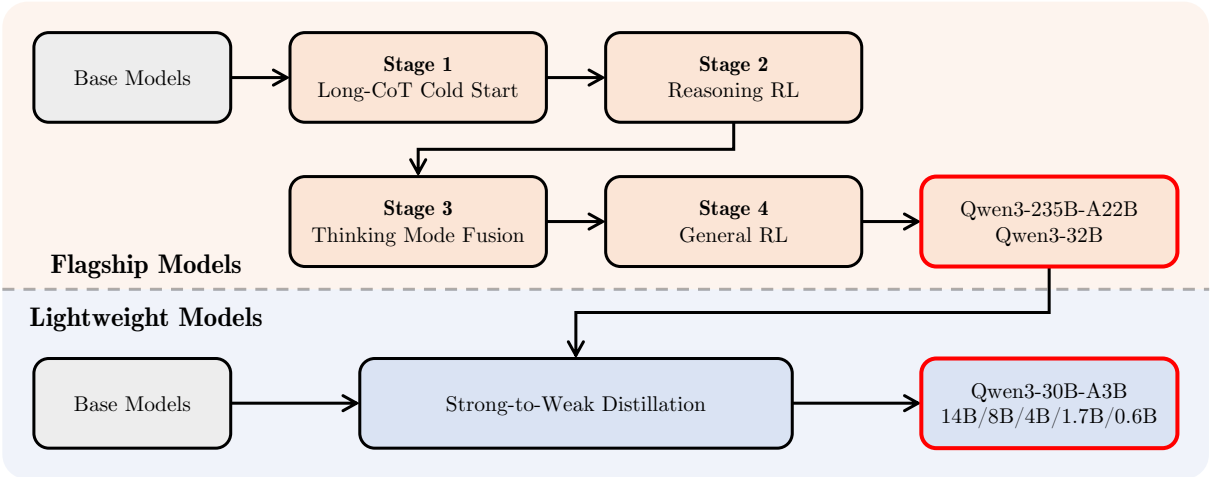

Figure 1: Post-training pipeline of the Qwen3 series models.