- **Coding Tasks**: EvalPlus (Liu et al., 2023a) (0-shot) (Average of HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), Humaneval+, MBPP+) (Liu et al., 2023a), MultiPL-E (Cassano et al., 2023) (0-shot) (Python, C++, JAVA, PHP, TypeScript, C#, Bash, JavaScript), MBPP-3shot (Austin et al., 2021), CRUX-O of CRUXEval (1-shot) (Gu et al., 2024).
- **Multilingual Tasks**: MGSM (Shi et al., 2023) (8-shot, CoT), MMMLU (OpenAI, 2024) (5-shot), INCLUDE (Romanou et al., 2024) (5-shot).

For the base model baselines, we compare the Qwen3 series base models with the Qwen2.5 base models (Yang et al., 2024b) and other leading open-source base models, including DeepSeek-V3 Base (Liu et al., 2024a), Gemma-3 (Team et al., 2025), Llama-3 (Dubey et al., 2024), and Llama-4 (Meta-AI, 2025) series base models, in terms of scale of parameters. All models are evaluated using the same evaluation pipeline and the widely-used evaluation settings to ensure fair comparison.

Summary of Evaluation Results Based on the overall evaluation results, we highlight some key conclusions of Qwen3 base models.

(1) Compared with the previously open-source SOTA dense and MoE base models (such as DeepSeek-V3 Base, Llama-4-Maverick Base, and Qwen2.5-72B-Base), Qwen3-235B-A22B-Base outperforms these models in most tasks with significantly fewer total parameters or activated parameters.

(2) For the Qwen3 MoE base models, our experimental results indicate that: (a) Using the same pre-training data, Qwen3 MoE base models can achieve similar performance to Qwen3 dense base models with only **1/5** activated parameters. (b) Due to the improvements of the Qwen3 MoE architecture, the scale-up of the training tokens, and more advanced training strategies, the Qwen3 MoE base models can outperform the Qwen2.5 MoE base models with less than **1/2** activated parameters and fewer total parameters. (c) Even with **1/10** of the activated parameters of the Qwen2.5 dense base model, the Qwen3 MoE base model can achieve comparable performance, which brings us significant advantages in inference and training costs.

(3) The overall performance of the Qwen3 dense base models is comparable to the Qwen2.5 base models at higher parameter scales. For example, Qwen3-1.7B/4B/8B/14B/32B-Base achieve comparable performance to Qwen2.5-3B/7B/14B/32B/72B-Base, respectively. Especially in STEM, coding, and reasoning benchmarks, the performance of Qwen3 dense base models even surpasses Qwen2.5 base models at higher parameter scales.

The detailed results are as follows.

Qwen3-235B-A22B-Base We compare Qwen3-235B-A22B-Base to our previous similar-sized MoE Qwen2.5-Plus-Base (Yang et al., 2024b) and other leading open-source base models: Llama-4-Maverick (Meta-AI, 2025), Qwen2.5-72B-Base (Yang et al., 2024b), DeepSeek-V3 Base (Liu et al., 2024a). From the results in Table 3, the Qwen3-235B-A22B-Base model attains the highest performance scores across most of the evaluated benchmarks. We further compare Qwen3-235B-A22B-Base with other baselines separately for the detailed analysis.

(1) Compared with the recently open-source model Llama-4-Maverick-Base, which has about **twice** the number of parameters, Qwen3-235B-A22B-Base still performs better on most benchmarks.

(2) Compared with the previously state-of-the-art open-source model DeepSeek-V3-Base, Qwen3-235B-A22B-Base outperforms DeepSeek-V3-Base on 14 out of 15 evaluation benchmarks with only about **1/3** the total number of parameters and **2/3** activated parameters, demonstrating the powerful and cost-effectiveness of our models.

(3) Compared with our previous MoE Qwen2.5-Plus of similar size, Qwen3-235B-A22B-Base significantly outperforms it with fewer parameters and activated parameters, which shows the remarkable advantages of Qwen3 in pre-training data, training strategy, and model architecture.

(4) Compared with our previous flagship open-source dense model Qwen2.5-72B-Base, Qwen3-235B-A22B-Base surpasses the latter in all benchmarks and uses fewer than **1/3** of the activated parameters. Meanwhile, due to the advantage of the model architecture, the inference costs and training costs on each trillion tokens of Qwen3-235B-A22B-Base are much cheaper than those of Qwen2.5-72B-Base.

Qwen3-32B-Base Qwen3-32B-Base is our largest dense model among the Qwen3 series. We compare it to the baselines of similar sizes, including Gemma-3-27B (Team et al., 2025) and Qwen2.5-32B (Yang et al., 2024b). In addition, we introduce two strong baselines: the recently open-source MoE model Llama-4-Scout, which has three times the parameters of Qwen3-32B-Base but half the activated parameters;