Table 10: **Multilingual benchmarks and the included languages.** The languages are identified in IETF language tags.

Benchmark	# Langs	Languages
Multi-IF	8	en, es, fr, hi, it, pt, ru, zh
INCLUDE	44	ar, az, be, bg, bn, de, el, es, et, eu, fa, fi, fr, he, hi, hr, hu, hy, id, it, ja, ka, kk, ko, lt, mk, ml, ms, ne, nl, pl, pt, ru, sq, sr, ta, te, tl, tr, uk, ur, uz, vi, zh
MMMLU	14	ar, bn, de, en, es, fr, hi, id, it, ja, ko, pt, sw, zh
MT-AIME2024	55	af, ar, bg, bn, ca, cs, cy, da, de, el, en, es, et, fa, fi, fr, gu, he, hi, hr, hu, id, it, ja, kn, ko, lt, lv, mk, ml, mr, ne, nl, no, pa, pl, pt, ro, ru, sk, sl, so, sq, sv, sw, ta, te, th, tl, tr, uk, ur, vi, zh-Hans, zh-Hant
PolyMath	18	ar, bn, de, en, es, fr, id, it, ja, ko, ms, pt, ru, sw, te, th, vi, zh
MLogiQA	10	ar, en, es, fr, ja, ko, pt, th, vi, zh

categorized into several dimensions:

- **General Tasks**: We utilize benchmarks including MMLU-Redux (Gema et al., 2024), GPQA-Diamond (Rein et al., 2023), C-Eval (Huang et al., 2023), and LiveBench (2024-11-25) (White et al., 2024). For GPQA-Diamond, we sample 10 times for each query and report the averaged accuracy.
- **Alignment Tasks**: To evaluate how well the model aligns with human preferences, we employ a suite of specialized benchmarks. For instruction-following performance, we report the strict-prompt accuracy of IFEval (Zhou et al., 2023). To assess alignment with human preferences on general topics, we utilize Arena-Hard (Li et al., 2024) and AlignBench v1.1 (Liu et al., 2023b). For writing tasks, we rely on Creative Writing V3 (Paech, 2024) and WritingBench (Wu et al., 2025) to evaluate the model's proficiency and creativity.
- **Math & Text Reasoning**: For evaluating mathematical and logical reasoning skills, we employ high-level math benchmarks including MATH-500 (Lightman et al., 2023), AIME'24 and AIME'25 (AIME, 2025), and text reasoning tasks including ZebraLogic (Lin et al., 2025) and AutoLogi (Zhu et al., 2025). For AIME problems, each year's questions include Part I and Part II, totaling 30 questions. For each question, we sample 64 times and take the average accuracy as the final score.
- **Agent & Coding**: To test the model's proficiency in coding and agent-based tasks, we use BFCL v3 (Yan et al., 2024), LiveCodeBench (v5, 2024.10-2025.02) (Jain et al., 2024), and Codeforces Ratings from CodeElo (Quan et al., 2025). For BFCL, all Qwen3 models are evaluated using the FC format, and yarn was used to deploy the models to a context length of 64k for Multi-Turn evaluation. Some baselines are derived from the BFCL leaderboard, taking the higher scores between FC and Prompt formats. For models not reported on the leaderboard, the Prompt formats are evaluated. For LiveCodeBench, for the non-thinking mode, we use the officially recommended prompt, while for the thinking mode, we adjust the prompt template to allow the model to think more freely, by removing the restriction `You will not return anything except for the program`. To evaluate the performance gap between models and competitive programming experts, we use CodeForces to calculate Elo ratings. In our benchmark, each problem is solved by generating up to eight independent reasoning attempts.
- **Multilingual Tasks**: For multilingual capabilities, we evaluate four kinds of tasks: instruction following, knowledge, mathematics, and logical reasoning. Instruction following is assessed using Multi-IF (He et al., 2024), which focuses on 8 key languages. Knowledge assessment consisted of two types: regional knowledge evaluated through INCLUDE (Romanou et al., 2024), covering 44 languages, and general knowledge assessed with MMMLU (OpenAI, 2024) across 14 languages, excluding the unoptimized Yoruba language; for these two benchmarks, we sample only 10% of the original data to improve evaluation efficiency. The mathematics task employ MT-AIME2024 (Son et al., 2025), encompassing 55 languages, and PolyMath (Wang et al., 2025), which includes 18 languages. Logical reasoning is evaluated using MlogiQA, covering 10 languages, sourced from Zhang et al. (2024).

For all Qwen3 models in the thinking mode, we utilize a sampling temperature of 0.6, a top-p value of 0.95, and a top-k value of 20. Additionally, for Creative Writing v3 and WritingBench, we apply a presence penalty of 1.5 to encourage the generation of more diverse content. For Qwen3 models in the non-thinking mode, we configure the sampling hyperparameters with temperature = 0.7, top-p = 0.8, top-k = 20, and presence penalty = 1.5. For both the thinking and non-thinking modes, we set the max output length to 32,768 tokens, except AIME'24 and AIME'25 where we extend this length to 38,912 tokens to provide sufficient thinking space.