Qwen3 Technical Report

Qwen Team

https://huggingface.co/Qwen
https://modelscope.cn/organization/qwen
https://github.com/QwenLM/Qwen3

Abstract

In this work, we present Qwen3, the latest version of the Qwen model family. Qwen3 comprises a series of large language models (LLMs) designed to advance performance, efficiency, and multilingual capabilities. The Qwen3 series includes models of both dense and Mixture-of-Expert (MoE) architectures, with parameter scales ranging from 0.6 to 235 billion. A key innovation in Qwen3 is the integration of thinking mode (for complex, multi-step reasoning) and non-thinking mode (for rapid, context-driven responses) into a unified framework. This eliminates the need to switch between different models——such as chat-optimized models (e.g., GPT-4o) and dedicated reasoning models (e.g., QwQ-32B)——and enables dynamic mode switching based on user queries or chat templates. Meanwhile, Qwen3 introduces a thinking budget mechanism, allowing users to allocate computational resources adaptively during inference, thereby balancing latency and performance based on task complexity. Moreover, by leveraging the knowledge from the flagship models, we significantly reduce the computational resources required to build smaller-scale models, while ensuring their highly competitive performance. Empirical evaluations demonstrate that Qwen3 achieves state-of-the-art results across diverse benchmarks, including tasks in code generation, mathematical reasoning, agent tasks, etc., competitive against larger MoE models and proprietary models. Compared to its predecessor Qwen2.5, Qwen3 expands multilingual support from 29 to 119 languages and dialects, enhancing global accessibility through improved cross-lingual understanding and generation capabilities. To facilitate reproducibility and community-driven research and development, all Qwen3 models are publicly accessible under Apache 2.0.

1 Introduction

The pursuit of artificial general intelligence (AGI) or artificial super intelligence (ASI) has long been a goal for humanity. Recent advancements in large foundation models, e.g., GPT-4o (OpenAI, 2024), Claude 3.7 (Anthropic, 2025), Gemini 2.5 (DeepMind, 2025), DeepSeek-V3 (Liu et al., 2024a), Llama-4 (Meta-AI, 2025), and Qwen2.5 (Yang et al., 2024b), have demonstrated significant progress toward this objective. These models are trained on vast datasets spanning trillions of tokens across diverse domains and tasks, effectively distilling human knowledge and capabilities into their parameters. Furthermore, recent developments in reasoning models, optimized through reinforcement learning, highlight the potential for foundation models to enhance inference-time scaling and achieve higher levels of intelligence, e.g., o3 (OpenAI, 2025), DeepSeek-R1 (Guo et al., 2025). While most state-of-the-art models remain proprietary, the rapid growth of open-source communities has substantially reduced the performance gap between open-weight and closed-source models. Notably, an increasing number of top-tier models (Meta-AI, 2025; Liu et al., 2024a; Guo et al., 2025; Yang et al., 2024b) are now being released as open-source, fostering broader research and innovation in artificial intelligence.

In this work, we introduce Qwen3, the latest series in our foundation model family, Qwen. Qwen3 is a collection of open-weight large language models (LLMs) that achieve state-of-the-art performance across a wide variety of tasks and domains. We release both dense and Mixture-of-Experts (MoE) models, with the number of parameters ranging from 0.6 billion to 235 billion, to meet the needs of different downstream applications. Notably, the flagship model, Qwen3-235B-A22B, is an MoE model with a total of 235 billion parameters and 22 billion activated ones per token. This design ensures both high performance and efficient inference.

Qwen3 introduces several key advancements to enhance its functionality and usability. First, it integrates two distinct operating modes, thinking mode and non-thinking mode, into a single model. This allows users to switch between these modes without alternating between different models, e.g., switching from Qwen2.5 to QwQ (Qwen Team, 2024). This flexibility ensures that developers and users can adapt the model's behavior to suit specific tasks efficiently. Additionally, Qwen3 incorporates thinking budgets, providing users with fine-grained control over the level of reasoning effort applied by the model during task execution. This capability is crucial to the optimization of computational resources and performance, tailoring the model's thinking behavior to meet varying complexity in real-world applications. Furthermore, Qwen3 has been pre-trained on 36 trillion tokens covering up to 119 languages and dialects, effectively enhancing its multilingual capabilities. This broadened language support amplifies its potential for deployment in global use cases and international applications. These advancements together establish Qwen3 as a cutting-edge open-source large language model family, capable of effectively addressing complex tasks across various domains and languages.

The pre-training process for Qwen3 utilizes a large-scale dataset consisting of approximately 36 trillion tokens, curated to ensure linguistic and domain diversity. To efficiently expand the training data, we employ a multi-modal approach: Qwen2.5-VL (Bai et al., 2025) is finetuned to extract text from extensive PDF documents. We also generate synthetic data using domain-specific models: Qwen2.5-Math (Yang et al., 2024c) for mathematical content and Qwen2.5-Coder (Hui et al., 2024) for code-related data. The pre-training process follows a three-stage strategy. In the first stage, the model is trained on about 30 trillion tokens to build a strong foundation of general knowledge. In the second stage, it is further trained on knowledge-intensive data to enhance reasoning abilities in areas like science, technology, engineering, and mathematics (STEM) and coding. Finally, in the third stage, the model is trained on long-context data to increase its maximum context length from 4,096 to 32,768 tokens.

To better align foundation models with human preferences and downstream applications, we employ a multi-stage post-training approach that empowers both thinking (reasoning) and non-thinking modes. In the first two stages, we focus on developing strong reasoning abilities through long chain-of-thought (CoT) cold-start finetuning and reinforcement learning focusing on mathematics and coding tasks. In the final two stages, we combine data with and without reasoning paths into a unified dataset for further fine-tuning, enabling the model to handle both types of input effectively, and we then apply general-domain reinforcement learning to improve performance across a wide range of downstream tasks. For smaller models, we use strong-to-weak distillation, leveraging both off-policy and on-policy knowledge transfer from larger models to enhance their capabilities. Distillation from advanced teacher models significantly outperforms reinforcement learning in performance and training efficiency.

We evaluate both pre-trained and post-trained versions of our models across a comprehensive set of benchmarks spanning multiple tasks and domains. Experimental results show that our base pre-trained models achieve state-of-the-art performance. The post-trained models, whether in thinking or non-thinking mode, perform competitively against leading proprietary models and large mixture-of-experts (MoE) models such as o1, o3-mini, and DeepSeek-V3. Notably, our models excel in coding, mathematics, and agent-related tasks. For example, the flagship model Qwen3-235B-A22B achieves 85.7 on AIME'24

and 81.5 on AIME'25 (AIME, 2025), 70.7 on LiveCodeBench v5 (Jain et al., 2024), 2,056 on CodeForces, and 70.8 on BFCL v3 (Yan et al., 2024). In addition, other models in the Qwen3 series also show strong performance relative to their size. Furthermore, we observe that increasing the thinking budget for thinking tokens leads to a consistent improvement in the model's performance across various tasks.

In the following sections, we describe the design of the model architecture, provide details on its training procedures, present the experimental results of pre-trained and post-trained models, and finally, conclude this technical report by summarizing the key findings and outlining potential directions for future research.

2 Architecture

The Qwen3 series includes 6 dense models, namely Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Qwen3-14B, and Qwen3-32B, and 2 MoE models, Qwen3-30B-A3B and Qwen3-235B-A22B. The flagship model, Qwen3-235B-A22B, has a total of 235B parameters with 22B activated ones. Below, we elaborate on the architecture of the Qwen3 models.

The architecture of the Qwen3 dense models is similar to Qwen2.5 (Yang et al., 2024b), including using Grouped Query Attention (GQA, Ainslie et al., 2023), SwiGLU (Dauphin et al., 2017), Rotary Positional Embeddings (RoPE, Su et al., 2024), and RMSNorm (Jiang et al., 2023) with pre-normalization. Besides, we remove QKV-bias used in Qwen2 (Yang et al., 2024a) and introduce QK-Norm (Dehghani et al., 2023) to the attention mechanism to ensure stable training for Qwen3. Key information on model architecture is provided in Table 1.

The Qwen3 MoE models share the same fundamental architecture as the Qwen3 dense models. Key information on model architecture is provided in Table 2. We follow Qwen2.5-MoE (Yang et al., 2024b) and implement fine-grained expert segmentation (Dai et al., 2024). The Qwen3 MoE models have 128 total experts with 8 activated experts per token. Unlike Qwen2.5-MoE, the Qwen3-MoE design excludes shared experts. Furthermore, we adopt the global-batch load balancing loss (Qiu et al., 2025) to encourage expert specialization. These architectural and training innovations have yielded substantial improvements in model performance across downstream tasks.

Qwen3 models utilize Qwen's tokenizer (Bai et al., 2023), which implements byte-level byte-pair encoding (BBPE, Brown et al., 2020; Wang et al., 2020; Sennrich et al., 2016) with a vocabulary size of 151,669.

Table 1: Model architecture of Qwen3 dense models.

Models	Layers	Heads (Q / KV)	Tie Embedding	Context Length
Qwen3-0.6B	28	16 / 8	Yes	32K
Qwen3-1.7B	28	16 / 8	Yes	32K
Qwen3-4B	36	32 / 8	Yes	128K
Qwen3-8B	36	32 / 8	No	128K
Qwen3-14B	40	40 / 8	No	128K
Qwen3-32B	64	64 / 8	No	128K

Table 2: Model architecture of Qwen3 MoE models.

Models	Layers	Heads (Q / KV)	# Experts (Total / Activated)	Context Length
Qwen3-30B-A3B	48	32 / 4	128 / 8	128K
Qwen3-235B-A22B	94	64 / 4	128 / 8	128K

3 Pre-training

In this section, we describe the construction of our pretraining data, the details of our pretraining approach, and present experimental results from evaluating the base models on standard benchmarks.

3.1 Pre-training Data

Compared with Qwen2.5 (Yang et al., 2024b), we have significantly expanded the scale and diversity of our training data. Specifically, we collected twice as many pre-training tokens—covering three times more languages. All Qwen3 models are trained on a large and diverse dataset consisting of **119 languages and dialects**, with a total of **36 trillion tokens**. This dataset includes high-quality content in various

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).

- **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;

Table 3: **Comparison among Qwen3-235B-A22B-Base and other representative strong open-source baselines. The highest, the second-best scores are shown in bold and underlined, respectively.**

	Qwen2.5-72B Base	Qwen2.5-Plus Base	Llama-4-Maverick Base	DeepSeek-V3 Base	Qwen3-235B-A22B Base
Architecture	Dense	MoE	MoE	MoE	MoE
# Total Params	72B	271B	402B	671B	235B
# Activated Params	72B	37B	17B	37B	22B
General Tasks					
MMLU	86.06	85.02	85.16	87.19	**87.81**
MMLU-Redux	83.91	82.69	84.05	86.14	**87.40**
MMLU-Pro	58.07	63.52	63.91	59.84	**68.18**
SuperGPQA	36.20	37.18	40.85	41.53	**44.06**
BBH	86.30	85.60	83.62	86.22	**88.87**
Math & STEM Tasks					
GPQA	45.88	41.92	43.94	41.92	**47.47**
GSM8K	91.50	91.89	87.72	87.57	**94.39**
MATH	62.12	62.78	63.32	62.62	**71.84**
Coding Tasks					
EvalPlus	65.93	61.43	68.38	63.75	**77.60**
MultiPL-E	58.70	62.16	57.28	62.26	**65.94**
MBPP	76.00	74.60	75.40	74.20	**81.40**
CRUX-O	66.20	68.50	77.00	76.60	**79.00**
Multilingual Tasks					
MGSM	82.40	82.21	79.69	82.68	**83.53**
MMMLU	84.40	83.49	83.09	85.88	**86.70**
INCLUDE	69.05	66.97	73.47	**75.17**	73.46

Table 4: **Comparison among Qwen3-32B-Base and other strong open-source baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

	Qwen2.5-32B Base	Qwen2.5-72B Base	Gemma-3-27B Base	Llama-4-Scout Base	Qwen3-32B Base
Architecture	Dense	Dense	Dense	MoE	Dense
# Total Params	32B	72B	27B	109B	32B
# Activated Params	32B	72B	27B	17B	32B
General Tasks					
MMLU	83.32	**86.06**	78.69	78.27	83.61
MMLU-Redux	81.97	**83.91**	76.53	71.09	83.41
MMLU-Pro	55.10	58.07	52.88	56.13	**65.54**
SuperGPQA	33.55	36.20	29.87	26.51	**39.78**
BBH	84.48	86.30	79.95	82.40	**87.38**
Math & STEM Tasks					
GPQA	47.97	45.88	26.26	40.40	**49.49**
GSM8K	92.87	91.50	81.20	85.37	**93.40**
MATH	57.70	**62.12**	51.78	51.66	61.62
Coding Tasks					
EvalPlus	66.25	65.93	55.78	59.90	**72.05**
MultiPL-E	58.30	58.70	45.03	47.38	**67.06**
MBPP	73.60	76.00	68.40	68.60	**78.20**
CRUX-O	67.80	66.20	60.00	61.90	**72.50**
Multilingual Tasks					
MGSM	78.12	82.40	73.74	79.93	**83.06**
MMMLU	82.40	**84.40**	77.62	74.83	83.83
INCLUDE	64.35	**69.05**	68.94	68.09	67.87

Table 5: **Comparison among Qwen3-14B-Base, Qwen3-30B-A3B-Base, and other strong open-source baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

	Gemma-3-12B Base	Qwen2.5-14B Base	Qwen2.5-32B Base	Qwen2.5-Turbo Base	Qwen3-14B Base	Qwen3-30B-A3B Base
Architecture	Dense	Dense	Dense	MoE	Dense	MoE
# Total Params	12B	14B	32B	42B	14B	30B
# Activated Params	12B	14B	32B	6B	14B	3B
General Tasks						
MMLU	73.87	79.66	**83.32**	79.50	81.05	81.38
MMLU-Redux	70.70	76.64	**81.97**	77.11	79.88	81.17
MMLU-Pro	44.91	51.16	55.10	55.60	61.03	**61.49**
SuperGPQA	24.61	30.68	33.55	31.19	34.27	**35.72**
BBH	74.28	78.18	**84.48**	76.10	81.07	81.54
Math & STEM Tasks						
GPQA	31.31	32.83	**47.97**	41.41	39.90	43.94
GSM8K	78.01	90.22	**92.87**	88.32	92.49	91.81
MATH	44.43	55.64	57.70	55.60	**62.02**	59.04
Coding Tasks						
EvalPlus	52.65	60.70	66.25	61.23	**72.23**	71.45
MultiPL-E	43.03	54.79	58.30	53.24	61.69	**66.53**
MBPP	60.60	69.00	73.60	67.60	73.40	**74.40**
CRUX-O	52.00	61.10	67.80	60.20	**68.60**	67.20
Multilingual Tasks						
MGSM	64.35	74.68	78.12	70.45	**79.20**	79.11
MMMLU	72.50	78.34	**82.40**	79.76	79.69	81.46
INCLUDE	63.34	60.26	64.35	59.25	64.55	**67.00**

Table 6: **Comparison among Qwen8B-Base and other strong open-source baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

	Llama-3-8B Base	Qwen2.5-7B Base	Qwen2.5-14B Base	Qwen3-8B Base
Architecture	Dense	Dense	Dense	Dense
# Total Params	8B	7B	14B	8B
# Activated Params	8B	7B	14B	8B
General Tasks				
MMLU	66.60	74.16	**79.66**	76.89
MMLU-Redux	61.59	71.06	**76.64**	76.17
MMLU-Pro	35.36	45.00	51.16	**56.73**
SuperGPQA	20.54	26.34	30.68	**31.64**
BBH	57.70	70.40	78.18	**78.40**
Math & STEM Tasks				
GPQA	25.80	36.36	32.83	**44.44**
GSM8K	55.30	85.36	**90.22**	89.84
MATH	20.50	49.80	55.64	**60.80**
Coding Tasks				
EvalPlus	44.13	62.18	60.70	**67.65**
MultiPL-E	31.45	50.73	54.79	**58.75**
MBPP	48.40	63.40	69.00	**69.80**
CRUX-O	36.80	48.50	61.10	**62.00**
Multilingual Tasks				
MGSM	38.92	63.60	74.68	**76.02**
MMMLU	59.65	71.34	**78.34**	75.72
IINCLUDE	44.94	53.98	**60.26**	59.40

Table 7: **Comparison among Qwen3-4B-Base and other strong open-source baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

	Gemma-3-4B Base	Qwen2.5-3B Base	Qwen2.5-7B Base	Qwen3-4B Base
Architecture	Dense	Dense	Dense	Dense
# Total Params	4B	3B	7B	4B
# Activated Params	4B	3B	7B	4B
General Tasks				
MMLU	59.51	65.62	**74.16**	72.99
MMLU-Redux	56.91	63.68	71.06	**72.79**
MMLU-Pro	29.23	34.61	45.00	**50.58**
SuperGPQA	17.68	20.31	26.34	**28.43**
BBH	51.70	56.30	70.40	**72.59**
Math & STEM Tasks				
GPQA	24.24	26.26	36.36	**36.87**
GSM8K	43.97	79.08	85.36	**87.79**
MATH	26.10	42.64	49.80	**54.10**
Coding Tasks				
EvalPlus	43.23	46.28	62.18	**63.53**
MultiPL-E	28.06	39.65	50.73	**53.13**
MBPP	46.40	54.60	63.40	**67.00**
CRUX-O	34.00	36.50	48.50	**55.00**
Multilingual Tasks				
MGSM	33.11	47.53	63.60	**67.74**
MMMLU	59.62	65.55	71.34	**71.42**
INCLUDE	49.06	45.90	53.98	**56.29**

Table 8: **Comparison among Qwen3-1.7B-Base, Qwen3-0.6B-Base, and other strong open-source baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

	Qwen2.5-0.5B Base	Qwen3-0.6B Base	Gemma-3-1B Base	Qwen2.5-1.5B Base	Qwen3-1.7B Base
Architecture	Dense	Dense	Dense	Dense	Dense
# Total Params	0.5B	0.6B	1B	1.5B	1.7B
# Activated Params	0.5B	0.6B	1B	1.5B	1.7B
General Tasks					
MMLU	47.50	52.81	26.26	60.90	**62.63**
MMLU-Redux	45.10	51.26	25.99	58.46	**61.66**
MMLU-Pro	15.69	24.74	9.72	28.53	**36.76**
SuperGPQA	11.30	15.03	7.19	17.64	**20.92**
BBH	20.30	41.47	28.13	45.10	**54.47**
Math & STEM Tasks					
GPQA	24.75	26.77	24.75	24.24	**28.28**
GSM8K	41.62	59.59	2.20	68.54	**75.44**
MATH	19.48	32.44	3.66	35.00	**43.50**
Coding Tasks					
EvalPlus	31.85	36.23	8.98	44.80	**52.70**
MultiPL-E	18.70	24.58	5.15	33.10	**42.71**
MBPP	29.80	36.60	9.20	43.60	**55.40**
CRUX-O	12.10	27.00	3.80	29.60	**36.40**
Multilingual Tasks					
MGSM	12.07	30.99	1.74	32.82	**50.71**
MMMLU	31.53	50.16	26.57	60.27	**63.27**
INCLUDE	24.74	34.26	25.62	39.55	**45.57**

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

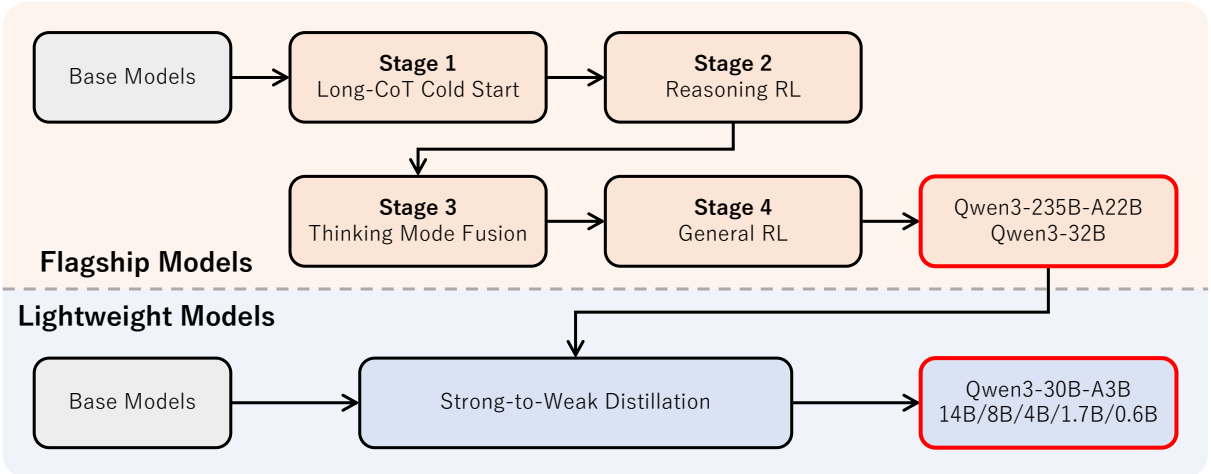

Figure 1: Post-training pipeline of the Qwen3 series models.

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

Table 9: **Examples of SFT data for thinking and non-thinking modes during the thinking mode fusion stage.** For the thinking mode, the `/think` flag can be omitted since it represents the default behavior. This feature has been implemented in the chat template[1] supported by the Hugging Face's tokenizer, where the thinking mode can be disabled using an additional parameter `enable_thinking=False`.

Thinking Mode	Non-Thinking Mode
<|im_start|>user {query} /think<|im_end|> <|im_start|>assistant <think> {thinking_content} </think> {response}<|im_end|>	<|im_start|>user {query} /no_think<|im_end|> <|im_start|>assistant <think> </think> {response}<|im_end|>

stable, which is crucial for maintaining stable training. As a result, we achieve consistent improvements in both training reward and validation performance over the course of a single RL run, without any manual intervention on hyperparameters. For instance, the AIME'24 score of the Qwen3-235B-A22B model increases from 70.1 to 85.1 over a total of 170 RL training steps.

4.3 Thinking Mode Fusion

The goal of the Thinking Mode Fusion stage is to integrate the "non-thinking" capabilities into the previously developed "thinking" model. This approach allows developers to manage and control reasoning behaviors, while also reducing the cost and complexity of deploying separate models for thinking and non-thinking tasks. To achieve this, we conduct continual supervised fine-tuning (SFT) on the Reasoning RL model and design a chat template to fuse the two modes. Moreover, we find that models capable of handling both modes proficiently perform consistently well under different thinking budgets.

Construction of SFT data. The SFT dataset combines both the "thinking" and "non-thinking" data. To ensure that the performance of the Stage 2 model is not compromised by the additional SFT, the "thinking" data is generated via rejection sampling on Stage 1 queries using the Stage 2 model itself. The "non-thinking" data, on the other hand, is carefully curated to cover a diverse range of tasks, including coding, mathematics, instruction-following, multilingual tasks, creative writing, question answering, and role-playing. Additionally, we employ automatically generated checklists for assessing the response quality of "non-thinking" data. To enhance the performance on tasks with low-resource languages, we particularly increase the proportion of translation tasks.

Chat Template Design. To better integrate the two modes and enable users to dynamically switch the model's thinking process, we design chat templates for Qwen3, as shown in Table 9. Specifically, for samples in thinking mode and non-thinking mode, we introduce `/think` and `/no_think` flags in the user query or system message, respectively. This allows the model to follow the user's input and select the appropriate thinking mode accordingly. For non-thinking mode samples, we retain an empty thinking block in the assistant's response. This design ensures internal format consistency within the model and allows developers to prevent the model from engaging in thinking behavior by concatenating an empty think block in the chat template. By default, the model operates in thinking mode; therefore, we add some thinking mode training samples where the user queries do not include `/think` flags. For more complex multi-turn dialogs, we randomly insert multiple `/think` and `/no_think` flags into users' queries, with the model response adhering to the last flag encountered.

Thinking Budget. An additional advantage of Thinking Mode Fusion is that, once the model learns to respond in both non-thinking and thinking modes, it naturally develops the ability to handle intermediate cases—generating responses based on incomplete thinking. This capability lays the foundation for implementing budget control over the model's thinking process. Specifically, when the length of the model's thinking reaches a user-defined threshold, we manually halt the thinking process and insert the stop-thinking instruction: "`Considering the limited time by the user, I have to give the solution based on the thinking directly now.\n</think>.\n\n`". After this instruction is inserted, the model proceeds to generate a final response based on its accumulated reasoning up to that point. It is worth noting that this ability is not explicitly trained but emerges naturally as a result of applying Thinking Mode Fusion.

4.4 General RL

The General RL stage aims to broadly enhance the models' capabilities and stability across diverse scenarios. To facilitate this, we have established a sophisticated **reward system** covering **over 20 distinct tasks**, each with customized scoring criteria. These tasks specifically target enhancements in the following core capabilities:

- **Instruction Following**: This capability ensures that models accurately interpret and follow user instructions, including requirements related to content, format, length, and the use of structured output, delivering responses that align with user expectations.
- **Format Following**: In addition to explicit instructions, we expect the model to adhere to specific formatting conventions. For instance, it should respond appropriately to the `/think` and `/no_think` flags by switching between thinking and non-thinking modes, and consistently use designated tokens (e.g., `<think>` and `</think>`) to separate the thinking and response parts in the final output.
- **Preference Alignment**: For open-ended queries, preference alignment focuses on improving the model's helpfulness, engagement, and style, ultimately delivering a more natural and satisfying user experience.
- **Agent Ability**: This involves training the model to correctly invoke tools via designated interfaces. During the RL rollout, the model is allowed to perform complete multi-turn interaction cycles with real environment execution feedback, thereby improving its performance and stability in long-horizon decision-making tasks.
- **Abilities for Specialized Scenarios**: In more specialized scenarios, we design tasks tailored to the specific context. For example, in Retrieval-Augmented Generation (RAG) tasks, we incorporate reward signals to guide the model toward generating accurate and contextually appropriate responses, thereby minimizing the risk of hallucination.

To provide feedback for the aforementioned tasks, we utilized three distinct types of rewards:

(1) **Rule-based Reward**: The rule-based reward has been widely used in the reasoning RL stage, and is also useful for general tasks such as instruction following (Lambert et al., 2024) and format adherence. Well-designed rule-based rewards can assess the correctness of model outputs with high precision, preventing issues like reward hacking.

(2) **Model-based Reward with Reference Answer**: In this approach, we provide a reference answer for each query and prompt Qwen2.5-72B-Instruct to score the model's response based on this reference. This method allows for more flexible handling of diverse tasks without requiring strict formatting, avoiding false negatives that can occur with purely rule-based rewards.

(3) **Model-based Reward without Reference Answer**: Leveraging human preference data, we train a reward model to assign scalar scores to model responses. This approach, which does not depend on a reference answer, can handle a broader range of queries while effectively enhancing the model's engagement and helpfulness.

4.5 Strong-to-Weak Distillation

The Strong-to-Weak Distillation pipeline is specifically designed to optimize lightweight models, encompassing 5 dense models (Qwen3-0.6B, 1.7B, 4B, 8B, and 14B) and one MoE model (Qwen3-30B-A3B). This approach enhances model performance while effectively imparting robust mode-switching capabilities. The distillation process is divided into two primary phases:

(1) **Off-policy Distillation**: At this initial phase, we combine the outputs of teacher models generated with both `/think` and `/no_think` modes for response distillation. This helps lightweight student models develop basic reasoning skills and the ability to switch between different modes of thinking, laying a solid foundation for the next on-policy training phase.

(2) **On-policy Distillation**: In this phase, the student model generates on-policy sequences for fine-tuning. Specifically, prompts are sampled, and the student model produces responses in either `/think` or `/no_think` mode. The student model is then fine-tuned by aligning its logits with those of a teacher model (Qwen3-32B or Qwen3-235B-A22B) to minimize the KL divergence.

4.6 Post-training Evaluation

To comprehensively evaluate the quality of instruction-tuned models, we adopted automatic benchmarks to assess model performance under both thinking and non-thinking modes. These benchmarks are

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.

Table 11: **Comparison among Qwen3-235B-A22B (Thinking) and other reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		OpenAI-o1	DeepSeek-R1	Grok-3-Beta (Think)	Gemini2.5-Pro	Qwen3-235B-A22B
	Architecture	-	MoE	-	-	MoE
	# Activated Params	-	37B	-	-	22B
	# Total Params	-	671B	-	-	235B
General Tasks	MMLU-Redux	92.8	92.9	-	**93.7**	92.7
	GPQA-Diamond	78.0	71.5	80.2	**84.0**	71.1
	C-Eval	85.5	**91.8**	-	82.9	89.6
	LiveBench 2024-11-25	75.7	71.6	-	**82.4**	77.1
Alignment Tasks	IFEval strict prompt	**92.6**	83.3	-	89.5	83.4
	Arena-Hard	92.1	92.3	-	**96.4**	95.6
	AlignBench v1.1	8.86	8.76	-	**9.03**	8.94
	Creative Writing v3	81.7	85.5	-	**86.0**	84.6
	WritingBench	7.69	7.71	-	**8.09**	8.03
Math & Text Reasoning	MATH-500	96.4	97.3		**98.8**	98.0
	AIME'24	74.3	79.8	83.9	**92.0**	85.7
	AIME'25	79.2	70.0	77.3	**86.7**	81.5
	ZebraLogic	81.0	78.7	-	**87.4**	80.3
	AutoLogi	79.8	86.1	-	85.4	**89.0**
Agent & Coding	BFCL v3	67.8	56.9	-	62.9	**70.8**
	LiveCodeBench v5	63.9	64.3	70.6	70.4	**70.7**
	CodeForces (Rating / Percentile)	1891 / 96.7%	2029 / 98.1%	-	2001 / 97.9%	**2056 / 98.2%**
Multilingual Tasks	Multi-IF	48.8	67.7	-	**77.8**	71.9
	INCLUDE	84.6	82.7	-	**85.1**	78.7
	MMMLU 14 languages	**88.4**	86.4	-	86.9	84.3
	MT-AIME2024	67.4	73.5	-	76.9	**80.8**
	PolyMath	38.9	47.1	-	52.2	**54.7**
	MLogiQA	75.5	73.8	-	75.6	**77.1**

Table 12: **Comparison among Qwen3-235B-A22B (Non-thinking) and other non-reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		GPT-4o -2024-11-20	DeepSeek-V3	Qwen2.5-72B -Instruct	LLaMA-4 -Maverick	Qwen3-235B-A22B
	Architecture	-	MoE	Dense	MoE	MoE
	# Activated Params	-	37B	72B	17B	22B
	# Total Params	-	671B	72B	402B	235B
General Tasks	MMLU-Redux	87.0	89.1	86.8	**91.8**	89.2
	GPQA-Diamond	46.0	59.1	49.0	**69.8**	62.9
	C-Eval	75.5	**86.5**	84.7	83.5	86.1
	LiveBench 2024-11-25	52.2	60.5	51.4	59.5	**62.5**
Alignment Tasks	IFEval strict prompt	86.5	86.1	84.1	**86.7**	83.2
	Arena-Hard	85.3	85.5	81.2	82.7	**96.1**
	AlignBench v1.1	8.42	8.64	7.89	7.97	**8.91**
	Creative Writing v3	**81.1**	74.0	61.8	61.3	80.4
	WritingBench	7.11	6.49	7.06	5.46	**7.70**
Math & Text Reasoning	MATH-500	77.2	90.2	83.6	90.6	**91.2**
	AIME'24	11.1	39.2	18.9	38.5	**40.1**
	AIME'25	7.6	**28.8**	15.0	15.9	24.7
	ZebraLogic	27.4	**42.1**	26.6	40.0	37.7
	AutoLogi	65.9	76.1	66.1	75.2	**83.3**
Agent & Coding	BFCL v3	**72.5**	57.6	63.4	52.9	68.0
	LiveCodeBench v5	32.7	33.1	30.7	**37.2**	35.3
	CodeForces (Rating / Percentile)	864 / 35.4%	1134 / 54.1%	859 / 35.0%	712 / 24.3%	**1387 / 75.7%**
Multilingual Tasks	Multi-IF	65.6	55.6	65.3	**75.5**	70.2
	INCLUDE	78.8	76.7	69.6	**80.9**	75.6
	MMMLU 14 languages	80.3	81.1	76.9	**82.5**	79.8
	MT-AIME2024	9.2	20.9	12.7	27.0	**32.4**
	PolyMath	13.7	20.4	16.9	26.1	**27.0**
	MLogiQA	57.4	58.9	59.3	59.9	**67.6**

Summary of Evaluation Results From the evaluation results, we summarize several key conclusions of the finalized Qwen3 models as follows:

(1) Our flagship model, Qwen3-235B-A22B, demonstrates the state-of-the-art overall performance among open-source models in both the thinking and non-thinking modes, surpassing strong baselines such as DeepSeek-R1 and DeepSeek-V3. Qwen3-235B-A22B is also highly competitive to closed-source leading models, such as OpenAI-o1, Gemini2.5-Pro, and GPT-4o, showcasing its profound reasoning capabilities and comprehensive general abilities.

(2) Our flagship dense model, Qwen3-32B, outperforms our previous strongest reasoning model, QwQ-32B, in most of the benchmarks, and performs comparably to the closed-source OpenAI-o3-mini, indicating its compelling reasoning capabilities. Qwen3-32B is also remarkably performant in the non-thinking mode and surpasses our previous flagship non-reasoning dense model, Qwen2.5-72B-Instruct.

(3) Our lightweight models, including Qwen3-30B-A3B, Qwen3-14B, and other smaller dense ones, possess consistently superior performance to the open-source models with a close or larger amount of parameters, proving the success of our Strong-to-Weak Distillation approach.

The detailed results are as follows.

Qwen3-235B-A22B For our flagship model Qwen3-235B-A22B, we compare it with the leading reasoning and non-reasoning models. For the thinking mode, we take OpenAI-o1 (OpenAI, 2024), DeepSeek-R1 (Guo et al., 2025), Grok-3-Beta (Think) (xAI, 2025), and Gemini2.5-Pro (DeepMind, 2025) as the reasoning baselines. For the non-thinking mode, we take GPT-4o-2024-11-20 (OpenAI, 2024), DeepSeek-V3 (Liu et al., 2024a), Qwen2.5-72B-Instruct (Yang et al., 2024b), and LLaMA-4-Maverick (Meta-AI, 2025) as the non-reasoning baselines. We present the evaluation results in Table 11 and 12.

(1) From Table 11, with only 60% activated and 35% total parameters, Qwen3-235B-A22B (Thinking) outperforms DeepSeek-R1 on **17/23** the benchmarks, particularly on the reasoning-demanded tasks (e.g., mathematics, agent, and coding), demonstrating the state-of-the-art reasoning capabilities of Qwen3-235B-A22B among open-source models. Moreover, Qwen3-235B-A22B (Thinking) is also highly competitive to the closed-source OpenAI-o1, Grok-3-Beta (Think), and Gemini2.5-Pro, substantially narrowing the gap in the reasoning capabilities between open-source and close-source models.

(2) From Table 12, Qwen3-235B-A22B (Non-thinking) exceeds the other leading open-source models, including DeepSeek-V3, LLaMA-4-Maverick, and our previous flagship model Qwen2.5-72B-Instruct, and also surpasses the closed-source GPT-4o-2024-11-20 in **18/23** the benchmarks, indicating its inherent strong capabilities even when not enhanced with the deliberate thinking process.

Qwen3-32B For our flagship dense model, Qwen3-32B, we take DeepSeek-R1-Distill-Llama-70B, OpenAI-o3-mini (medium), and our previous strongest reasoning model, QwQ-32B (Qwen Team, 2025), as the baselines in the thinking mode. We also take GPT-4o-mini-2024-07-18, LLaMA-4-Scout, and our previous flagship model, Qwen2.5-72B-Instruct, as the baselines in the non-thinking mode. We present the evaluation results in Table 13 and 14.

(1) From Table 13, Qwen3-32B (Thinking) outperforms QwQ-32B on **17/23** the benchmarks, making it the new state-of-the-art reasoning model at the sweet size of 32B. Moreover, Qwen3-32B (Thinking) also competes with the closed-source OpenAI-o3-mini (medium) with better alignment and multilingual performance.

(2) From Table 14, Qwen3-32B (Non-thinking) exhibits superior performance to all the baselines on almost all the benchmarks. Particularly, Qwen3-32B (Non-thinking) performs on par with Qwen2.5-72B-Instruct on the general tasks with significant advantages on the alignment, multilingual, and reasoning-related tasks, again proving the fundamental improvements of Qwen3 over our previous Qwen2.5 series models.

Qwen3-30B-A3B & Qwen3-14B For Qwen3-30B-A3B and Qwen3-14B, we compare them with DeepSeek-R1-Distill-Qwen-32B and QwQ-32B in the thinking mode, and Phi-4 (Abdin et al., 2024), Gemma-3-27B-IT (Team et al., 2025), and Qwen2.5-32B-Instruct in the non-thinking mode, respectively. We present the evaluation results in Table 15 and 16.

(1) From Table 15, Qwen3-30B-A3B and Qwen3-14B (Thinking) are both highly competitive to QwQ-32B, especially on the reasoning-related benchmarks. It is noteworthy that Qwen3-30B-A3B achieves comparable performance to QwQ-32B with a smaller model size and less than

Table 13: **Comparison among Qwen3-32B (Thinking) and other reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		DeepSeek-R1 -Distill-Llama-70B	QwQ-32B	OpenAI-o3-mini (medium)	Qwen3-32B
	Architecture	Dense	Dense	-	Dense
	# Activated Params	70B	32B	-	32B
	# Total Params	70B	32B	-	32B
General Tasks	MMLU-Redux	89.3	90.0	90.0	**90.9**
	GPQA-Diamond	65.2	65.6	**76.8**	68.4
	C-Eval	71.8	**88.4**	75.1	87.3
	LiveBench 2024-11-25	54.5	72.0	70.0	**74.9**
Alignment Tasks	IFEval strict prompt	79.3	83.9	**91.5**	85.0
	Arena-Hard	60.6	89.5	89.0	**93.8**
	AlignBench v1.1	6.74	8.70	8.38	**8.72**
	Creative Writing v3	62.1	**82.4**	74.8	81.0
	WritingBench	6.08	7.86	7.52	**7.90**
Math & Text Reasoning	MATH-500	94.5	**98.0**	**98.0**	97.2
	AIME'24	70.0	79.5	79.6	**81.4**
	AIME'25	56.3	69.5	**74.8**	72.9
	ZebraLogic	71.3	76.8	**88.9**	88.8
	AutoLogi	83.5	**88.1**	86.3	87.3
Agent & Coding	BFCL v3	49.3	66.4	64.6	**70.3**
	LiveCodeBench v5	54.5	62.7	**66.3**	65.7
	CodeForces (Rating / Percentile)	1633 / 91.4%	1982 / 97.7%	**2036 / 98.1%**	1977 / 97.7%
Multilingual Tasks	Multi-IF	57.6	68.3	48.4	**73.0**
	INCLUDE	62.1	69.7	73.1	**73.7**
	MMMLU 14 languages	69.6	**80.9**	79.3	80.6
	MT-AIME2024	29.3	68.0	73.9	**75.0**
	PolyMath	29.4	45.9	38.6	**47.4**
	MLogiQA	60.3	75.5	71.1	**76.3**

Table 14: **Comparison among Qwen3-32B (Non-thinking) and other non-reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		GPT-4o-mini -2024-07-18	LLaMA-4 -Scout	Qwen2.5-72B -Instruct	Qwen3-32B
	Architecture	-	MoE	Dense	Dense
	# Activated Params	-	17B	72B	32B
	# Total Params	-	109B	72B	32B
General Tasks	MMLU-Redux	81.5	86.3	**86.8**	85.7
	GPQA-Diamond	40.2	**57.2**	49.0	54.6
	C-Eval	66.3	78.2	**84.7**	83.3
	LiveBench 2024-11-25	41.3	47.6	51.4	**59.8**
Alignment Tasks	IFEval strict prompt	80.4	**84.7**	84.1	83.2
	Arena-Hard	74.9	70.5	81.2	**92.8**
	AlignBench v1.1	7.81	7.49	7.89	**8.58**
	Creative Writing v3	70.3	55.0	61.8	**78.3**
	WritingBench	5.98	5.49	7.06	**7.54**
Math & Text Reasoning	MATH-500	78.2	82.6	83.6	**88.6**
	AIME'24	8.1	28.6	18.9	**31.0**
	AIME'25	8.8	10.0	15.0	**20.2**
	ZebraLogic	20.1	24.2	26.6	**29.2**
	AutoLogi	52.6	56.8	66.1	**78.5**
Agent & Coding	BFCL v3	**64.0**	45.4	63.4	63.0
	LiveCodeBench v5	27.9	29.8	30.7	**31.3**
	CodeForces (Rating / Percentile)	1113 / 52.6%	981 / 43.7%	859 / 35.0%	**1353 / 71.0%**
Multilingual Tasks	Multi-IF	62.4	64.2	65.3	**70.7**
	INCLUDE	66.0	**74.1**	69.6	70.9
	MMMLU 14 languages	72.1	**77.5**	76.9	76.5
	MT-AIME2024	6.0	19.1	12.7	**24.1**
	PolyMath	12.0	20.9	16.9	**22.5**
	MLogiQA	42.6	53.9	59.3	**62.9**

Table 15: **Comparison among Qwen3-30B-A3B / Qwen3-14B (Thinking) and other reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		DeepSeek-R1 -Distill-Qwen-32B	QwQ-32B	Qwen3-14B	Qwen3-30B-A3B
	Architecture	Dense	Dense	Dense	MoE
	# Activated Params	32B	32B	14B	3B
	# Total Params	32B	32B	14B	30B
General Tasks	MMLU-Redux	88.2	**90.0**	88.6	89.5
	GPQA-Diamond	62.1	65.6	64.0	**65.8**
	C-Eval	82.2	**88.4**	86.2	86.6
	LiveBench 2024-11-25	45.6	72.0	71.3	**74.3**
Alignment Tasks	IFEval strict prompt	72.5	83.9	85.4	**86.5**
	Arena-Hard	60.8	89.5	**91.7**	91.0
	AlignBench v1.1	7.25	**8.70**	8.56	**8.70**
	Creative Writing v3	55.0	**82.4**	80.3	79.1
	WritingBench	6.13	**7.86**	7.80	7.70
Math & Text Reasoning	MATH-500	94.3	**98.0**	96.8	**98.0**
	AIME'24	72.6	79.5	79.3	**80.4**
	AIME'25	49.6	69.5	70.4	**70.9**
	ZebraLogic	69.6	76.8	88.5	**89.5**
	AutoLogi	74.6	88.1	**89.2**	88.7
Agent & Coding	BFCL v3	53.5	66.4	**70.4**	69.1
	LiveCodeBench v5	54.5	62.7	**63.5**	62.6
	CodeForces (Rating / Percentile)	1691 / 93.4%	**1982 / 97.7%**	1766 / 95.3%	1974 / 97.7%
Multilingual Tasks	Multi-IF	31.3	68.3	**74.8**	72.2
	INCLUDE	68.0	69.7	71.7	**71.9**
	MMMLU 14 languages	78.6	**80.9**	77.9	78.4
	MT-AIME2024	44.6	68.0	73.3	**73.9**
	PolyMath	35.1	45.9	45.8	**46.1**
	MLogiQA	63.3	**75.5**	71.1	70.1

Table 16: **Comparison among Qwen3-30B-A3B / Qwen3-14B (Non-thinking) and other non-reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		Phi-4	Gemma-3 -27B-IT	Qwen2.5-32B -Instruct	Qwen3-14B	Qwen3-30B-A3B
	Architecture	Dense	Dense	Dense	Dense	MoE
	# Activated Params	14B	27B	32B	14B	3B
	# Total Params	14B	27B	32B	14B	30B
General Tasks	MMLU-Redux	**85.3**	82.6	83.9	82.0	84.1
	GPQA-Diamond	**56.1**	42.4	49.5	54.8	54.8
	C-Eval	66.9	66.6	80.6	81.0	**82.9**
	LiveBench 2024-11-25	41.6	49.2	50.0	**59.6**	59.4
Alignment Tasks	IFEval strict prompt	62.1	80.6	79.5	**84.8**	83.7
	Arena-Hard	75.4	86.8	74.5	86.3	**88.0**
	AlignBench v1.1	7.61	7.80	7.71	8.52	**8.55**
	Creative Writing v3	51.2	**82.0**	54.6	73.1	68.1
	WritingBench	5.73	7.22	5.90	**7.24**	7.22
Math & Text Reasoning	MATH-500	80.8	**90.0**	84.6	**90.0**	89.8
	AIME'24	22.9	32.6	18.8	31.7	**32.8**
	AIME'25	17.3	**24.0**	12.8	23.3	21.6
	ZebraLogic	32.3	24.6	26.1	33.0	**33.2**
	AutoLogi	66.2	64.2	65.5	**82.0**	81.5
Agent & Coding	BFCL v3	47.0	59.1	**62.8**	61.5	58.6
	LiveCodeBench v5	25.2	26.9	26.4	29.0	**29.8**
	CodeForces (Rating / Percentile)	**1280 / 65.3%**	1063 / 49.3%	903 / 38.2%	1200 / 58.6%	1267 / 64.1%
Multilingual Tasks	Multi-IF	49.5	69.8	63.2	**72.9**	70.8
	INCLUDE	65.3	**71.4**	67.5	67.8	67.8
	MMMLU 14 languages	74.7	**76.1**	74.2	72.6	73.8
	MT-AIME2024	13.1	23.0	15.3	23.2	**24.6**
	PolyMath	17.4	20.3	18.3	22.0	**23.3**
	MLogiQA	53.1	58.5	58.0	**58.9**	53.3

Table 17: **Comparison among Qwen3-8B / Qwen3-4B (Thinking) and other reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		DeepSeek-R1 -Distill-Qwen-14B	DeepSeek-R1 -Distill-Qwen-32B	Qwen3-4B	Qwen3-8B
	Architecture	Dense	Dense	Dense	Dense
	# Activated Params	14B	32B	4B	8B
	# Total Params	14B	32B	4B	8B
General Tasks	MMLU-Redux	84.1	**88.2**	83.7	87.5
	GPQA-Diamond	59.1	**62.1**	55.9	62.0
	C-Eval	78.1	82.2	77.5	**83.4**
	LiveBench 2024-11-25	52.3	45.6	63.6	**67.1**
Alignment Tasks	IFEval strict prompt	72.6	72.5	81.9	**85.0**
	Arena-Hard	48.0	60.8	76.6	**85.8**
	AlignBench v1.1	7.43	7.25	8.30	**8.46**
	Creative Writing v3	54.2	55.0	61.1	**75.0**
	WritingBench	6.03	6.13	7.35	**7.59**
Math & Text Reasoning	MATH-500	93.9	94.3	97.0	**97.4**
	AIME'24	69.7	72.6	73.8	**76.0**
	AIME'25	44.5	49.6	65.6	**67.3**
	ZebraLogic	59.1	69.6	81.0	**84.8**
	AutoLogi	78.6	74.6	87.9	**89.1**
Agent & Coding	BFCL v3	49.5	53.5	65.9	**68.1**
	LiveCodeBench v5	45.5	54.5	54.2	**57.5**
	CodeForces (Rating / Percentile)	1574 / 89.1%	1691 / 93.4%	1671 / 92.8%	**1785 / 95.6%**
Multilingual Tasks	Multi-IF	29.8	31.3	66.3	**71.2**
	INCLUDE	59.7	**68.0**	61.8	67.8
	MMMLU 14 languages	73.8	**78.6**	69.8	74.4
	MT-AIME2024	33.7	44.6	60.7	**65.4**
	PolyMath	28.6	35.1	40.0	**42.7**
	MLogiQA	53.6	63.3	65.9	**69.0**

Table 18: **Comparison among Qwen3-8B / Qwen3-4B (Non-thinking) and other non-reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		LLaMA-3.1-8B -Instruct	Gemma-3 -12B-IT	Qwen2.5-7B -Instruct	Qwen2.5-14B -Instruct	Qwen3-4B	Qwen3-8B
	Architecture	Dense	Dense	Dense	Dense	Dense	Dense
	# Activated Params	8B	12B	7B	14B	4B	8B
	# Total Params	8B	12B	7B	14B	4B	8B
General Tasks	MMLU-Redux	61.7	77.8	75.4	**80.0**	77.3	79.5
	GPQA-Diamond	32.8	40.9	36.4	**45.5**	41.7	39.3
	C-Eval	52.0	61.1	76.2	**78.0**	72.2	77.9
	LiveBench 2024-11-25	26.0	43.7	34.9	42.2	48.4	**53.5**
Alignment Tasks	IFEval strict prompt	75.0	80.2	71.2	81.0	81.2	**83.0**
	Arena-Hard	30.1	**82.6**	52.0	68.3	66.2	79.6
	AlignBench v1.1	6.01	7.77	7.27	7.67	8.10	**8.38**
	Creative Writing v3	52.8	**79.9**	49.8	55.8	53.6	64.5
	WritingBench	4.57	7.05	5.82	5.93	6.85	**7.15**
Math & Text Reasoning	MATH-500	54.8	85.6	77.6	83.4	84.8	**87.4**
	AIME'24	6.3	22.4	9.1	15.2	25.0	**29.1**
	AIME'25	2.7	18.8	12.1	13.6	19.1	**20.9**
	ZebraLogic	12.8	17.8	12.0	19.7	**35.2**	26.7
	AutoLogi	30.9	58.9	42.9	57.4	76.3	**76.5**
Agent & Coding	BFCL v3	49.6	50.6	55.8	58.7	57.6	**60.2**
	LiveCodeBench v5	10.8	**25.7**	14.4	21.9	21.3	22.8
	CodeForces (Rating / Percentile)	473 / 14.9%	462 / 14.7%	191 / 0.0%	904 / 38.3%	842 / 33.7%	**1110 / 52.4%**
Multilingual Tasks	Multi-IF	52.1	65.6	47.7	55.5	61.3	**69.2**
	INCLUDE	34.0	**65.3**	53.6	63.5	53.8	62.5
	MMMLU 14 languages	44.4	70.0	61.4	**70.3**	61.7	66.9
	MT-AIME2024	0.4	**16.7**	5.5	8.5	13.9	16.6
	PolyMath	5.8	17.6	11.9	15.0	16.6	**18.8**
	MLogiQA	41.9	**54.5**	49.5	51.3	49.9	51.4

Table 19: **Comparison among Qwen3-1.7B / Qwen3-0.6B (Thinking) and other reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		DeepSeek-R1 -Distill-Qwen-1.5B	DeepSeek-R1 -Distill-Llama-8B	Qwen3-0.6B	Qwen3-1.7B
	Architecture	Dense	Dense	Dense	Dense
	# Activated Params	1.5B	8B	0.6B	1.7B
	# Total Params	1.5B	8B	0.6B	1.7B
General Tasks	MMLU-Redux	45.4	66.4	55.6	**73.9**
	GPQA-Diamond	33.8	**49.0**	27.9	40.1
	C-Eval	27.1	50.4	50.4	**68.1**
	LiveBench 2024-11-25	24.9	40.6	30.3	**51.1**
Alignment Tasks	IFEval strict prompt	39.9	59.0	59.2	**72.5**
	Arena-Hard	4.5	17.6	8.5	**43.1**
	AlignBench v1.1	5.00	6.24	6.10	**7.60**
	Creative Writing v3	16.4	**51.1**	30.6	48.0
	WritingBench	4.03	5.42	5.61	**7.02**
Math & Text Reasoning	MATH-500	83.9	89.1	77.6	**93.4**
	AIME'24	28.9	**50.4**	10.7	48.3
	AIME'25	22.8	27.8	15.1	**36.8**
	ZebraLogic	4.9	37.1	30.3	**63.2**
	AutoLogi	19.1	63.4	61.6	**83.2**
Agent & Coding	BFCL v3	14.0	21.5	46.4	**56.6**
	LiveCodeBench v5	13.2	**42.5**	12.3	33.2
Multilingual Tasks	Multi-IF	13.3	27.0	36.1	**51.2**
	INCLUDE	21.9	34.5	35.9	**51.8**
	MMMLU 14 languages	27.3	40.1	43.1	**59.1**
	MT-AIME2024	12.4	13.2	7.8	**36.1**
	PolyMath	14.5	10.8	11.4	**25.2**
	MLogiQA	29.0	32.8	40.9	**56.0**

Table 20: **Comparison among Qwen3-1.7B / Qwen3-0.6B (Non-thinking) and other non-reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		Gemma-3 -1B-IT	Phi-4-mini	Qwen2.5-1.5B -Instruct	Qwen2.5-3B -Instruct	Qwen3-0.6B	Qwen3-1.7B
	Architecture	Dense	Dense	Dense	Dense	Dense	Dense
	# Activated Params	1.0B	3.8B	1.5B	3.1B	0.6B	1.7B
	# Total Params	1.0B	3.8B	1.5B	3.1B	0.6B	1.7B
General Tasks	MMLU-Redux	33.3	**67.9**	50.7	64.4	44.6	64.4
	GPQA-Diamond	19.2	25.2	29.8	**30.3**	22.9	28.6
	C-Eval	28.5	40.0	53.3	**68.2**	42.6	61.0
	LiveBench 2024-11-25	14.4	25.3	18.0	23.8	21.8	**35.6**
Alignment Tasks	IFEval strict prompt	54.5	**68.6**	42.5	58.2	54.5	68.2
	Arena-Hard	17.8	32.8	9.0	23.7	6.5	**36.9**
	AlignBench v1.1	5.3	6.00	5.60	6.49	5.60	**7.20**
	Creative Writing v3	**52.8**	10.3	31.5	42.8	28.4	43.6
	WritingBench	5.18	4.05	4.67	5.55	5.13	**6.54**
Math & Text Reasoning	MATH-500	46.4	67.6	55.0	67.2	55.2	**73.0**
	AIME'24	0.9	8.1	0.9	6.7	3.4	**13.4**
	AIME'25	0.8	5.3	0.4	4.2	2.6	**9.8**
	ZebraLogic	1.9	2.7	3.4	4.8	4.2	**12.8**
	AutoLogi	16.4	28.8	22.5	29.9	37.4	**59.8**
Agent & Coding	BFCL v3	16.3	31.3	47.8	50.4	44.1	**52.2**
	LiveCodeBench v5	1.8	10.4	5.3	9.2	3.6	**11.6**
Multilingual Tasks	Multi-IF	32.8	40.5	20.2	32.3	33.3	**44.7**
	INCLUDE	32.7	**43.8**	33.1	**43.8**	34.4	42.6
	MMMLU 14 languages	32.5	51.4	40.4	**51.8**	37.1	48.3
	MT-AIME2024	0.2	0.9	0.7	1.6	1.5	**4.9**
	PolyMath	3.5	6.7	5.0	7.3	4.6	**10.3**
	MLogiQA	31.8	39.5	40.9	39.5	37.3	**41.1**

1/10 activated parameters, demonstrating the effectiveness of our Strong-to-Weak Distillation approach in endowing lightweight models with profound reasoning capabilities.

(2) From Table 16, Qwen3-30B-A3B and Qwen3-14B (Non-thinking) surpass the non-reasoning baselines in most of the benchmarks. They exceed our previous Qwen2.5-32B-Instruct model with significantly fewer activated and total parameters, allowing for more efficient and cost-effective performance.

Qwen3-8B / 4B / 1.7B / 0.6B For Qwen3-8B and Qwen3-4B, we compare them with DeepSeek-R1-Distill-Qwen-14B and DeepSeek-R1-Distill-Qwen-32B in the thinking mode, and LLaMA-3.1-8B-Instruct (Dubey et al., 2024), Gemma-3-12B-IT (Team et al., 2025), Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct in the non-thinking mode, respectively. For Qwen3-1.7B and Qwen3-0.6B, we compare them with DeepSeek-R1-Distill-Qwen-1.5B and DeepSeek-R1-Distill-Llama-8B in the thinking mode, and Gemma-3-1B-IT, Phi-4-mini, Qwen2.5-1.5B-Instruct, and Qwen2.5-3B-Instruct in the non-thinking mode, respectively. We present the evaluation results of Qwen3-8B and Qwen3-4B in Table 17 and 18 and those of Qwen3-1.7B and Qwen3-0.6B in Table 19 and 20, respectively. Overall, these edge-side models exhibit impressive performance and outperform baselines even with more parameters, including our previous Qwen2.5 models, in either the thinking or the non-thinking mode. These results, once again, demonstrate the efficacy of our Strong-to-Weak Distillation approach, making it possible for us to build the lightweight Qwen3 models with remarkably reduced costs and efforts.

4.7 Discussion

The Effectiveness of Thinking Budget To verify that Qwen3 can enhance its intelligence level by leveraging an increased thinking budget, we adjust the allocated thinking budget on four benchmarks across Mathematics, Coding, and STEM domains. The resulting scaling curves are presented in Figure 2, Qwen3 demonstrates scalable and smooth performance improvements correlated to the allocated thinking budget. Moreover, we observe that if we further extend the output length beyond 32K, the model's performance is expected to improve further in the future. We leave this exploration as future work.

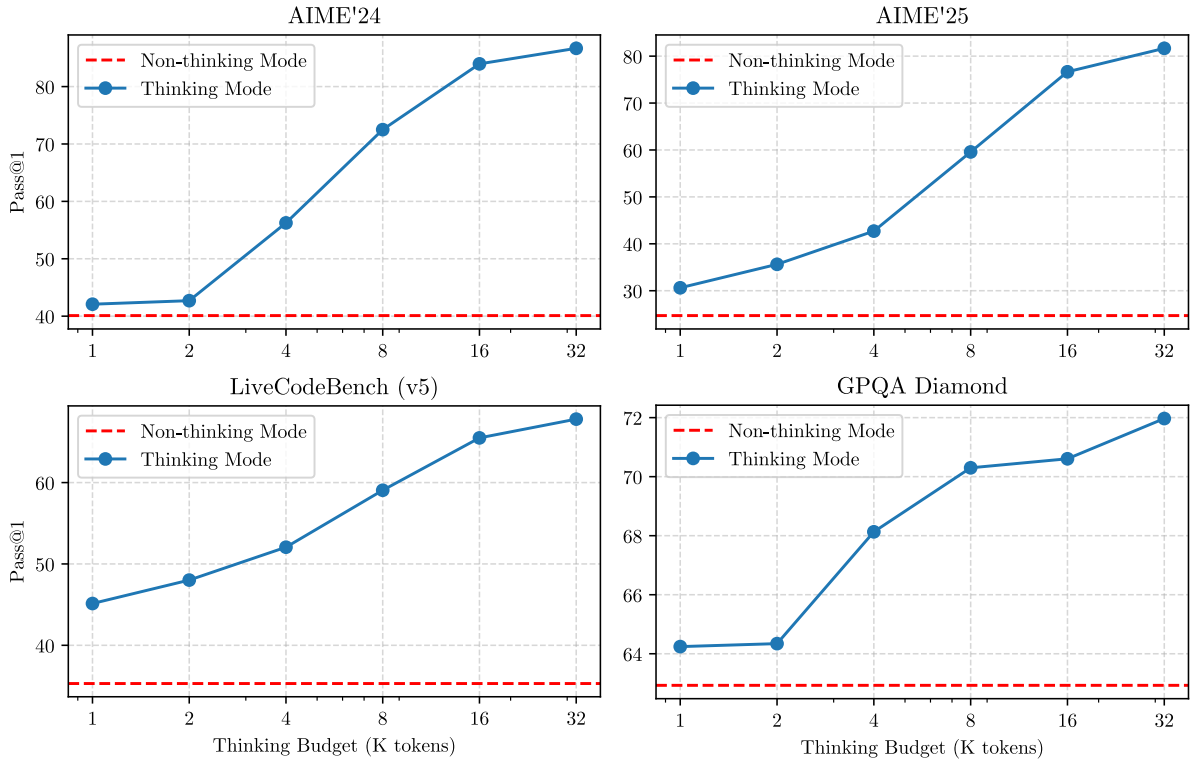

Figure 2: Performance of Qwen3-235B-A22B with respect to the thinking budget.

The Effectiveness and Efficiency of On-Policy Distillation We evaluate the effectiveness and efficiency of on-policy distillation by comparing the performance and computational cost—measured in GPU hours—after undergoing distillation versus direct reinforcement learning, both starting from the same off-policy distilled 8B checkpoint. For simplicity, we focus solely on math and code-related queries in

this comparison. The results, summarized in Table 21, show that distillation achieves significantly better performance than reinforcement learning while requiring approximately only 1/10 of the GPU hours. Furthermore, distillation from teacher logits enables the student model to expand its exploration space and enhance its reasoning potential, as evidenced by the improved pass@64 scores on the AIME'24 and AIME'25 benchmarks after distillation, compared to the initial checkpoint. In contrast, reinforcement learning does not lead to any improvement in pass@64 scores. These observations highlight the advantages of leveraging a stronger teacher model in guiding student model learning.

Table 21: Comparison of reinforcement learning and on-policy distillation on Qwen3-8B. Numbers in parentheses indicate pass@64 scores.

Method	AIME'24	AIME'25	MATH500	LiveCodeBench v5	MMLU -Redux	GPQA -Diamond	GPU Hours
Off-policy Distillation	55.0 (90.0)	42.8 (83.3)	92.4	42.0	86.4	55.6	-
+ Reinforcement Learning	67.6 (90.0)	55.5 (83.3)	94.8	52.9	86.9	61.3	17,920
+ On-policy Distillation	**74.4** (**93.3**)	**65.5** (**86.7**)	**97.0**	**60.3**	**88.3**	**63.3**	1,800

The Effects of Thinking Mode Fusion and General RL To evaluate the effectiveness of Thinking Mode Fusion and General Reinforcement Learning (RL) during the post-training, we conduct evaluations on various stages of the Qwen-32B model. In addition to the datasets mentioned earlier, we introduce several in-house benchmarks to monitor other capabilities. These benchmarks include:

- **CounterFactQA**: Contains counterfactual questions where the model needs to identify that the questions are not factual and avoid generating hallucinatory answers.
- **LengthCtrl**: Includes creative writing tasks with length requirements; the final score is based on the difference between the generated content length and the target length.
- **ThinkFollow**: Involves multi-turn dialogues with randomly inserted `/think` and `/no_think` flags to test whether the model can correctly switch thinking modes based on user queries.
- **ToolUse**: Evaluates the stability of the model in single-turn, multi-turn, and multi-step tool calling processes. The score includes accuracy in intent recognition, format accuracy, and parameter accuracy during the tool calling process.

Table 22: Performance of Qwen3-32B after Reasoning RL (Stage 2), Thinking Mode Fusion (Stage 3), and General RL (Stage 4). Benchmarks with * are in-house datasets.

		Stage 2 Reasoning RL	Stage 3 Thinking Mode Fusion		Stage 4 General RL	
	Benchmark	Thinking	Thinking	Non-Thinking	Thinking	Non-Thinking
General Tasks	LiveBench 2024-11-25	68.6	70.9+2.3	57.1	74.9+4.0	59.8+2.8
	Arena-Hard	86.8	89.4+2.6	88.5	93.8+4.4	92.8+4.3
	CounterFactQA*	50.4	61.3+10.9	64.3	68.1+6.8	66.4+2.1
Instruction & Format Following	IFEval strict prompt	73.0	78.4+5.4	78.4	85.0+6.6	83.2+4.8
	Multi-IF	61.4	64.6+3.2	65.2	73.0+8.4	70.7+5.5
	LengthCtrl*	62.6	70.6+8.0	84.9	73.5+2.9	87.3+2.4
	ThinkFollow*	-		88.7		98.9+10.2
Agent	BFCL v3	69.0	68.4-0.6	61.5	70.3+1.9	63.0+1.5
	ToolUse*	63.3	70.4+7.1	73.2	85.5+15.1	86.5+13.3
Knowledge & STEM	MMLU-Redux	91.4	91.0-0.4	86.7	90.9-0.1	85.7-1.0
	GPQA-Diamond	68.8	69.0+0.2	50.4	68.4-0.6	54.6+4.3
Math & Coding	AIME'24	83.8	81.9-1.9	28.5	81.4-0.5	31.0+2.5
	LiveCodeBench v5	68.4	67.2-1.2	31.1	65.7-1.5	31.3+0.2

The results are shown in Table 22, where we can draw the following conclusions:

(1) Stage 3 integrates the non-thinking mode into the model, which already possesses thinking capabilities after the first two stages of training. The ThinkFollow benchmark score of 88.7 indicates that the model has developed an initial ability to switch between modes, though it still occasionally makes errors. Stage 3 also enhances the model's general and instruction-following capabilities in thinking mode, with CounterFactQA improving by 10.9 points and LengthCtrl by 8.0 points.

(2) Stage 4 further strengthens the model's general, instruction-following, and agent capabilities in both thinking and non-thinking modes. Notably, the ThinkFollow score improves to 98.9, ensuring accurate mode switching.

(3) For Knowledge, STEM, Math, and Coding tasks, Thinking Mode Fusion and General RL do not bring significant improvements. In contrast, for challenging tasks like AIME'24 and LiveCodeBench, the performance in thinking mode actually decreases after these two training stages. We conjecture this degradation is due to the model being trained on a broader range of general tasks, which may compromise its specialized capabilities in handling complex problems. During the development of Qwen3, we choose to accept this performance trade-off to enhance the model's overall versatility.

5 Conclusion

In this technical report, we introduce Qwen3, the latest version of the Qwen series. Qwen3 features both thinking mode and non-thinking mode, allowing users to dynamically manage the number of tokens used for complex thinking tasks. The model was pre-trained on an extensive dataset containing 36 trillion tokens, enabling it to understand and generate text in 119 languages and dialects. Through a series of comprehensive evaluations, Qwen3 has shown strong performance across a range of standard benchmarks for both pre-trained and post-trained models, including tasks related to code generation, mathematics, reasoning, and agents.

In the near future, our research will focus on several key areas. We will continue to scale up pretraining by using data that is both higher in quality and more diverse in content. At the same time, we will work on improving model architecture and training methods for the purposes of effective compression, scaling to extremely long contexts, etc. In addition, we plan to increase computational resources for reinforcement learning, with a particular emphasis on agent-based RL systems that learn from environmental feedback. This will allow us to build agents capable of tackling complex tasks that require inference time scaling.

6 Authors

Core Contributors: An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, Zihan Qiu

Contributors: Bei Chen, Biao Sun, Bin Luo, Bin Zhang, Binghai Wang, Bowen Ping, Boyi Deng, Chang Si, Chaojie Yang, Chen Cheng, Chenfei Wu, Chengpeng Li, Chengyuan Li, Fan Hong, Guobin Zhao, Hang Zhang, Hangrui Hu, Hanyu Zhao, Hao Lin, Hao Xiang, Haoyan Huang, Hongkun Hao, Humen Zhong, Jialin Wang, Jiandong Jiang, Jianqiang Wan, Jianyuan Zeng, Jiawei Chen, Jie Zhang, Jin Xu, Jinkai Wang, Jinyang Zhang, Jinzheng He, Jun Tang, Kai Zhang, Ke Yi, Keming Lu, Keqin Chen, Langshi Chen, Le Jiang, Lei Zhang, Linjuan Wu, Man Yuan, Mingkun Yang, Minmin Sun, Mouxiang Chen, Na Ni, Nuo Chen, Peng Liu, Peng Wang, Peng Zhu, Pengcheng Zhang, Pengfei Wang, Qiaoyu Tang, Qing Fu, Qiuyue Wang, Rong Zhang, Rui Hu, Runji Lin, Shen Huang, Shuai Bai, Shutong Jiang, Sibo Song, Siqi Zhang, Song Chen, Tao He, Ting He, Tingfeng Hui, Wei Ding, Wei Liao, Wei Lin, Wei Zhang, Weijia Xu, Wenbin Ge, Wenmeng Zhou, Wenyuan Yu, Xianyan Jia, Xianzhong Shi, Xiaodong Deng, Xiaoming Huang, Xiaoyuan Li, Ximing Zhou, Xinyao Niu, Xipin Wei, Xuejing Liu, Yang Liu, Yang Yao, Yang Zhang, Yanpeng Li, Yantao Liu, Yidan Zhang, Yikai Zhu, Yiming Wang, Yiwen Hu, Yong Jiang, Yong Li, Yongan Yue, Yu Guan, Yuanzhi Zhu, Yunfei Chu, Yunlong Feng, Yuxin Zhou, Yuxuan Cai, Zeyao Ma, Zhaohai Li, Zheng Li, Zhengyang Tang, Zheren Fu, Zhi Li, Zhibo Yang, Zhifang Guo, Zhipeng Zhang, Zhiying Xu, Zhiyu Yin, Zhongshen Zeng, Zile Qiao, Ziye Meng

A Appendix

A.1 Additional Evaluation Results

A.1.1 Long-Context Ability

Table 23: **Performance of Qwen3 Models on the RULER benchmark.**

	Model	RULER						
		Avg.	**4K**	**8K**	**16K**	**32K**	**64K**	**128K**
	Qwen2.5-7B-Instruct	85.4	96.7	95.1	93.7	89.4	82.3	55.1
	Qwen2.5-14B-Instruct	91.4	97.7	96.8	95.9	93.4	86.7	78.1
	Qwen2.5-32B-Instruct	92.9	96.9	97.1	95.5	95.5	90.3	82.0
	Qwen2.5-72B-Instruct	**95.1**	**97.7**	**97.2**	**97.7**	**96.5**	93.0	88.4
Non-Thinking Mode	**Qwen3-4B**	85.2	95.1	93.6	91.0	87.8	77.8	66.0
	Qwen3-8B	89.1	96.3	96.0	91.8	91.2	82.1	77.4
	Qwen3-14B	94.6	98.0	97.8	96.4	96.1	94.0	85.1
	Qwen3-32B	93.7	98.4	96.0	96.2	94.4	91.8	85.6
	Qwen3-30B-A3B	91.6	96.5	97.0	95.3	92.4	89.1	79.2
	Qwen3-235B-A22B	95.0	**97.7**	**97.2**	96.4	95.1	**93.3**	**90.6**
Thinking Mode	**Qwen3-4B**	83.5	92.7	88.7	86.5	83.2	83.0	67.2
	Qwen3-8B	84.4	94.7	94.4	86.1	80.8	78.3	72.0
	Qwen3-14B	90.1	95.4	93.6	89.8	91.9	90.6	79.0
	Qwen3-32B	91.0	94.7	93.7	91.6	92.5	90.0	83.5
	Qwen3-30B-A3B	86.6	94.1	92.7	89.0	86.6	82.1	75.0
	Qwen3-235B-A22B	92.2	95.1	94.8	93.0	92.3	92.0	86.0

For evaluating long-context processing capabilities, we report the results on the RULER benchmark (Hsieh et al., 2024) in Table 23. To enable length extrapolation, we utilize YARN (Peng et al., 2023) with a `scaling_factor=4`. In thinking mode, we set the thinking budget to 8192 tokens to mitigate overly verbose reasoning on the extremely long inputs.

The results show that:

1. In non-thinking mode, Qwen3 outperforms Qwen2.5 models of a similar size in long-context processing tasks.
2. In thinking mode, the model's performance slightly degrades. We hypothesize that the thinking content does not provide significant benefits for these retrieval tasks, which do not rely on reasoning and may instead interfere with the retrieval process. We are committed to enhancing the long-context capability in the thinking mode in future versions.

A.1.2 Multilingual Ability

Table 24-35 presents the detailed benchmark scores across various languages, including Spanish, French, Portuguese, Italian, Arabic, Japanese, Korean, Indonesian, Russian, Vietnamese, German, and Thai. The results of these tables demonstrate that the Qwen3 series models achieve competitive performance across all evaluated benchmarks, showcasing their strong multilingual capabilities.

To evaluate the performance of Qwen3 across a broader range of languages, we utilize Belebele (Bandarkar et al., 2023), a benchmark for natural language understanding. We conduct evaluations on 80 supported languages from the benchmark, excluding 42 unoptimized languages, as shown in Table 36 (organized by language family). The performance comparison between Qwen3 and other baseline models on the Belebele benchmark is presented in Table 37. The results show that Qwen3 achieves comparable performance to similarly-sized Gemma models while outperforming Qwen2.5 significantly.

Table 24: **Benchmark scores for language: Spanish (es)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	Multi-IF	MLogiQA	INCLUDE	MMMLU	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	**80.1**	70.0	**96.4**	**88.7**	**90.0**	54.4	**79.9**
	QwQ-32B	70.0	75.0	81.8	84.5	76.7	52.2	73.4
	Qwen3-235B-A22B	74.2	**76.2**	89.1	86.7	86.7	**57.3**	78.4
	Qwen3-32B	74.7	68.8	90.9	82.8	76.7	51.8	74.3
	Qwen3-30B-A3B	74.9	71.2	80.0	81.9	76.7	48.5	72.2
	Qwen3-14B	76.2	67.5	83.6	81.1	73.3	50.3	72.0
	Qwen3-8B	74.1	70.0	78.2	79.2	70.0	43.7	69.2
	Qwen3-4B	69.1	68.8	72.7	75.7	66.7	42.3	65.9
	Qwen3-1.7B	56.0	55.0	72.7	64.5	46.7	30.2	54.2
	Qwen3-0.6B	39.2	42.5	54.5	48.8	13.3	14.3	35.4
Non-thinking	GPT-4o-2024-1120	67.5	52.5	**89.1**	80.6	10.0	15.5	52.5
	Gemma-3-27b-IT	73.5	57.5	**89.1**	77.7	30.0	22.4	58.4
	Qwen2.5-72B-Instruct	66.7	61.3	80.0	80.1	20.0	18.8	54.5
	Qwen3-235B-A22B	71.7	**66.2**	83.6	**83.7**	33.3	**29.5**	**61.3**
	Qwen3-32B	72.1	65.0	83.6	80.4	26.7	24.7	58.8
	Qwen3-30B-A3B	72.1	53.8	85.5	78.3	33.3	25.0	58.0
	Qwen3-14B	**76.2**	63.7	78.2	77.4	**40.0**	25.0	60.1
	Qwen3-8B	73.1	50.0	80.0	73.7	16.7	21.3	52.5
	Qwen3-4B	65.8	50.0	60.0	68.3	13.3	17.3	45.8
	Qwen3-1.7B	47.9	43.8	50.9	54.3	10.0	11.6	36.4
	Qwen3-0.6B	35.5	37.5	43.6	39.5	3.3	5.8	27.5

Table 25: **Benchmark scores for language: French (fr)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	Multi-IF	MLogiQA	INCLUDE	MMMLU	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	**80.5**	73.8	**85.7**	**88.3**	80.0	52.8	76.8
	QwQ-32B	72.4	78.8	76.2	84.0	80.0	49.4	73.5
	Qwen3-235B-A22B	77.3	78.8	**85.7**	86.6	**86.7**	**57.4**	**78.8**
	Qwen3-32B	76.7	**81.2**	76.2	82.1	83.3	47.1	74.4
	Qwen3-30B-A3B	75.2	67.5	83.3	81.0	76.7	46.9	71.8
	Qwen3-14B	77.6	71.2	73.8	80.4	73.3	44.2	70.1
	Qwen3-8B	73.8	66.2	**85.7**	77.9	70.0	45.3	69.8
	Qwen3-4B	71.3	63.7	71.4	74.5	66.7	40.2	64.6
	Qwen3-1.7B	52.6	56.2	54.8	64.8	60.0	28.7	52.8
	Qwen3-0.6B	36.1	48.8	47.6	48.4	6.7	14.0	33.6
Non-thinking	GPT-4o-2024-1120	67.8	56.2	85.7	**81.8**	10.0	15.3	52.8
	Gemma-3-27b-IT	73.9	57.5	73.8	78.3	23.3	21.5	54.7
	Qwen2.5-72B-Instruct	72.1	55.0	81.0	80.2	26.7	15.7	55.1
	Qwen3-235B-A22B	73.2	**65.0**	**88.1**	81.1	**36.7**	**28.1**	**62.0**
	Qwen3-32B	75.8	60.0	73.8	79.5	30.0	23.0	57.0
	Qwen3-30B-A3B	75.6	52.5	69.0	77.9	26.7	27.3	54.8
	Qwen3-14B	**78.4**	63.7	73.8	75.1	33.3	24.4	58.1
	Qwen3-8B	71.9	52.5	71.4	71.7	20.0	21.4	51.5
	Qwen3-4B	64.2	47.5	61.9	67.6	20.0	19.2	46.7
	Qwen3-1.7B	46.1	43.8	64.3	53.2	3.3	11.6	37.0
	Qwen3-0.6B	32.8	35.0	38.1	39.4	6.7	4.6	26.1

Table 26: **Benchmark scores for language: Portuguese (pt)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	Multi-IF	MLogiQA	INCLUDE	MMMLU	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	**80.5**	73.8	**83.9**	**88.9**	73.3	52.2	75.4
	QwQ-32B	70.5	70.0	80.4	84.0	80.0	48.7	72.3
	Qwen3-235B-A22B	73.6	**78.8**	78.6	86.2	**86.7**	**58.3**	**77.0**
	Qwen3-32B	74.1	76.2	76.8	82.6	80.0	52.4	73.7
	Qwen3-30B-A3B	76.1	71.2	71.4	81.0	76.7	49.3	71.0
	Qwen3-14B	77.3	68.8	75.0	81.6	83.3	46.7	72.1
	Qwen3-8B	73.9	67.5	75.0	78.6	56.7	44.8	66.1
	Qwen3-4B	70.6	62.5	71.4	75.1	73.3	44.2	66.2
	Qwen3-1.7B	55.6	60.0	53.6	64.6	46.7	28.2	51.4
	Qwen3-0.6B	38.7	33.8	42.9	47.5	10.0	12.7	30.9
Non-thinking	GPT-4o-2024-1120	66.8	57.5	78.6	80.7	10.0	15.0	51.4
	Gemma-3-27b-IT	72.9	55.0	75.0	77.1	**33.3**	20.9	55.7
	Qwen2.5-72B-Instruct	68.8	55.0	71.4	82.2	23.3	11.3	52.0
	Qwen3-235B-A22B	72.5	**67.5**	**82.1**	**83.5**	**33.3**	**28.3**	**61.2**
	Qwen3-32B	71.1	61.3	73.2	80.6	30.0	23.9	56.7
	Qwen3-30B-A3B	72.3	47.5	67.9	77.8	26.7	24.0	52.7
	Qwen3-14B	**75.5**	58.8	75.0	76.5	26.7	25.8	56.4
	Qwen3-8B	71.9	56.2	71.4	72.9	20.0	19.7	52.0
	Qwen3-4B	66.1	50.0	73.2	66.7	10.0	18.1	47.4
	Qwen3-1.7B	49.5	33.8	39.3	52.9	6.7	12.8	32.5
	Qwen3-0.6B	36.6	37.5	42.9	37.5	3.3	5.7	27.2

Table 27: **Benchmark scores for language: Italian (it)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	Multi-IF	INCLUDE	MMMLU	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	**80.9**	**100.0**	**87.2**	**90.0**	54.1	**82.4**
	QwQ-32B	71.2	96.4	84.9	76.7	49.3	75.7
	Qwen3-235B-A22B	73.7	96.4	85.7	80.0	**57.4**	78.6
	Qwen3-32B	76.6	90.9	81.6	80.0	49.7	75.8
	Qwen3-30B-A3B	75.9	94.5	81.9	80.0	48.1	76.1
	Qwen3-14B	79.0	94.5	80.2	70.0	47.0	74.1
	Qwen3-8B	74.6	89.1	77.5	76.7	46.1	72.8
	Qwen3-4B	69.8	83.6	74.4	76.7	44.5	69.8
	Qwen3-1.7B	54.6	74.5	64.2	53.3	29.6	55.2
	Qwen3-0.6B	37.8	45.5	45.9	6.7	13.3	29.8
Non-thinking	GPT-4o-2024-1120	67.6	**98.2**	80.7	13.3	15.2	55.0
	Gemma-3-27b-IT	74.6	90.9	78.4	23.3	20.5	57.5
	Qwen2.5-72B-Instruct	67.2	94.5	80.7	16.7	16.7	55.2
	Qwen3-235B-A22B	72.9	92.7	**82.6**	**33.3**	**28.6**	**62.0**
	Qwen3-32B	71.4	92.7	79.5	30.0	23.0	59.3
	Qwen3-30B-A3B	73.9	87.3	77.7	**33.3**	24.8	59.4
	Qwen3-14B	**75.8**	89.1	75.7	26.7	27.6	59.0
	Qwen3-8B	72.1	85.5	72.9	13.3	23.8	53.5
	Qwen3-4B	63.0	78.2	67.8	23.3	19.3	50.3
	Qwen3-1.7B	46.1	70.9	53.4	6.7	11.9	37.8
	Qwen3-0.6B	35.1	43.6	39.0	0.0	4.5	24.4

Table 28: **Benchmark scores for language: Arabic (ar)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	MLogiQA	INCLUDE	MMMLU	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	75.0	**89.3**	**87.8**	76.7	52.6	**76.3**
	QwQ-32B	75.0	67.9	81.8	80.0	41.3	69.2
	Qwen3-235B-A22B	**80.0**	71.4	83.6	76.7	**53.7**	73.1
	Qwen3-32B	66.2	73.2	80.1	**86.7**	47.0	70.6
	Qwen3-30B-A3B	66.2	66.1	77.2	83.3	47.3	68.0
	Qwen3-14B	71.2	67.9	77.4	83.3	46.6	69.3
	Qwen3-8B	65.0	67.9	74.4	76.7	44.9	65.8
	Qwen3-4B	62.5	55.4	67.7	66.7	41.2	58.7
	Qwen3-1.7B	55.0	44.6	53.2	36.7	25.8	43.1
	Qwen3-0.6B	40.0	41.1	38.9	10.0	11.7	28.3
Non-thinking	GPT-4o-2024-1120	51.2	**78.6**	**80.9**	13.3	12.9	47.4
	Gemma-3-27b-IT	56.2	62.5	74.4	26.7	22.8	48.5
	Qwen2.5-72B-Instruct	56.2	66.1	77.2	6.7	14.7	44.2
	Qwen3-235B-A22B	**66.2**	67.9	79.5	**40.0**	**28.2**	**56.4**
	Qwen3-32B	55.0	69.6	75.7	23.3	25.4	49.8
	Qwen3-30B-A3B	48.8	64.3	71.6	30.0	22.6	47.5
	Qwen3-14B	52.5	60.7	69.5	23.3	23.5	45.9
	Qwen3-8B	45.0	58.9	64.6	13.3	16.4	39.6
	Qwen3-4B	52.5	42.9	56.7	13.3	15.3	36.1
	Qwen3-1.7B	31.2	37.5	43.6	3.3	9.4	25.0
	Qwen3-0.6B	40.0	39.3	35.4	0.0	3.8	23.7

Table 29: **Benchmark scores for language: Japanese (ja)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	MLogiQA	INCLUDE	MMMLU	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	72.5	74.5	83.8	**83.3**	**55.4**	73.9
	QwQ-32B	73.8	86.3	82.3	53.3	39.9	67.1
	Qwen3-235B-A22B	**75.0**	**94.1**	**84.8**	73.3	52.7	**76.0**
	Qwen3-32B	70.0	90.2	80.2	76.7	47.7	73.0
	Qwen3-30B-A3B	66.2	88.2	79.9	73.3	47.4	71.0
	Qwen3-14B	68.8	88.2	79.4	66.7	45.7	69.8
	Qwen3-8B	71.2	86.3	74.9	73.3	44.7	70.1
	Qwen3-4B	63.7	80.4	72.5	53.3	40.7	62.1
	Qwen3-1.7B	53.8	74.5	61.8	36.7	28.5	51.1
	Qwen3-0.6B	47.5	47.1	45.1	13.3	14.5	33.5
Non-thinking	GPT-4o-2024-1120	60.0	92.2	**81.9**	10.0	12.5	51.3
	Gemma-3-27b-IT	66.2	86.3	76.5	20.0	17.3	53.3
	Qwen2.5-72B-Instruct	55.0	**94.1**	77.7	16.7	17.7	52.2
	Qwen3-235B-A22B	**67.5**	92.2	80.9	26.7	**26.9**	**58.8**
	Qwen3-32B	58.8	92.2	78.0	20.0	20.5	53.9
	Qwen3-30B-A3B	51.2	82.4	74.9	30.0	20.6	51.8
	Qwen3-14B	55.0	84.3	73.8	**33.3**	19.8	53.2
	Qwen3-8B	47.5	82.4	69.9	20.0	18.5	47.7
	Qwen3-4B	46.2	76.5	64.8	13.3	15.1	43.2
	Qwen3-1.7B	40.0	68.6	46.3	3.3	11.6	34.0
	Qwen3-0.6B	37.5	37.3	37.9	3.3	3.7	23.9

Table 30: **Benchmark scores for language: Korean (ko)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	MLogiQA	INCLUDE	MMMLU	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	75.0	**88.0**	**85.9**	76.7	50.0	**75.1**
	QwQ-32B	**76.2**	72.0	81.8	60.0	40.0	66.0
	Qwen3-235B-A22B	71.2	80.0	84.7	**80.0**	**55.7**	74.3
	Qwen3-32B	71.2	74.0	79.2	**80.0**	48.5	70.6
	Qwen3-30B-A3B	68.8	72.0	78.6	76.7	46.6	68.5
	Qwen3-14B	67.5	74.0	79.6	76.7	46.0	68.8
	Qwen3-8B	60.0	80.0	74.7	76.7	42.3	66.7
	Qwen3-4B	66.2	74.0	68.8	70.0	40.6	63.9
	Qwen3-1.7B	53.8	66.0	57.8	43.3	25.2	49.2
	Qwen3-0.6B	33.8	52.0	41.5	13.3	11.8	30.5
Non-thinking	GPT-4o-2024-1120	**63.7**	**80.0**	**80.5**	13.3	12.9	50.1
	Gemma-3-27b-IT	58.8	76.0	75.9	20.0	18.3	49.8
	Qwen2.5-72B-Instruct	58.8	68.0	76.7	6.7	17.7	45.6
	Qwen3-235B-A22B	**63.7**	76.0	79.8	**33.3**	**27.9**	**56.1**
	Qwen3-32B	60.0	74.0	77.2	26.7	21.2	51.8
	Qwen3-30B-A3B	52.5	72.0	72.5	16.7	20.7	46.9
	Qwen3-14B	52.5	68.0	73.3	20.0	18.7	46.5
	Qwen3-8B	52.5	76.0	66.5	23.3	16.3	46.9
	Qwen3-4B	46.2	74.0	59.9	13.3	16.6	42.0
	Qwen3-1.7B	48.8	58.0	46.0	6.7	9.0	33.7
	Qwen3-0.6B	40.0	52.0	36.9	0.0	5.5	26.9

Table 31: **Benchmark scores for language: Indonesian (id)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	INCLUDE	MMMLU	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	80.0	86.3	**83.3**	51.3	**75.2**
	QwQ-32B	76.4	83.7	73.3	47.3	70.2
	Qwen3-235B-A22B	80.0	**87.2**	80.0	**53.5**	**75.2**
	Qwen3-32B	80.0	82.0	76.7	45.6	71.1
	Qwen3-30B-A3B	**81.8**	80.4	80.0	44.9	71.8
	Qwen3-14B	78.2	79.6	70.0	45.3	68.3
	Qwen3-8B	72.7	77.7	70.0	43.8	66.0
	Qwen3-4B	70.9	72.3	66.7	41.2	62.8
	Qwen3-1.7B	63.6	61.2	36.7	26.8	47.1
	Qwen3-0.6B	36.4	46.6	10.0	12.6	26.4
Non-thinking	GPT-4o-2024-1120	80.0	81.1	10.0	14.7	46.4
	Gemma-3-27b-IT	76.4	75.9	13.3	22.6	47.0
	Qwen2.5-72B-Instruct	74.5	78.8	10.0	16.6	45.0
	Qwen3-235B-A22B	**81.8**	**81.9**	**33.3**	**27.5**	**56.1**
	Qwen3-32B	**81.8**	77.2	23.3	24.3	51.6
	Qwen3-30B-A3B	70.9	76.4	30.0	25.9	50.8
	Qwen3-14B	70.9	74.1	26.7	24.6	49.1
	Qwen3-8B	78.2	69.6	20.0	21.6	47.4
	Qwen3-4B	67.3	66.5	13.3	19.0	41.5
	Qwen3-1.7B	52.7	49.0	3.3	10.8	29.0
	Qwen3-0.6B	52.7	40.0	3.3	5.1	25.3

Table 32: **Benchmark scores for language: Russian (ru)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	Multi-IF	INCLUDE	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	**68.1**	**80.4**	70.0	52.3	67.7
	QwQ-32B	61.2	73.2	76.7	43.6	63.7
	Qwen3-235B-A22B	62.2	**80.4**	**80.0**	**53.1**	**68.9**
	Qwen3-32B	62.5	73.2	63.3	46.5	61.4
	Qwen3-30B-A3B	60.7	76.8	73.3	45.4	64.0
	Qwen3-14B	63.6	**80.4**	66.7	46.4	64.3
	Qwen3-8B	62.9	69.6	63.3	37.7	58.4
	Qwen3-4B	52.8	69.6	56.7	36.6	53.9
	Qwen3-1.7B	37.8	46.4	20.0	22.8	31.8
	Qwen3-0.6B	26.4	46.4	3.3	7.0	20.8
Non-thinking	GPT-4o-2024-1120	52.0	**80.4**	20.0	13.7	41.5
	Gemma-3-27b-IT	57.3	71.4	23.3	21.6	43.4
	Qwen2.5-72B-Instruct	54.1	67.9	20.0	13.3	38.8
	Qwen3-235B-A22B	56.7	75.0	**40.0**	**26.1**	**49.4**
	Qwen3-32B	58.6	71.4	30.0	23.3	45.8
	Qwen3-30B-A3B	58.0	73.2	30.0	21.1	45.6
	Qwen3-14B	**60.3**	71.4	26.7	24.2	45.6
	Qwen3-8B	59.3	58.9	20.0	22.8	40.2
	Qwen3-4B	46.1	58.9	13.3	17.8	34.0
	Qwen3-1.7B	34.8	41.1	3.3	13.2	23.1
	Qwen3-0.6B	25.5	46.4	0.0	5.8	19.4

Table 33: **Benchmark scores for language: Vietnamese (vi)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	MLogiQA	INCLUDE	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	72.5	**89.1**	70.0	52.1	70.9
	QwQ-32B	71.2	69.1	70.0	49.2	64.9
	Qwen3-235B-A22B	**75.0**	87.3	**83.3**	**55.1**	**75.2**
	Qwen3-32B	67.5	81.8	**83.3**	44.0	69.2
	Qwen3-30B-A3B	68.8	78.2	76.7	46.1	67.4
	Qwen3-14B	72.5	72.7	73.3	45.8	66.1
	Qwen3-8B	65.0	72.7	73.3	42.9	63.5
	Qwen3-4B	68.8	63.6	60.0	42.2	58.6
	Qwen3-1.7B	52.5	61.8	30.0	26.9	42.8
	Qwen3-0.6B	33.8	38.2	6.7	9.8	22.1
Non-thinking	GPT-4o-2024-1120	57.5	81.8	10.0	13.0	40.6
	Gemma-3-27b-IT	52.5	74.5	33.3	20.6	45.2
	Qwen2.5-72B-Instruct	61.3	72.7	26.7	18.6	44.8
	Qwen3-235B-A22B	**70.0**	**83.6**	**36.7**	**27.1**	**54.4**
	Qwen3-32B	60.0	81.8	23.3	21.8	46.7
	Qwen3-30B-A3B	52.5	81.8	20.0	24.7	44.8
	Qwen3-14B	63.7	67.3	20.0	21.6	43.2
	Qwen3-8B	48.8	65.5	20.0	19.1	38.4
	Qwen3-4B	48.8	65.5	20.0	19.0	38.3
	Qwen3-1.7B	36.2	60.0	3.3	10.9	27.6
	Qwen3-0.6B	30.0	36.4	3.3	3.9	18.4

Table 34: **Benchmark scores for language: German (de)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	INCLUDE	MMMLU	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	50.0	85.6	**86.7**	53.8	69.0
	QwQ-32B	57.1	83.8	76.7	51.0	67.2
	Qwen3-235B-A22B	**71.4**	**86.0**	83.3	**55.4**	**74.0**
	Qwen3-32B	64.3	81.9	**86.7**	48.1	70.2
	Qwen3-30B-A3B	64.3	81.9	80.0	46.6	68.2
	Qwen3-14B	57.1	80.9	70.0	48.1	64.0
	Qwen3-8B	64.3	78.1	66.7	43.6	63.2
	Qwen3-4B	57.1	74.0	73.3	43.1	61.9
	Qwen3-1.7B	64.3	63.4	36.7	26.8	47.8
	Qwen3-0.6B	57.1	47.6	10.0	13.7	32.1
Non-thinking	GPT-4o-2024-1120	57.1	80.4	10.0	13.5	40.2
	Gemma-3-27b-IT	57.1	76.1	26.7	20.2	45.0
	Qwen2.5-72B-Instruct	64.3	79.9	16.7	19.3	45.0
	Qwen3-235B-A22B	**71.4**	**81.7**	**40.0**	**25.9**	**54.8**
	Qwen3-32B	57.1	77.2	30.0	21.9	46.6
	Qwen3-30B-A3B	57.1	77.7	23.3	25.2	45.8
	Qwen3-14B	57.1	76.0	30.0	24.5	46.9
	Qwen3-8B	64.3	70.8	20.0	19.9	43.8
	Qwen3-4B	64.3	66.0	26.7	16.4	43.4
	Qwen3-1.7B	42.9	53.2	10.0	10.6	29.2
	Qwen3-0.6B	42.9	37.8	3.3	5.7	22.4

Table 35: **Benchmark scores for language: Thai (th)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	MLogiQA	MT-AIME24	PolyMath	Average
Thinking	Gemini2.5-Pro	73.8	80.0	50.7	68.2
	QwQ-32B	**75.0**	60.0	41.3	58.8
	Qwen3-235B-A22B	73.8	**86.7**	**53.6**	**71.4**
	Qwen3-32B	73.8	76.7	46.9	65.8
	Qwen3-30B-A3B	63.7	80.0	45.2	63.0
	Qwen3-14B	65.0	76.7	44.4	62.0
	Qwen3-8B	68.8	70.0	41.3	60.0
	Qwen3-4B	60.0	60.0	39.4	53.1
	Qwen3-1.7B	48.8	33.3	23.7	35.3
	Qwen3-0.6B	33.8	13.3	11.4	19.5
Non-thinking	GPT-4o-2024-1120	52.5	10.0	11.9	24.8
	Gemma-3-27b-IT	50.0	16.7	19.0	28.6
	Qwen2.5-72B-Instruct	58.8	6.7	17.4	27.6
	Qwen3-235B-A22B	**61.3**	23.3	**27.6**	**37.4**
	Qwen3-32B	**61.3**	13.3	22.2	32.3
	Qwen3-30B-A3B	50.0	**30.0**	22.3	34.1
	Qwen3-14B	47.5	23.3	22.1	31.0
	Qwen3-8B	42.5	10.0	17.2	23.2
	Qwen3-4B	43.8	13.3	16.1	24.4
	Qwen3-1.7B	42.5	6.7	9.5	19.6
	Qwen3-0.6B	37.5	0.0	3.6	13.7

Table 36: Language families and language codes supported by Qwen3 in Belebele Benchmark

Language family	# Langs	Language code (ISO 639-3_ISO 15924)
Indo-European	40	por_Latn, deu_Latn, tgk_Cyrl, ces_Latn, nob_Latn, dan_Latn, snd_Arab, spa_Latn, isl_Latn, slv_Latn, eng_Latn, ory_Orya, hrv_Latn, ell_Grek, ukr_Cyrl, pan_Guru, srp_Cyrl, npi_Deva, mkd_Cyrl, guj_Gujr, nld_Latn, swe_Latn, hin_Deva, rus_Cyrl, asm_Beng, cat_Latn, als_Latn, sin_Sinh, urd_Arab, mar_Deva, lit_Latn, slk_Latn, ita_Latn, pol_Latn, bul_Cyrl, afr_Latn, ron_Latn, fra_Latn, ben_Beng, hye_Armn
Sino-Tibetan	3	zho_Hans, mya_Mymr, zho_Hant
Afro-Asiatic	8	heb_Hebr, apc_Arab, acm_Arab, ary_Arab, ars_Arab, arb_Arab, mlt_Latn, erz_Arab
Austronesian	7	ilo_Latn, ceb_Latn, tgl_Latn, sun_Latn, jav_Latn, war_Latn, ind_Latn
Dravidian	4	mal_Mlym, kan_Knda, tel_Telu, tam_Taml
Turkic	4	kaz_Cyrl, azj_Latn, tur_Latn, uzn_Latn
Tai-Kadai	2	tha_Thai, lao_Laoo
Uralic	3	fin_Latn, hun_Latn, est_Latn
Austroasiatic	2	vie_Latn, khm_Khmr
Other	7	eus_Latn, kor_Hang, hat_Latn, swh_Latn, kea_Latn, jpn_Jpan, kat_Geor

Table 37: **Comparison of Belebele Benchmark performance between Qwen3 and other baseline models.** Scores are highlighted with the highest in **bold** and the second-best underlined.

Model	Indo-European	Sino-Tibetan	Afro-Asiatic	Austronesian	Dravidian	Turkic	Tai-Kadai	Uralic	Austroasiatic	Other
Gemma-3-27B-IT	89.2	86.3	**85.9**	84.1	83.5	86.8	81.0	91.0	86.5	**87.0**
Qwen2.5-32B-Instruct	85.5	82.3	80.4	70.6	67.8	80.8	74.5	87.0	79.0	72.6
QwQ-32B	86.1	83.7	81.9	71.3	69.3	80.3	77.0	88.0	83.0	74.0
Qwen3-32B (Thinking)	**90.7**	**89.7**	84.8	**86.7**	**84.5**	**89.3**	83.5	**91.3**	**88.0**	83.1
Qwen3-32B (Non-thinking)	89.1	88.0	82.3	83.7	84.0	85.0	**85.0**	88.7	**88.0**	81.3
Gemma-3-12B-IT	85.8	83.3	**83.4**	79.3	79.0	82.8	77.5	89.0	**83.0**	81.6
Qwen2.5-14B-Instruct	82.7	78.9	80.4	69.1	66.2	74.2	72.2	83.9	77.9	70.4
Qwen3-14B (Thinking)	**88.6**	**87.3**	82.4	**82.4**	**81.0**	**83.8**	**83.5**	**91.0**	82.5	**81.7**
Qwen3-14B (Non-thinking)	87.4	82.7	80.1	80.7	78.0	81.8	80.5	87.7	81.5	77.0
Gemma-3-4B-IT	71.8	72.0	63.5	61.7	64.8	64.0	61.5	70.7	71.0	62.6
Qwen2.5-3B-Instruct	58.0	62.3	57.2	47.9	36.9	45.1	49.8	50.6	56.8	48.4
Qwen3-4B (Thinking)	**82.2**	**77.7**	**74.1**	**73.0**	**74.3**	**76.3**	**68.5**	**83.0**	**74.5**	**67.9**
Qwen3-4B (Non-thinking)	76.0	77.0	65.6	65.6	65.5	64.0	60.5	74.0	74.0	61.0
Gemma-3-1B-IT	36.5	36.0	30.0	29.1	28.8	27.3	28.0	32.7	33.0	30.9
Qwen2.5-1.5B-Instruct	41.5	43.0	39.6	34.8	28.6	29.7	39.4	33.8	42.0	36.0
Qwen3-1.7B (Thinking)	**69.7**	**66.0**	**59.4**	**58.6**	**52.8**	**57.8**	**53.5**	**70.3**	**63.5**	**53.4**
Qwen3-1.7B (Non-thinking)	58.8	62.7	50.8	53.0	43.3	48.0	46.0	54.3	54.0	43.9

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

AIME. AIME problems and solutions, 2025. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query Transformer models from multi-head checkpoints. In *EMNLP*, pp. 4895–4901. Association for Computational Linguistics, 2023.

Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. Training-free long-context scaling of large language models. *CoRR*, abs/2402.17463, 2024.

Anthropic. Claude 3.7 Sonnet, 2025. URL https://www.anthropic.com/news/claude-3-7-sonnet.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *CoRR*, abs/2108.07732, 2021.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *CoRR*, abs/2309.16609, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The Belebele benchmark: A parallel reading comprehension dataset in 122 language variants. *CoRR*, abs/2308.16884, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q. Feldman, Arjun Guha, Michael Greenberg, and Abhinav Jangda. MultiPL-E: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Trans. Software Eng.*, 49(7):3675–3691, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. *CoRR*, abs/2401.06066, 2024.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 933–941. PMLR, 2017.

Google DeepMind. Gemini 2.5, 2025. URL https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7480–7512. PMLR, 2023.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The Llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. DoGE: Domain reweighting with generalization estimation. *arXiv preprint arXiv:2310.15393*, 2023.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with MMLU? *CoRR*, abs/2406.04127, 2024.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I. Wang. CRUXEval: A benchmark for code reasoning, understanding and execution. *arXiv preprint arXiv:2401.03065*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-IF: Benchmarking LLMs on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021b.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *CoRR*, abs/2404.06654, 2024.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *NeurIPS*, 2023.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2.5-Coder technical report. *CoRR*, abs/2409.12186, 2024.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. *CoRR*, abs/2403.07974, 2024.

Zixuan Jiang, Jiaqi Gu, Hanqing Zhu, and David Z. Pan. Pre-RMSNorm and Pre-CRMSNorm Transformers: Equivalent and efficient pre-LN Transformers. *CoRR*, abs/2305.14858, 2023.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *CoRR*, abs/2411.15124, 2024.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline. *CoRR*, abs/2406.11939, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *CoRR*, abs/2305.20050, 2023.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. ZebraLogic: On the scaling limits of LLMs for logical reasoning. *CoRR*, abs/2502.01100, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? Rigorous evaluation of large language models for code generation. In *NeurIPS*, 2023a.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. RegMix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*, 2024b.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. AlignBench: Benchmarking Chinese alignment of large language models. *CoRR*, abs/2311.18743, 2023b.

Meta-AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

OpenAI. Hello GPT-4o, 2024. URL https://openai.com/index/hello-gpt-4o/.

OpenAI. Multilingual massive multitask language understanding, 2024. URL https://huggingface.co/datasets/openai/MMMLU.

OpenAI. Learning to reason with LLMs, 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

OpenAI. Introducing openai o3 and o4-mini, 2025. URL https://openai.com/index/introducing-o3-and-o4-mini/.

Samuel J. Paech. Creative writing v3, 2024. URL https://eqbench.com/creative_writing.html.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. *CoRR*, abs/2309.00071, 2023.

Zihan Qiu, Zeyu Huang, Bo Zheng, Kaiyue Wen, Zekun Wang, Rui Men, Ivan Titov, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Demons in the detail: On implementing load balancing loss for training specialized mixture-of-expert models. *CoRR*, abs/2501.11873, 2025.

Shanghaoran Quan, Jiaxi Yang, Bowen Yu, Bo Zheng, Dayiheng Liu, An Yang, Xuancheng Ren, Bofei Gao, Yibo Miao, Yunlong Feng, Zekun Wang, Jian Yang, Zeyu Cui, Yang Fan, Yichang Zhang, Binyuan Hui, and Junyang Lin. CodeElo: Benchmarking competition-level code generation of LLMs with human-comparable Elo ratings. *CoRR*, abs/2501.01257, 2025.

Qwen Team. QwQ: Reflect deeply on the boundaries of the unknown, November 2024. URL https://qwenlm.github.io/blog/qwq-32b-preview/.

Qwen Team. QwQ-32B: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *CoRR*, abs/2311.12022, 2023.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzeminski, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia Soltani Moakhar, Ran Tamir, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. INCLUDE: evaluating multilingual language understanding with regional knowledge. *CoRR*, abs/2411.19799, 2024.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL (1)*. The Association for Computer Linguistics, 2016.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *ICLR*. OpenReview.net, 2023.

Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. Linguistic generalizability of test-time scaling in mathematical reasoning. *CoRR*, abs/2502.17407, 2025.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced Transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *ACL (Findings)*, pp. 13003–13051. Association for Computational Linguistics, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. Neural machine translation with byte-level subwords. In *AAAI*, pp. 9154–9160. AAAI Press, 2020.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. PolyMath: Evaluating mathematical reasoning in multilingual contexts, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *CoRR*, abs/2406.01574, 2024.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-free LLM benchmark. *CoRR*, abs/2406.19314, 2024.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. WritingBench: A comprehensive benchmark for generative writing. *CoRR*, abs/2503.05244, 2025.

xAI. Grok 3 beta — the age of reasoning agents, 2025. URL https://x.ai/news/grok-3.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818, 2023.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. *CoRR*, abs/2309.16039, 2023.

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Berkeley function calling leaderboard. https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *CoRR*, abs/2407.10671, 2024a.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024b.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *CoRR*, abs/2409.12122, 2024c.

Yidan Zhang, Boyi Deng, Yu Wan, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, and Jingren Zhou. P-MMEval: A parallel multilingual multitask benchmark for consistent evaluation of LLMs. *CoRR*, abs/2411.09116, 2024.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911, 2023.

Qin Zhu, Fei Huang, Runyu Peng, Keming Lu, Bowen Yu, Qinyuan Cheng, Xipeng Qiu, Xuanjing Huang, and Junyang Lin. AutoLogi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models. *CoRR*, abs/2502.16906, 2025.