本文由 AINLP 公众号整理翻译，更多 LLM 资源请扫码关注!

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.

Qwen3 技术报告

Qwen 团队

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

摘要

在本工作中，我们推出了Qwen3，这是Qwen模型系列的最新版本。Qwen3包括一系列旨在提升性能、效率和多语言能力的大型语言模型（LLMs）。Qwen3系列涵盖了密集架构和专家混合（MoE）架构的模型，参数规模从0.6亿到2350亿不等。Qwen3的一个关键创新是将思考模式（用于复杂、多步骤推理）和非思考模式（用于快速、上下文驱动的响应）集成到一个统一框架中。这消除了在不同模型之间切换的需求——例如，优化聊天的模型（如GPT-4o）和专用推理模型（如QwQ-32B）——并实现根据用户查询或聊天模板进行动态模式切换。同时，Qwen3引入了思考预算机制，允许用户在推理过程中自适应分配计算资源，从而根据任务复杂度平衡延迟和性能。此外，借助旗舰模型的知识，我们显著减少了构建较小模型所需的计算资源，同时确保其具有高度竞争力的性能。实证评估显示，Qwen3在代码生成、数学推理、代理任务等多项任务中实现了最先进的结果，在与更大规模的MoE模型和专有模型的竞争中表现出色。与前身Qwen2.5相比，Qwen3将多语言支持从29种扩展到119种语言和方言，通过改进跨语言理解和生成能力，增强了全球的可访问性。为了促进可复现性和社区驱动的研发，所有Qwen3模型均以Apache 2.0许可证公开提供。

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

1 引言

人类长期以来一直以追求人工通用智能（AGI）或人工超级智能（ASI）为目标。近年来，大型基础模型的突破，例如 GPT-4o（OpenAI，2024）、Claude 3.7（Anthropic，2025）、Gemini 2.5（DeepMind，2025）、DeepSeek-V3（Liu 等，2024a）、Llama-4（Meta-AI，2025）以及 Qwen2.5（Yang 等，2024b），在实现这一目标方面取得了显著进展。这些模型在涵盖数万亿标记的庞大数据集上进行训练，涉及各种领域和任务，有效地将人类的知识和能力浓缩到其参数中。此外，近期在推理模型方面的进展，通过强化学习进行优化，凸显了基础模型提升推理时扩展能力和实现更高智能水平的潜力，例如 o3（OpenAI，2025）、DeepSeek-R1（Guo 等，2025）。虽然大多数最先进的模型仍然是专有的，但开源社区的快速发展大大缩小了开源模型与闭源模型之间的性能差距。值得注意的是，越来越多的顶级模型（Meta-AI，2025；Liu 等，2024a；Guo 等，2025；Yang 等，2024b）正被作为开源项目发布，促进了人工智能领域的更广泛研究与创新。

在本工作中，我们介绍了Qwen3，这是我们基础模型系列Qwen的最新系列。Qwen3是一组开源权重的大型语言模型（LLMs），在各种任务和领域中都实现了最先进的性能。我们发布了密集模型和专家混合（MoE）模型，参数数量从6亿到2350亿不等，以满足不同下游应用的需求。值得注意的是，旗舰模型Qwen3-235B-A22B是一个MoE模型，总参数数为2350亿，每个标记激活的参数为220亿。这一设计确保了高性能和高效推理。

Qwen3 引入了多项关键创新，以增强其功能和易用性。首先，它将思考模式和非思考模式两种不同的操作模式集成到一个模型中。这允许用户在这两种模式之间切换，而无需在不同模型之间切换，例如，从 Qwen2.5 切换到 QwQ（Qwen 团队，2024）。这种灵活性确保开发者和用户能够高效地根据具体任务调整模型的行为。此外，Qwen3 引入了思考预算，为用户提供了对模型在执行任务时所施加的推理努力水平的细粒度控制。这一能力对于优化计算资源和性能至关重要，可以根据实际应用中的不同复杂度调整模型的思考行为。此外，Qwen3 已在覆盖多达119种语言和方言的36万亿个标记上进行了预训练，有效提升了其多语言能力。这一扩展的语言支持增强了其在全球应用和国际场景中的部署潜力。这些创新共同确立了 Qwen3 作为一个前沿的开源大型语言模型系列，能够有效应对各个领域和多种语言中的复杂任务。

Qwen3的预训练过程使用了一个规模庞大的数据集，包含大约36万亿个标记，经过精心策划以确保语言和领域的多样性。为了高效扩展训练数据，我们采用多模态方法：Qwen2.5-VL（Bai等，2025）经过微调以提取大量PDF文档中的文本。我们还使用领域特定模型生成合成数据：用于数学内容的Qwen2.5-Math（Yang等，2024c）和用于代码相关数据的Qwen2.5-Coder（Hui等，2024）。预训练过程遵循三阶段策略。第一阶段，模型在大约30万亿个标记上训练，以建立坚实的通用知识基础。第二阶段，模型在知识密集型数据上进行进一步训练，以增强在科学、技术、工程和数学（STEM）以及编码等领域的推理能力。最后，在第三阶段，模型在长上下文数据上进行训练，将其最大上下文长度从{v*}增加到{v*}。

为了更好地使基础模型与人类偏好和下游应用保持一致，我们采用多阶段后训练方法，既支持思考（推理）模式，也支持非思考模式。在前两个阶段，我们专注于通过长链式思考（CoT）冷启动微调和以数学与编码任务为重点的强化学习，培养强大的推理能力。在最后两个阶段，我们将带有和不带有推理路径的数据合并成一个统一的数据集进行进一步微调，使模型能够有效处理这两种类型的输入，然后应用通用领域的强化学习，以提升在各种下游任务中的表现。对于较小的模型，我们采用强到弱的蒸馏方法，利用来自较大模型的离策略和在策略知识转移，增强其能力。从先进教师模型的蒸馏在性能和训练效率方面都显著优于强化学习。

我们在涵盖多个任务和领域的全面基准测试中，评估了我们模型的预训练版本和后训练版本。实验结果显示，我们的基础预训练模型达到了最先进的性能。无论是在思考模式还是非思考模式下，后训练模型在与领先的专有模型和大型混合专家（MoE）模型（如 o1、o3-mini 和 DeepSeek-V3）竞争中表现出色。值得注意的是，我们的模型在编码、数学和代理相关任务中表现尤为出色。例如，旗舰模型 Qwen3-235B-A22B 在 AIME’24 上的得分达到了 85.7。

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

在 AIME’25（AIME，2025）中获得81.5，在 LiveCodeBench v5（Jain 等，2024）中获得70.7，在 CodeForces 上获得2,056，在 BFCL v3（Yan 等，2024）中获得70.8。此外，Qwen3 系列中的其他模型也显示出相对于其规模的强劲性能。此外，我们观察到，增加思考令牌的思考预算会在各种任务中持续改善模型的性能。

在接下来的章节中，我们将描述模型架构的设计，提供其训练过程的详细信息，展示预训练和后训练模型的实验结果，最后通过总结关键发现并概述未来研究的潜在方向，结束本技术报告。

2 架构

Qwen3 系列包括 6 个密集模型，分别是 Qwen3-0.6B、Qwen3-1.7B、Qwen3-4B、Qwen3-8B、Qwen3-14B 和 Qwen3-32B，以及 2 个 MoE 模型，Qwen3-30B-A3B 和 Qwen3-235B-A22B。旗舰模型 Qwen3-235B-A22B 拥有总共 {v*} 参数，其中激活参数为 {v*}。下面，我们将详细介绍 Qwen3 模型的架构。

Qwen3 密集模型的架构与 Qwen2.5 (Yang et al., 2024b) 类似，包括使用分组查询注意力（GQA，Ainslie et al., 2023）、SwiGLU（Dauphin et al., 2017）、旋转位置嵌入（RoPE，Su et al., 2024）以及采用预归一化的 RMSNorm（Jiang et al., 2023）。此外，我们移除了在 Qwen2 中使用的 QKV 偏置（Yang et al., 2024a），并引入 QK-Norm（Dehghani et al., 2023）到注意力机制中，以确保 Qwen3 的训练稳定性。关于模型架构的关键信息详见表 1。

Qwen3 MoE模型与Qwen3稠密模型共享相同的基本架构。关于模型架构的关键信息见表2。我们遵循Qwen2.5-MoE（Yang等人，2024b）并实现了细粒度的专家划分（Dai等人，2024）。Qwen3 MoE模型共有128个专家，每个Token激活8个专家。与Qwen2.5-MoE不同，Qwen3-MoE设计不包括共享专家。此外，我们采用全局批次负载均衡损失（Qiu等人，2025）以促进专家专业化。这些架构和训练创新在下游任务中显著提升了模型性能。

Qwen3 模型采用 Qwen 的分词器（Bai 等人，2023），该分词器实现了字节级字节对编码（BBPE，Brown 等人，2020；Wang 等人，2020；Sennrich 等人，2016），词汇表大小为 151,669。

表1：Qwen3 密集模型的架构。

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

表2：Qwen3 MoE模型的架构。

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 预训练

在本节中，我们将描述预训练数据的构建过程、预训练方法的细节，并展示在标准基准上评估基础模型的实验结果。

3.1 预训练数据

与Qwen2.5（Yang等，2024b）相比，我们在训练数据的规模和多样性方面有了显著扩展。具体而言，我们收集的预训练标记数量是原来的两倍——涵盖的语言数量是原来的三倍。所有Qwen3模型都在一个庞大且多样化的数据集上进行训练，该数据集包含119种语言和方言，总计36万亿个标记。该数据集包括各种高质量内容

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).

诸如编码、STEM（科学、技术、工程和数学）、推理任务、书籍、多语种文本和合成数据等领域。

为了进一步扩展预训练数据语料库，我们首先使用 Qwen2.5-VL 模型（Bai 等，2025）对大量类似 PDF 的文档进行文本识别。然后，使用 Qwen2.5 模型（Yang 等，2024b）对识别出的文本进行优化，以提高其质量。通过这两个步骤，我们能够获得额外的一组高质量文本标记，总量达到数万亿。此外，我们还采用 Qwen2.5（Yang 等，2024b）、Qwen2.5-Math（Yang 等，2024c）和 Qwen2.5-Coder（Hui 等，2024）模型，合成数万亿的不同格式的文本标记，包括教科书、问答、指令和代码片段，涵盖数十个领域。最后，我们通过引入更多多语言数据和增加更多语言，进一步扩展预训练语料库。与 Qwen2.5 使用的预训练数据相比，支持的语言数量已从 29 大幅增加到 119，增强了模型的语言覆盖范围和跨语言能力。

我们开发了一个多语言数据标注系统，旨在提升训练数据的质量和多样性。该系统已应用于我们的规模化预训练数据集，标注了超过30万亿个标记，涵盖教育价值、领域、行业和安全等多个维度。这些详细的标注支持更有效的数据筛选和组合。与之前的研究（Xie 等，2023；Fan 等，2023；Liu 等，2024b）在数据源或领域层面优化数据混合的方法不同，我们的方法通过在具有细粒度数据标签的小型代理模型上进行大量消融实验，在实例层面优化数据混合。

3.2 预训练阶段

Qwen3 模型通过三个阶段的预训练过程：

(1) 通用阶段 (S1)：在第一个预训练阶段，所有Qwen3模型使用序列长度为4,096个标记，在超过30万亿个标记上进行训练。在此阶段，模型已全面预训练掌握语言能力和通用世界知识，训练数据涵盖119种语言和方言。(2) 推理阶段 (S2)：为了进一步提升推理能力，我们通过增加STEM、编码、推理和合成数据的比例，优化了本阶段的预训练语料库。模型在序列长度为4,096个标记的基础上，使用大约5万亿个高质量标记进行进一步预训练。在此阶段，我们还加快了学习率的衰减速度。

(3) 长上下文阶段：在最后的预训练阶段，我们收集高质量的长上下文语料库，以扩展Qwen3模型的上下文长度。所有模型都在数百亿个标记上进行预训练，序列长度为32,768个标记。长上下文语料库包括75%的文本长度在16,384到32,768之间，25%的文本长度在4,096到16,384之间。沿袭Qwen2.5（Yang等，2024b），我们使用ABF技术（Xiong等，2023）将RoPE的基础频率从10,000提高到1,000,000。同时，我们引入YARN（Peng等，2023）和Dual Chunk Attention（DCA，An等，2024），在推理过程中实现序列长度能力的四倍提升。

类似于 Qwen2.5（Yang 等人，2024b），我们基于上述提到的三个预训练阶段，开发了用于预测最优超参数（例如学习率调度器和批量大小）的缩放规律。通过大量实验，我们系统地研究了模型架构、训练数据、训练阶段与最优训练超参数之间的关系。最后，我们为每个密集模型或 MoE 模型设定了预测的最优学习率和批量大小策略。

3.3 预训练评估

我们对Qwen3系列的基础语言模型进行了全面评估。基础模型的评估主要关注其在常识、推理、数学、科学知识、编码和多语言能力方面的表现。预训练基础模型的评估数据集包括15个基准测试：

- 一般任务：MMLU（Hendrycks 等，2021a）（5-shot），MMLU-Pro（Wang 等，2024）（5-shot，CoT），MMLU-redux（Gema 等，2024）（5-shot），BBH（Suzgun 等，2023）（3-shot，CoT），SuperGPQA（Du 等，2025）（5-shot，CoT）。
- 数学与STEM任务：GPQA（Rein 等，2023）（5次示范，链式推理），GSM8K（Cobbe 等，2021）（4次示范，链式推理），MATH（Hendrycks 等，2021b）（4次示范，链式推理）。

- **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;

- 编码任务：EvalPlus（Liu 等，2023a）（零样本）（HumanEval（Chen 等，2021）、MBPP（Austin 等，2021）、Humaneval+、MBPP+的平均值）（Liu 等，2023a），MultiPL-E（Cassano 等，2023）（零样本）（Python、C++、JAVA、PHP、TypeScript、C#、Bash、JavaScript），MBPP-3shot（Austin 等，2021），CRUX-O 来自 CRUXEval（1-shot）（Gu 等，2024）。
- 多语言任务：MGSM（Shi 等，2023）（8次示范，链式推理），MMMLU（OpenAI，2024）（5次示范），INCLUDE（Romanou 等，2024）（5次示范）。

对于基础模型基线，我们将Qwen3系列基础模型与Qwen2.5基础模型（Yang et al., 2024b）以及其他领先的开源基础模型进行比较，包括DeepSeek-V3 Base（Liu et al., 2024a）、Gemma-3（Team et al., 2025）、Llama-3（Dubey et al., 2024）和Llama-4（Meta-AI, 2025）系列基础模型，主要比较参数规模。所有模型均使用相同的评估流程和广泛使用的评估设置进行评估，以确保公平的比较。

基于整体评估结果，我们总结了Qwen3基础模型的一些关键结论。

(1) 与之前开源的SOTA密集和MoE基础模型（如DeepSeek-V3 Base、Llama-4-Maverick Base和Qwen2.5-72B-Base）相比，Qwen3-235B-A22B-Base在大多数任务中表现优于这些模型，参数总数或激活参数显著更少。 (2) 对于Qwen3 MoE基础模型，我们的实验结果表明： (a) 使用相同的预训练数据，Qwen3 MoE基础模型仅用1/5的激活参数即可达到与Qwen3密集基础模型相似的性能。 (b) 由于Qwen3 MoE架构的改进、训练Token的规模扩大以及更先进的训练策略，Qwen3 MoE基础模型可以在激活参数少于1/2且总参数更少的情况下，优于Qwen2.5 MoE基础模型。 (c) 即使激活参数只有Qwen2.5密集基础模型的1/10，Qwen3 MoE基础模型也能实现相当的性能，这为推理和训练成本带来了显著优势。

(3) Qwen3 密集基础模型的整体性能在较高参数规模下与 Qwen2.5 基础模型相当。例如，Qwen3-1.7B/4B/8B/14B/32B-Base 在性能上分别与 Qwen2.5-3B/7B/14B/32B/72B-Base 相当。特别是在 STEM、编码和推理基准测试中，Qwen3 密集基础模型的表现甚至超过了在更高参数规模下的 Qwen2.5 基础模型。

详细结果如下。

Qwen3-235B-A22B-Base 我们将Qwen3-235B-A22B-Base与我们之前的类似规模的MoE Qwen2.5-Plus-Base（Yang et al., 2024b）以及其他领先的开源基础模型进行比较：Llama-4-Maverick（Meta-AI，2025）、Qwen2.5-72B-Base（Yang et al., 2024b）、DeepSeek-V3 Base（Liu et al., 2024a）。从表3中的结果来看，Qwen3-235B-A22B-Base模型在大多数评估基准中都取得了最高的性能分数。我们还将Qwen3-235B-A22B-Base与其他基线模型进行单独比较，以进行详细分析。

(1) 与最近开源的模型 Llama-4-Maverick-Base（参数大约是其的两倍）相比，Qwen3-235B-A22B-Base 在大多数基准测试中仍然表现更优。 (2) 与之前的最先进开源模型 DeepSeek-V3-Base 相比，Qwen3-235B-A22B-Base 在 15 个评估基准中有 14 个优于 DeepSeek-V3-Base，参数总数只有其的约 1/3，激活参数为其的 2/3，充分展示了我们模型的强大和性价比。 (3) 与我们之前规模相似的 MoE Qwen2.5-Plus 相比，Qwen3-235B-A22B-Base 在参数和激活参数更少的情况下，显著优于它，显示出 Qwen3 在预训练数据、训练策略和模型架构方面的显著优势。 (4) 与我们之前的旗舰开源稠密模型 Qwen2.5-72B-Base 相比，Qwen3-235B-A22B-Base 在所有基准测试中都优于后者，且激活参数少于其的 1/3。同时，由于模型架构的优势，Qwen3-235B-A22B-Base 在每万亿 tokens 的推理成本和训练成本远低于 Qwen2.5-72B-Base。

Qwen3-32B-Base Qwen3-32B-Base 是我们在 Qwen3 系列中最大的密集模型。我们将其与类似规模的基线模型进行比较，包括 Gemma-3-27B（Team 等，2025）和 Qwen2.5-32B（Yang 等，2024b）。此外，我们还引入了两个强大的基线：最近开源的 MoE 模型 Llama- 4-Scout，它的参数数量是 {v*} 的三倍，但激活参数只有一半；

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

表3：Qwen3-235B-A22B-Base 与其他具有代表性的强大开源基线的比较。最高分和第二高分分别用加粗和下划线标出。

	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

表4：Qwen3-32B-Base 与其他强大的开源基线的比较。最高分和第二高分分别用加粗和下划线标出。

	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

表5：Qwen3-14B-Base、Qwen3-30B-A3B-Base 与其他强大的开源基线的比较。最高分和第二高分分别用加粗和下划线显示。

	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**

表6：Qwen8B-Base与其他强大开源基线的比较。最高分和第二高分分别用加粗和下划线标出。

	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**

表7：Qwen3-4B-Base 与其他强大的开源基线的比较。最高和第二高的得分分别用加粗和下划线标出。

	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**

表8：Qwen3-1.7B-Base、Qwen3-0.6B-Base 与其他强大的开源基线的比较。最高分和第二高分分别用加粗和下划线标出。

	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

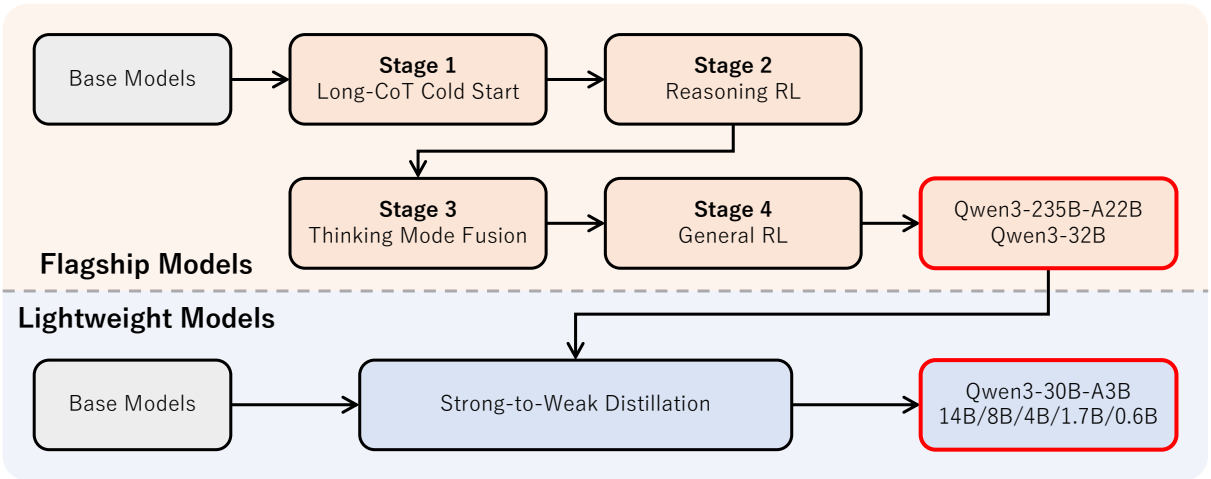

Figure 1: Post-training pipeline of the Qwen3 series models.

以及我们之前的旗舰开源密集模型Qwen2.5-72B-Base，其参数数量是Qwen3-32B-Base的两倍多。结果显示在表4中，支持三个关键结论：

(1) 与同等规模的模型相比，Qwen3-32B-Base在大多数基准测试中优于Qwen2.5-32B-Base和Gemma-3-27B Base。值得注意的是，Qwen3-32B-Base在MMLU-Pro上达到65.54，在SuperGPQA上达到39.78，显著优于其前身Qwen2.5-32B-Base。此外，Qwen3-32B-Base在编码基准测试中的得分明显高于所有基线模型。(2)令人惊讶的是，我们发现Qwen3-32B-Base在与Qwen2.5-72B-Base的比较中表现出竞争力。虽然Qwen3-32B-Base的参数数量不到Qwen2.5-72B-Base的一半，但在15个评估基准中的10个中优于Qwen2.5-72B-Base。在编码、数学和推理基准测试中，Qwen3-32B-Base具有显著优势。(3)与Llama-4-Scout-Base相比，Qwen3-32B-Base在所有15个基准测试中都显著优于它，参数数量只有Llama-4-Scout-Base的三分之一，但激活参数数量是其两倍。

Qwen3-14B-Base 和 Qwen3-30B-A3B-Base 的评估与类似规模的基线进行比较，包括 Gemma-3-12B Base 和 Qwen2.5-14B Base。同样，我们还引入了两个强基线：(1) Qwen2.5-Turbo（Yang 等人，2024b），它具有 42B 参数和 6B 激活参数。注意，其激活参数是 Qwen3-30B-A3B-Base 的两倍。(2) Qwen2.5-32B-Base，其激活参数是 Qwen3-30B-A3B 的 11 倍，是 Qwen3-14B 的两倍多。结果显示在表 5 中，我们可以得出以下结论。

(1) 与同等规模的模型相比，Qwen3-14B-Base在所有15个基准测试中都明显优于Qwen2.5-14B-Base和Gemma-3-12B-Base。 (2) 同样，Qwen3-14B-Base在参数不到Qwen2.5-32B-Base一半的情况下，也取得了非常具有竞争力的结果。 (3) 仅激活了1/5的非嵌入参数，Qwen3-30B-A3B在所有任务中都显著优于Qwen2.5-14B-Base，并且性能与Qwen3-14B-Base和Qwen2.5-32B-Base相当，这为我们在推理和训练成本方面带来了显著优势。

Qwen3-8B / 4B / 1.7B / 0.6B-Base 对于边缘端模型，我们以尺寸相似的 Qwen2.5、Llama-3 和 Gemma-3 基础模型作为基准。结果可以在表6、表7和表8中看到。所有的 Qwen3 8B / 4B / 1.7B / 0.6B-Base 模型在几乎所有基准测试中都持续保持强劲的性能。值得注意的是，Qwen3-8B / 4B / 1.7B-Base 模型甚至在超过一半的基准测试中优于更大尺寸的 Qwen2.5-14B / 7B / 3B 基础模型，尤其是在 STEM 相关和编码基准测试中，反映出 Qwen3 模型的显著提升。

4 训练后

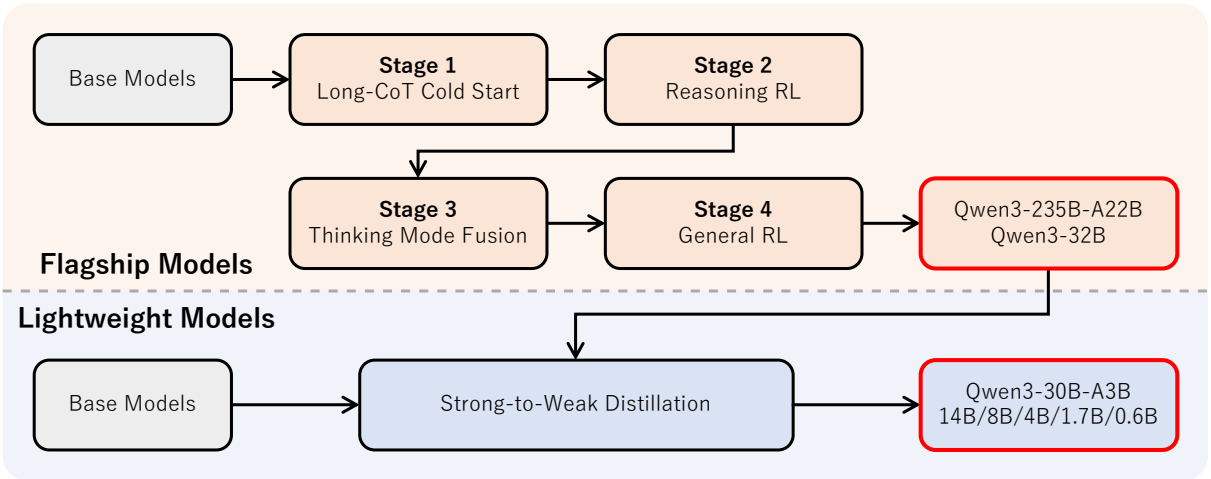

图 1：Qwen3 系列模型的训练后流程。

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

Qwen3的后训练流程策略性地设计了两个核心目标：

(1) 思维控制：这涉及到两种不同模式的整合，即“非思维”模式和“思维”模式，为用户提供选择模型是否进行推理的灵活性，并通过指定思维过程的令牌预算来控制思考的深度。

(2) 强到弱蒸馏：旨在简化和优化轻量级模型的后训练过程。通过利用大型模型的知识，我们大幅度降低了构建小规模模型所需的计算成本和开发工作量。

如图1所示，Qwen3系列的旗舰模型遵循一个复杂的四阶段训练过程。前两个阶段专注于培养模型的“思考”能力。接下来的两个阶段旨在将强大的“非思考”功能整合到模型中。

初步实验表明，将教师模型的输出logits直接蒸馏到轻量级学生模型中，可以有效提升它们的性能，同时保持对推理过程的细粒度控制。这种方法省去了为每个小规模模型单独进行耗时的四阶段训练过程的必要性。它带来了更好的即时性能，如更高的Pass@1分数所示，同时也增强了模型的探索能力，反映在Pass@64结果的提升。此外，它在训练效率方面也取得了显著优势，仅需四阶段训练方法的1/10 GPU时间。

在以下部分，我们将介绍四个阶段的训练过程，并详细说明强到弱蒸馏方法。

4.1 长链长尾冷启动

我们首先整理一个涵盖广泛类别的全面数据集，包括数学、代码、逻辑推理和一般STEM问题。数据集中的每个问题都配有经过验证的参考答案或基于代码的测试用例。该数据集作为长链式思维（long-CoT）训练“冷启动”阶段的基础。

数据集的构建涉及一个严格的两阶段过滤过程：查询过滤和响应过滤。在查询过滤阶段，我们使用 Qwen2.5-72B-Instruct 来识别并删除那些不易验证的查询。这包括包含多个子问题的查询或那些要求进行一般文本生成的查询。此外，我们还排除那些 Qwen2.5-72B-Instruct 在不使用 CoT 推理的情况下可以正确回答的查询。这有助于防止模型依赖表面猜测，确保只包含需要更深层次推理的复杂问题。此外，我们还使用 Qwen2.5-72B-Instruct 对每个查询的领域进行标注，以保持数据集中各领域的平衡代表性。

在保留验证查询集后，我们使用QwQ-32B（Qwen团队，2025）为每个剩余查询生成*N*个候选响应。当QwQ-32B持续未能生成正确的解决方案时，人工标注员会手动评估响应的准确性。对于具有正向Pass@*N*的查询，会应用更严格的筛选标准，以剔除以下响应：(1) 产生错误最终答案的，(2) 包含大量重复的，(3) 明显表明猜测而缺乏充分推理的，(4) 思考内容与总结内容不一致的，(5) 涉及不当的语言混用或风格转变的，或(6) 被怀疑与潜在验证集项目过于相似的。随后，经过精心筛选的细化数据集子集将用于推理模式的初始冷启动训练。此阶段的目标是为模型灌输基础推理模式，而不过分强调即时推理性能。这种方法确保模型的潜力不受限制，从而在随后的强化学习（RL）阶段获得更大的灵活性和改进。为了有效实现这一目标，最好在此准备阶段尽量减少训练样本数量和训练步骤。

4.2 推理强化学习

在推理强化学习阶段使用的查询-验证器对必须满足以下四个标准：(1) 它们在冷启动阶段未被使用。 (2) 它们对冷启动模型是可学习的。 (3) 它们尽可能具有挑战性。 (4) 它们涵盖了广泛的子领域。我们最终收集了总共 3,995 对查询-验证器对，并采用 GRPO (Shao 等人，2024) 来更新模型参数。我们观察到，使用较大的批量大小和每个查询的较多回合数，以及通过离策略训练以提高样本效率，对训练过程是有益的。我们还探讨了如何通过控制模型的熵，使其稳步增加或保持，从而平衡探索与利用。

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.

表9：思考模式融合阶段中思考和非思考模式的SFT数据示例。对于思考模式，可以省略/think标志，因为它代表默认行为。该功能已在由Hugging Face的分词器支持的聊天模板[1]中实现，其中可以通过额外参数enable_thinking=False禁用思考模式。

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|>

稳定，这对于保持稳定的训练至关重要。因此，我们在整个单次强化学习（RL）过程中，在训练奖励和验证性能方面都实现了持续的改进，而无需手动调整超参数。例如，Qwen3-235B-A22B模型的AIME’24得分在总共170个RL训练步骤中，从70.1提升到85.1。

4.3 思维模式融合

Thinking Mode Fusion 阶段的目标是将“非思考”能力整合到之前开发的“思考”模型中。这种方法允许开发者管理和控制推理行为，同时也降低了为思考和非思考任务部署单独模型的成本和复杂性。为此，我们对 Reasoning RL 模型进行持续的有监督微调（SFT），并设计了一个聊天模板以融合这两种模式。此外，我们发现能够熟练处理两种模式的模型在不同的思考预算下表现始终如一。

构建 SFT 数据。SFT 数据集结合了“思考”与“非思考”数据。为了确保 Stage 2 模型的性能不因额外的 SFT 而受到影响，“思考”数据通过在 Stage 1 查询上使用 Stage 2 模型进行拒绝采样生成。另一方面，“非思考”数据经过精心筛选，涵盖了多样的任务，包括编码、数学、指令执行、多语言任务、创意写作、问答以及角色扮演。此外，我们还采用自动生成的检查清单来评估“非思考”数据的响应质量。为了提升在低资源语言任务上的表现，我们特别增加了翻译任务的比例。

聊天模板设计。为了更好地整合两种模式并使用户能够动态切换模型的思考过程，我们为Qwen3设计了聊天模板，如表9所示。具体而言，对于思考模式和非思考模式的样本，分别在用户查询或系统消息中引入/think和/no_think标志。这允许模型根据用户的输入选择合适的思考模式。对于非思考模式的样本，我们在助手的回复中保留一个空的思考块。这一设计确保了模型内部格式的一致性，并允许开发者通过在聊天模板中拼接空的思考块，防止模型进行思考行为。默认情况下，模型在思考模式下运行；因此，我们添加了一些思考模式的训练样本，其中用户查询不包含/think标志。对于更复杂的多轮对话，我们会随机在用户查询中插入多个/think和/no_think标志，模型的回复将遵循最后遇到的标志。

思考预算。思维模式融合的另一个优势是，一旦模型学会在非思考和思考模式下响应，它自然会发展出处理中间情况的能力——基于不完整的思考生成响应。这一能力为实现对模型思考过程的预算控制奠定了基础。具体来说，当模型的思考长度达到用户定义的阈值时，我们会手动停止思考过程，并插入停止思考指令：“Considering the limited time by the user, I have to give the solution based on the thinking directly now.\n</think>.\n\n”。在插入此指令后，模型会根据其到目前为止的累积推理生成最终响应。值得注意的是，这种能力并非通过明确训练获得，而是作为应用思维模式融合的自然结果而出现的。

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

4.4 一般强化学习

通用强化学习阶段旨在广泛提升模型在各种场景下的能力和稳定性。为此，我们建立了一个复杂的奖励系统，涵盖超过20个不同任务，每个任务都具有定制的评分标准。这些任务特别针对以下核心能力的提升：

- 指令遵循：此能力确保模型能够准确理解并遵循用户指令，包括与内容、格式、长度以及结构化输出相关的要求，提供符合用户期望的回应。
- 格式如下：除了明确的指令外，我们还期望模型遵循特定的格式规范。例如，它应当根据 /think 和 / no_think 标志，适当地在思考模式和非思考模式之间切换，并在最终输出中始终使用指定的标记（例如 <think> 和 </think>）来分隔思考部分和回答部分。
- 偏好对齐：对于开放式问题，偏好对齐旨在提升模型的有用性、参与度和风格，最终提供更自然、更令人满意的用户体验。
- 代理能力：这涉及训练模型通过指定接口正确调用工具。在强化学习滚动过程中，模型被允许执行完整的多轮交互循环，并获得真实环境的执行反馈，从而提高其在长远决策任务中的表现和稳定性。
- 专门场景的能力：在更专业的场景中，我们设计了针对特定背景的任务。例如，在检索增强生成（RAG）任务中，我们引入奖励信号以引导模型生成准确且符合上下文的响应，从而最大程度地减少幻觉的风险。

为了对上述任务提供反馈，我们使用了三种不同类型的奖励：

(1) 基于规则的奖励：基于规则的奖励在推理强化学习阶段被广泛使用，也适用于诸如指令执行（Lambert 等人，2024）和格式遵循等一般任务。设计良好的基于规则的奖励可以高精度地评估模型输出的正确性，防止奖励黑客等问题。(2) 具有参考答案的基于模型的奖励：在这种方法中，我们为每个查询提供一个参考答案，并提示 Qwen2.5-72B-Instruct 根据此参考答案对模型的响应进行评分。这种方法可以更灵活地处理多样化的任务，无需严格的格式要求，避免了纯规则奖励可能导致的假阴性。(3) 无参考答案的基于模型的奖励：利用人类偏好数据，我们训练一个奖励模型，为模型响应分配标量分数。这种方法不依赖于参考答案，能够处理更广泛的查询，同时有效提升模型的参与度和帮助性。

4.5 强到弱蒸馏

强到弱蒸馏流程专为优化轻量级模型而设计，涵盖5个密集模型（Qwen3-0.6B、1.7B、4B、8B 和 14B）以及一个 MoE 模型（Qwen3-30B-A3B）。该方法提升了模型性能，同时有效赋予强大的模式切换能力。蒸馏过程分为两个主要阶段：

(1) 离策略蒸馏：在这个初始阶段，我们结合使用 /think 和 /no_think 模式生成的教师模型输出进行响应蒸馏。这有助于轻量级学生模型发展基本的推理能力以及在不同思维模式之间切换的能力，为下一阶段的在策略训练奠定坚实的基础。(2) 在策略蒸馏：在这个阶段，学生模型生成在策略序列以进行微调。具体来说，采样提示，学生模型在 /think 或 /no_think 模式下产生响应。然后，通过将其 logits 与教师模型（Qwen3-32B 或 Qwen3-235B-A22B）的 logits 对齐，最小化 KL 散度，进行微调。

4.6 训练后评估

为了全面评估指令调优模型的质量，我们采用了自动基准测试，以评估模型在思考和非思考模式下的性能。这些基准测试是

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.

表10：多语言基准测试及所包含的语言。语言以IETF语言标签标识。

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

分类为几个维度：

- 一般任务：我们使用包括 MMLU-Redux（Gema 等，2024）、GPQA-Diamond（Rein 等，2023）、C-Eval（Huang 等，2023）和 LiveBench（2024-11-25）（White 等，2024）在内的基准测试。对于 GPQA-Diamond，我们对每个查询采样 10 次，并报告平均准确率。
- 对齐任务：为了评估模型与人类偏好的匹配程度，我们采用一系列专业的基准测试。在指令遵循性能方面，我们报告 IFEval (Zhou et al., 2023) 的严格提示准确率。为了评估模型在一般话题上的偏好与人类的对齐情况，我们使用 Arena-Hard (Li et al., 2024) 和 AlignBench v1.1 (Liu et al., 2023b)。在写作任务方面，我们依赖 Creative Writing V3 (Paech, 2024) 和 WritingBench (Wu et al., 2025) 来评估模型的熟练度和创造力。
- 数学与文本推理：为了评估数学和逻辑推理能力，我们采用了高水平的数学基准，包括 MATH-500（Lightman 等，2023）、AIME’24 和 AIME’25（AIME，2025），以及文本推理任务，包括 ZebraLogic（Lin 等，2025）和 AutoLogi（Zhu 等，2025）。对于 AIME 问题，每年的题目包括第 I 部分和第 II 部分，共计 30 个问题。对于每个问题，我们采样 64 次，并取平均准确率作为最终得分。
- 代理与编码：为了测试模型在编码和基于代理的任务中的熟练程度，我们使用 BFCL v3（Yan 等，2024）、LiveCodeBench（v5，2024.10-2025.02）（Jain 等，2024）以及来自 CodeElo（Quan 等，2025）的 Codeforces 评级。对于 BFCL，所有 Qwen3 模型均使用 FC 格式进行评估，并使用 yarn 将模型部署到 64k 的上下文长度以进行多轮评估。一些基线来自 BFCL 排行榜，取 FC 和 Prompt 格式中的较高分数。对于未在排行榜上报告的模型，则评估 Prompt 格式。对于 LiveCodeBench，在非思考模式下，我们使用官方推荐的提示语，而在思考模式下，我们调整提示模板，允许模型更自由地思考，去除限制 `You will not return anything except for the program`。为了评估模型与竞赛编程专家之间的性能差距，我们使用 CodeForces 计算 Elo 评级。在我们的基准测试中，每个问题最多通过生成八次独立推理尝试来解决。
- 多语言任务：为了实现多语言能力，我们评估了四类任务：指令遵循、知识、数学和逻辑推理。指令遵循通过 Multi-IF (He et al., 2024) 进行评估，重点关注8种关键语言。知识评估包括两种类型：通过 INCLUDE (Romanou et al., 2024) 评估的区域知识，涵盖44种语言，以及通过 MMMLU (OpenAI, 2024) 在14种语言中评估的通用知识，不包括未优化的Yoruba语言；对于这两个基准，我们仅抽取原始数据的10%以提高评估效率。数学任务采用 MT-AIME2024 (Son et al., 2025)，涵盖55种语言，以及 PolyMath (Wang et al., 2025)，包括18种语言。逻辑推理通过 MlogiQA 进行评估，涵盖10种语言，数据来源于 Zhang et al. (2024)。

对于所有处于思考模式的Qwen3模型，我们采用采样温度0.6，top-p值0.95，以及top-k值20。此外，对于Creative Writing v3和WritingBench，我们应用1.5的存在惩罚，以鼓励生成更多样化的内容。对于非思考模式的Qwen3模型，我们将采样超参数配置为温度= 0.7，top-p= 0.8，top-k= 20，以及存在惩罚= 1.5。无论是思考模式还是非思考模式，我们都将最大输出长度设置为32,768个标记，除了AIME’24和AIME’25，我们将此长度扩展到38,912个标记，以提供足够的思考空间。

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**

表11：Qwen3-235B-A22B（Thinking）与其他推理基线的比较。最高分和第二高分分别用加粗和下划线标出。

		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**

表12：Qwen3-235B-A22B（非思考）与其他非推理基线的比较。最高和第二高分数分别用加粗和下划线显示。

		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

评估结果总结 从评估结果来看，我们总结了最终版Qwen3模型的几个关键结论如下:

(1) 我们的旗舰模型Qwen3-235B-A22B在思考模式和非思考模式下都展现出开源模型中的最先进整体性能，超越了DeepSeek-R1和DeepSeek-V3等强大基线。Qwen3-235B-A22B在与闭源领先模型如OpenAI-o1、Gemini2.5-Pro和GPT-4o的竞争中也具有很强的竞争力，展示了其深厚的推理能力和全面的通用能力。 (2) 我们的旗舰密集模型Qwen3-32B在大多数基准测试中优于我们之前最强的推理模型QwQ-32B，并且与闭源的OpenAI-o3-mini表现相当，显示出其令人信服的推理能力。Qwen3-32B在非思考模式下也表现出色，超越了我们之前的旗舰非推理密集模型Qwen2.5-72B-Instruct。 (3) 我们的轻量级模型，包括Qwen3-30B-A3B、Qwen3-14B以及其他较小的密集模型，在参数数量相近或更多的情况下，性能始终优于开源模型，证明了我们的Strong-to-Weak蒸馏方法的成功。

详细结果如下。

Qwen3-235B-A22B 对于我们的旗舰模型 Qwen3-235B-A22B，我们将其与领先的推理和非推理模型进行比较。在思考模式下，我们以 OpenAI-o1（OpenAI，2024）、DeepSeek-R1（Guo 等，2025）、Grok-3-Beta（Think）（xAI，2025）和 Gemini2.5-Pro（DeepMind，2025）作为推理基线。在非思考模式下，我们采用 GPT-4o-2024-11-20（OpenAI，2024）、DeepSeek-V3（Liu 等，2024a）、Qwen2.5-72B-Instruct（Yang 等，2024b）和 LLaMA-4-Maverick（Meta-AI，2025）作为非推理基线。我们在表 11 和 12 中展示了评估结果。

（1）从表11可以看出，Qwen3-235B-A22B（Thinking）仅激活了60%的参数，总参数量为35%，在17/23个基准测试中优于DeepSeek-R1，特别是在需要推理的任务（例如数学、代理和编码）上，展示了Qwen3-235B-A22B在开源模型中的最先进推理能力。此外，Qwen3-235B-A22B（Thinking）在与闭源的OpenAI-o1、Grok-3-Beta（Think）和Gemini2.5-Pro等模型的竞争中也表现出极强的实力，显著缩小了开源与闭源模型在推理能力上的差距。

(2) 从表12来看，Qwen3-235B-A22B（非思考）超过了其他领先的开源模型，包括DeepSeek-V3、LLaMA-4-Maverick以及我们之前的旗舰模型Qwen2.5-72B-Instruct，并且在18/23个基准测试中也优于闭源的GPT-4o-2024-11-20，表明其即使未经过刻意思考过程增强，也具有固有的强大能力。

Qwen3-32B 对于我们的旗舰密集模型 Qwen3-32B，我们在思考模式下以 DeepSeek-R1-Distill-Llama-70B、OpenAI-o3-mini（中等）以及我们之前最强的推理模型 QwQ-32B（Qwen 团队，2025）作为基线。我们还在非思考模式下采用 GPT-4o-mini-2024-07-18、LLaMA-4-Scout 以及我们之前的旗舰模型 Qwen2.5-72B-Instruct 作为基线。我们在表 13 和表 14 中展示了评估结果。

(1) 从表13可以看出，Qwen3-32B（Thinking）在17/23个基准测试中优于QwQ-32B，成为在32B这个理想规模上的新一代最先进推理模型。此外，Qwen3-32B（Thinking）在对齐和多语言性能方面也优于闭源的OpenAI-o3-mini（medium），具有竞争力。 (2) 从表14可以看出，Qwen3-32B（Non-thinking）在几乎所有基准测试中都表现优于所有基线模型。特别是，Qwen3-32B（Non-thinking）在通用任务上的表现与Qwen2.5-72B-Instruct持平，在对齐、多语言和推理相关任务上具有显著优势，再次证明了Qwen3相较于我们之前的Qwen2.5系列模型的根本性提升。

Qwen3-30B-A3B 与 Qwen3-14B 对比，我们在思考模式下将它们与 DeepSeek-R1-Distill-Qwen-32B 和 QwQ-32B 进行比较，在非思考模式下则分别与 Phi-4（Abdin 等，2024）、Gemma-3-27B-IT（Team 等，2025）以及 Qwen2.5-32B-Instruct 进行比较。我们在表 15 和 16 中展示了评估结果。

(1) 从表15来看，Qwen3-30B-A3B 和 Qwen3-14B（Thinking）在与 QwQ-32B 的竞争中都表现出色，尤其是在与推理相关的基准测试中。值得注意的是，Qwen3-30B-A3B 在模型规模更小、参数更少的情况下，达到了与 QwQ-32B 相当的性能。

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**

表13：Qwen3-32B（Thinking）与其他推理基线的比较。最高分和第二高分分别用加粗和下划线标出。

		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**

表14：Qwen3-32B（非思考）与其他非推理基线的比较。最高分和第二高分分别用加粗和下划线标出。

		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

表15：Qwen3-30B-A3B / Qwen3-14B（Thinking）与其他推理基线的比较。最高分和第二高分分别用加粗和下划线标出。

		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

表16：Qwen3-30B-A3B / Qwen3-14B（非思考）与其他非推理基线的比较。最高和第二高分数分别用加粗和下划线标出。

		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

表17：Qwen3-8B / Qwen3-4B（Thinking）与其他推理基线的比较。最高分和第二高分分别用加粗和下划线标出。

		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**

表18：Qwen3-8B / Qwen3-4B（非思考）与其他非推理基线的比较。最高和第二高分数分别用加粗和下划线标出。

		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**

表19：Qwen3-1.7B / Qwen3-0.6B（Thinking）与其他推理基线的比较。最高分和第二高分分别用加粗和下划线标出。

		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**

表20：Qwen3-1.7B / Qwen3-0.6B（非思考）与其他非推理基线的比较。最高分和第二高分分别用加粗和下划线标出。

		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.

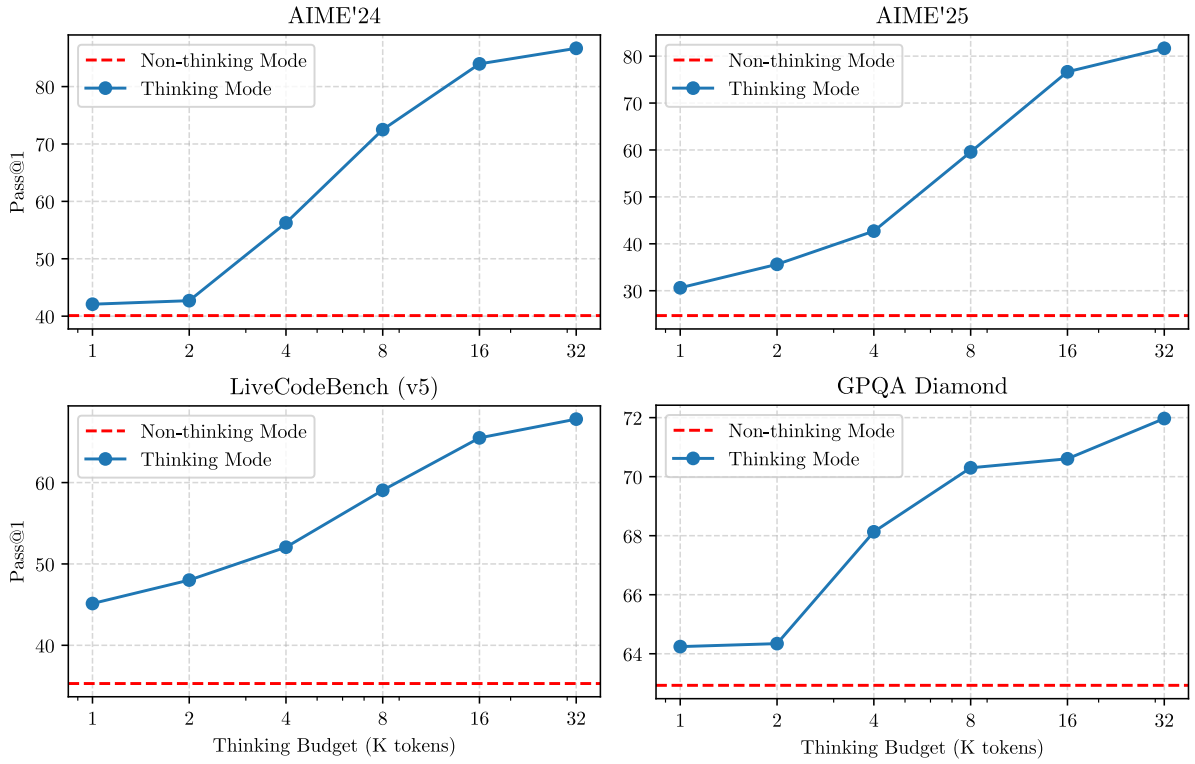

Figure 2: Performance of Qwen3-235B-A22B with respect to the thinking budget.

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

激活参数占比1/10，展示了我们强到弱蒸馏方法在赋予轻量级模型深刻推理能力方面的有效性。(2) 从表16来看，Qwen3-30B-A3B 和 Qwen3-14B（非思考）在大多数基准测试中超越了非推理基线。它们在激活参数和总参数明显更少的情况下，超越了我们之前的 Qwen2.5-32B-Instruct 模型，实现了更高效、更具成本效益的性能。

Qwen3-8B / 4B / 1.7B / 0.6B 对于Qwen3-8B和Qwen3-4B，我们在思考模式下将它们与DeepSeek-R1-Distill-Qwen-14B和DeepSeek-R1-Distill-Qwen-32B进行比较，在非思考模式下则分别与LLaMA-3.1-8B-Instruct（Dubey等，2024）、Gemma-3-12B-IT（Team等，2025）、Qwen2.5-7B-Instruct和Qwen2.5-14B-Instruct进行比较。对于Qwen3-1.7B和Qwen3-0.6B，我们在思考模式下将它们与DeepSeek-R1-Distill-Qwen-1.5B和DeepSeek-R1-Distill-Llama-8B进行比较，在非思考模式下则分别与Gemma-3-1B-IT、Phi-4-mini、Qwen2.5-1.5B-Instruct和Qwen2.5-3B-Instruct进行比较。我们在表17和18中展示了Qwen3-8B和Qwen3-4B的评估结果，表19和20中则是Qwen3-1.7B和Qwen3-0.6B的评估结果。总体而言，这些边缘端模型表现出令人印象深刻的性能，无论是在思考模式还是非思考模式下，都优于基线模型，包括我们之前的Qwen2.5系列模型，即使参数更多。这些结果再次证明了我们强到弱蒸馏方法的有效性，使我们能够以显著降低成本和努力的方式构建轻量级的Qwen3模型。

4.7 讨论

思维预算的有效性 为了验证Qwen3是否能够通过增加思维预算来提升其智能水平，我们在数学、编码和STEM领域的四个基准测试中调整了分配的思维预算。所得的缩放曲线如图2所示，Qwen3展现出与分配的思维预算相关的可扩展且平滑的性能提升。此外，我们观察到如果将输出长度进一步扩展到超过32K，模型的性能预计将在未来得到进一步提升。我们将这一探索留作未来的工作。

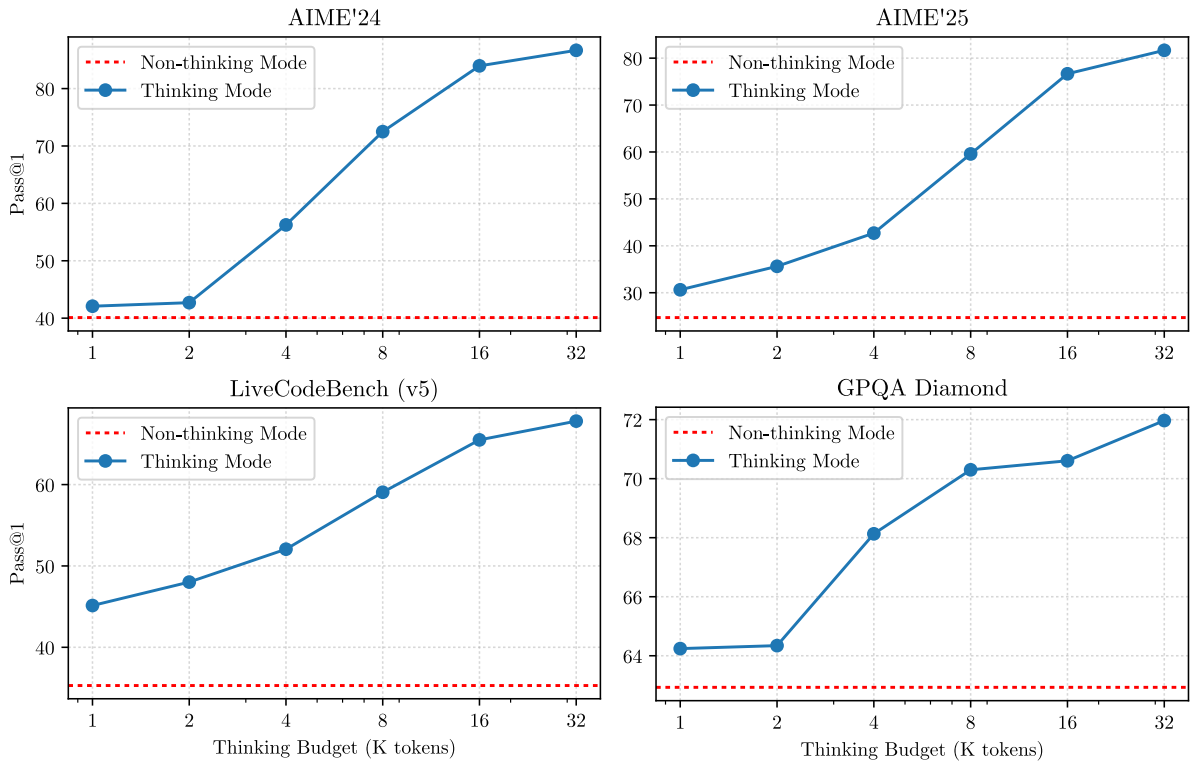

图2：Qwen3-235B-A22B在思考预算方面的性能。

在策略蒸馏的有效性和效率方面，我们通过比较经过蒸馏后与直接强化学习的性能和计算成本（以GPU小时为单位）来进行评估，二者都从相同的离策略蒸馏的8B检查点开始。为简便起见，我们仅关注数学和代码相关的查询。

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.

这个比较。结果总结在表21中，显示蒸馏在性能上明显优于强化学习，同时只需大约1/10的GPU时间。此外，从教师 logits 进行蒸馏使学生模型能够扩展其探索空间并增强其推理潜力，正如在蒸馏后在AIME’24和AIME’25基准测试中获得的pass@64分数的提升所证明，相较于初始检查点。相比之下，强化学习并未带来pass@64分数的任何提升。这些观察结果突显了利用更强的教师模型引导学生模型学习的优势。

表21：Qwen3-8B上强化学习与在策略蒸馏的比较。括号中的数字表示pass@64得分。

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

思维模式融合与通用强化学习的影响 为了评估思维模式融合和通用强化学习（RL）在后训练中的效果，我们对Qwen-32B模型的各个阶段进行了评估。除了前面提到的数据集外，我们还引入了几个内部基准测试，以监控其他能力。这些基准测试包括：

- CounterFactQA：包含反事实问题，模型需要识别这些问题不是真实的，并避免生成幻觉式的答案。
- LengthCtrl：包括具有长度要求的创意写作任务；最终得分基于生成内容长度与目标长度之间的差异。
- ThinkFollow：涉及多轮对话，随机插入`/think`和`/no_think`标志，以测试模型是否能根据用户查询正确切换思维模式。
- ToolUse：评估模型在单轮、多轮和多步工具调用过程中的稳定性。评分包括意图识别的准确性、格式准确性以及工具调用过程中参数的准确性。

表22：Qwen3-32B在推理RL（阶段2）、思维模式融合（阶段3）和通用RL（阶段4）后的性能。带*的基准测试为内部数据集。

		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

结果如表22所示，我们可以得出以下结论：

(1) 第三阶段将非思考模式整合到模型中，该模型在前两个阶段的训练后已具备思考能力。ThinkFollow基准得分为88.7，表明模型已初步具备切换模式的能力，尽管仍偶尔会出错。第三阶段还增强了模型在思考模式下的通用能力和指令遵循能力，CounterFactQA提高了10.9分，LengthCtrl提高了8.0分。

(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

(2) 第四阶段进一步增强了模型在思考和非思考模式下的通用、指令遵循和代理能力。值得注意的是，ThinkFollow 分数提升至 98.9，确保了模式切换的准确性。

(3) 对于知识、STEM、数学和编码任务，思维模式融合和通用强化学习并未带来显著的提升。相反，对于像 AIME’24 和 Live-CodeBench 这样具有挑战性的任务，经过这两个训练阶段后，思维模式的表现实际上有所下降。我们推测这种退化是由于模型在更广泛的通用任务上进行了训练，可能会影响其在处理复杂问题时的专业能力。在开发 Qwen3 时，我们选择接受这种性能的权衡，以提升模型的整体多样性。

5 结论

在本技术报告中，我们介绍了Qwen3，Qwen系列的最新版本。Qwen3具有思考模式和非思考模式，允许用户动态管理用于复杂思考任务的令牌数量。该模型在包含36万亿个令牌的庞大数据集上进行了预训练，使其能够理解和生成119种语言和方言的文本。通过一系列全面的评估，Qwen3在各种标准基准测试中表现出色，涵盖预训练模型和后训练模型，包括代码生成、数学、推理和代理等任务。

在不久的将来，我们的研究将集中在几个关键领域。我们将继续通过使用质量更高、内容更丰富的数据来扩大预训练的规模。同时，我们将致力于改进模型架构和训练方法，以实现有效的压缩、扩展到极长的上下文等目标。此外，我们计划增加强化学习的计算资源，特别强调基于代理的RL系统，从环境反馈中学习。这将使我们能够构建能够应对需要推理时间扩展的复杂任务的代理。

6 位作者

核心贡献者：安阳、李安峰、杨宝松、张贝辰、惠彬源、郑博、于博文、郜昌、黄成根、吕晨旭、郑楚杰、刘大恒、周凡、黄飞、胡峰、葛浩、韦浩然、林欢、唐佳龙、杨建、涂建宏、张建伟、杨建新、杨佳希、周静、周景仁、林俊阳、邓凯、鲍克勤、杨克新、于乐、邓良浩、李梅、薛明峰、李明泽、张佩、王鹏、朱秦、门锐、高清高、刘世轩、罗双、李天浩、唐天一、尹文彪、任兴章、王新宇、张新宇、任轩昌、范阳、苏阳、张义昌、张英儿、万瑜、刘玉琼、王泽坤、崔泽宇、张振儒、周志鹏、邱子涵

贡献者：陈贝，孙彪，罗斌，张斌，王炳海，平博文，邓博毅，司昌，杨超杰，程辰，吴辰菲，李成鹏，李成远，洪凡，赵国彬，张航，胡航瑞，赵汉宇，林浩，向浩，黄浩然，郝宏坤，中虎门，王佳林，江建东，万建强，曾建远，陈家伟，张杰，徐金，王金凯，张金凯，何金正，唐俊，张凯，易科，陆克明，陈克勤，陈晟世，吴乐，张磊，吴兰娟，袁满，杨明坤，孙敏敏，陈谋翔，倪娜，陈诺，刘鹏，王鹏，朱鹏，张鹏程，王鹏飞，唐巧瑜，傅青，王秋月，张荣，胡锐，林润基，黄深，白帅，江书通，宋思博，张思奇，陈松，何涛，贺婷，惠廷丰，丁伟，廖伟，林伟，张伟，徐维佳，葛文彬，周文萌，余文远，贾贤岩，史宪中，戴晓东，黄晓明，李晓远，周希明，牛新雅，魏西平，刘雪晶，刘洋，姚洋，张洋，李彦鹏，胡一文，姜永，李永，岳永安，关宇，朱远志，楚云飞，冯云龙，周宇新，蔡宇轩，马泽耀，李浩海，李正，唐正阳，傅哲仁，李志，杨志博，郭志鹏，徐志鹏，尹志勇，曾志深，乔子乐，孟子野

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.

A 附录

A.1 额外的评估结果

A.1.1 长上下文能力

表23：Qwen3模型在RULER基准测试中的性能。

	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

为了评估长上下文处理能力，我们在表23中报告了RULER基准（Hsieh等，2024）的结果。为了实现长度外推，我们使用了带有`scaling_factor=4`的YARN（Peng等，2023）。在思考模式下，我们将思考预算设置为8192个标记，以减轻在极长输入上的过度冗长推理。

结果显示：

1. 在非思考模式下，Qwen3 在长上下文处理任务中优于同等规模的 Qwen2.5 模型。2. 在思考模式下，模型的性能略有下降。我们假设，思考内容对这些不依赖推理的检索任务并没有显著的帮助，反而可能干扰检索过程。我们致力于在未来版本中增强思考模式下的长上下文能力。

A.1.2 多语言能力

表24-35展示了包括西班牙语、法语、葡萄牙语、意大利语、阿拉伯语、日语、韩语、印度尼西亚语、俄语、越南语、德语和泰语在内的各种语言的详细基准分数。这些表格的结果表明，Qwen3系列模型在所有评估基准中都表现出具有竞争力的性能，展示了它们强大的多语言能力。

为了评估Qwen3在更广泛的语言范围内的表现，我们采用Belebele（Bandarkar等，2023），这是一个自然语言理解的基准测试。我们对基准测试中支持的80种语言进行了评估，排除了42种未优化的语言，如表36所示（按语言家族分类）。Qwen3与其他基线模型在Belebele基准上的性能对比见表37。结果显示，Qwen3在性能上与同等规模的Gemma模型相当，同时显著优于Qwen2.5。

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

表24：语言：西班牙语（es）的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表25：语言：法语（fr）的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表26：语言：葡萄牙语（pt）的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表27：意大利语（it）语言的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表28：语言：阿拉伯语（ar）的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表29：语言：日语（ja）的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表30：语言：韩语（ko）的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表31：印尼语（id）的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表格 32：俄语（ru）语言的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表33：语言：越南语（vi）的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表 34：德语（de）语言的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表35：泰语（th）语言的基准分数。最高和第二高的分数分别用加粗和下划线显示。

	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

表36：Qwen3在Belebele基准测试中支持的语系和语言代码

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

表37：Qwen3与其他基线模型在Belebele基准测试中的性能比较。得分最高的用粗体标出，第二高的用下划线标出。

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.

参考文献

马拉·阿卜丁、乔蒂·阿内贾、哈基拉特·贝尔、Sébastien Bubeck、Ronen Eldan、Suriya Gunasekar、迈克尔·哈里森、拉塞尔·J·休伊特、莫贾恩·贾瓦赫里皮、皮耶罗·考夫曼等。Phi-4技术报告。*arXiv preprint arXiv:2412.08905*，2024年。

AIME。AIME题目与解答，2025年。网址 https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions。

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebr ón 和 Sumit Sanghai. GQA：从多头检查点训练通用多查询Transformer模型。在 *EMNLP*，第 4895–4901 页。计算语言学协会，2023。

陈新安，黄飞，张俊，龚善三，邱锡朋，周昌，孔灵鹏。无需训练的大规模语言模型长上下文扩展。*CoRR*，abs/2402.17463，2024。

人类学。克劳德 3.7 十四行诗，2025年。网址 https://www.anthropic.com/news/claude-3-7-sonnet。

雅各布·奥斯汀、奥古斯都·奥德纳、麦克斯韦·I·奈、马尔滕·博斯马、亨瑞克·米哈莱夫斯基、戴维·多汉、艾伦·江、凯瑞·J·蔡、迈克尔·特里、阮国·勒、查尔斯·萨顿。使用大型语言模型进行程序合成。*CoRR*，abs/2108.07732，2021。

白金泽，白帅，楚云飞，崔泽宇，邓凯，邓晓东，范阳，葛文彬，韩宇，黄飞，惠彬元，罗吉，李梅，林俊阳，林润基，刘大恒，刘高，卢成强，卢克明，马建新，门锐，任兴章，任轩成，谭传奇，谭思南，涂建宏，王鹏，王世杰，王伟，吴胜光，许本峰，徐金，杨安，杨浩，杨建，杨树生，杨阳，余博文，袁宏毅，袁正，张建伟，张兴轩，张宜昌，张振儒，周昌，周景仁，周晓欢，朱天航。Qwen技术报告。*CoRR*，abs/2309.16609，2023。

帅白，柯勤陈，雪晶刘，家林王，文彬葛，思博宋，凯当，鹏王，世杰王，俊唐等。Qwen2.5-VL技术报告。*arXiv preprint arXiv:2502.13923*，2025年。

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer 和 Madian Khabsa。Belebele 基准测试：一个包含 122 种语言变体的平行阅读理解数据集。*CoRR*，abs/2308.16884，2023。

汤姆 B. 布朗、本杰明·曼恩、尼克·莱德、梅拉妮·苏比亚、贾里德·卡普兰、普拉弗拉·达里沃尔、阿尔文德·尼拉坎坦、普拉纳夫·夏姆、吉里什·萨斯特里、阿曼达·阿斯克尔、桑迪尼·阿加瓦尔、阿里尔·赫伯特-沃斯、格雷琴·克鲁格、汤姆·亨尼根、雷文·查尔德、阿迪蒂亚·拉梅什、丹尼尔·M·齐格勒、杰弗里·吴、克莱门斯·温特、克里斯托弗·赫塞、马克·陈、埃里克·西格勒、马泰乌什·利特温、斯科特·格雷、本杰明·切斯、杰克·克拉克、克里斯托弗·伯纳、山姆·麦坎德利什、艾利克·拉德福德、伊利亚·苏茨克弗和达里奥·阿莫迪。语言模型是少-shot 学习者。在 *NeurIPS*，2020年。

费德里科·卡萨诺、约翰·高瓦尔、丹尼尔·阮、悉尼·阮、卢娜·菲普斯-科斯廷、唐纳德·平克尼、易明浩、紫阳天、卡罗琳·简·安德森、莫莉·Q·费尔德曼、阿尔俊·古哈、迈克尔·格林伯格和阿布希纳夫·江达。MultiPL-E：一种可扩展且多语种的神经代码生成基准方法。*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, 和 Wojciech Zaremba。评估在代码上训练的大型语言模型。*CoRR*，abs/2107.03374，2021。

卡尔·科布、维尼特·科萨拉朱、穆罕默德·巴维亚然、马克·陈、许宇俊、卢卡斯·凯撒、马蒂亚斯·普拉普特、杰瑞·特沃克、雅各布·希尔顿、中野礼一郎、克里斯托弗·赫塞和约翰·舒尔曼。训练验证器以解决数学应用题。*CoRR*，abs/2110.14168，2021。

戴岱迈、邓成奇、赵成刚、徐瑞祥、郝卓高、陈德利、李家石、曾旺丁、余兴凯、吴勇、谢振达、李志达、黄盼盼、罗福利、阮冲、隋志芳、梁文峰。DeepSeekMoE：迈向混合专家语言模型中的终极专家专业化。*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.

Yann N. Dauphin, Angela Fan, Michael Auli 和 David Grangier。使用门控卷积网络进行语言建模。在 *ICML*，*Proceedings of Machine Learning Research* 的第 70 卷，页码 933–941。PMLR，2017。

谷歌DeepMind。Gemini 2.5，2025年。网址 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 和 Neil Houlsby。将视觉变换器扩展到 220 亿参数。在 *ICML*，第 202 卷的 *Proceedings of Machine Learning Research*，第 7480–7512 页。PMLR，2023。

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, 等人。SuperGPQA：在285个研究生学科中扩展大规模语言模型评估。*arXiv preprint arXiv:2502.14739*，2025年。

阿比穆纽·杜贝，阿比纳夫·乔赫里，阿比纳夫·潘迪，阿比谢克·卡迪安，艾哈迈德·阿尔-达勒，艾莎·莱特曼，阿基尔·马图尔，艾伦·谢尔滕，艾米·杨，安吉拉·范，阿尼鲁德·戈亚尔，安东尼·哈特肖恩，奥博·杨，阿尔奇·米特拉，阿奇·斯拉万库马尔，阿尔捷姆·科雷涅夫，亚瑟·欣斯瓦克，阿伦· Rao，阿斯顿·张，奥尔埃利恩·罗德里格斯，奥斯汀·格雷格森，艾娃·斯帕塔鲁，巴蒂斯特·罗兹埃尔，贝瑟尼·比伦，平 Tang，鲍比·切恩，夏洛特·考歇托，查雅·纳雅克，克洛伊·毕，克里斯·马拉，克里斯·麦康奈尔，克里斯蒂安·凯勒，克里斯托弗·图雷，吴春阳，科琳·王，克里斯蒂安·坎顿·费雷尔，赛勒斯·尼古拉伊迪斯，达米安·阿隆西亚斯，丹尼尔·宋，丹妮尔·品茨，丹尼·利夫希茨，戴维·埃西奥布，德鲁夫·乔达里，德鲁夫·马哈詹，迭戈·加西亚-奥拉诺，迭戈·佩里诺，迪乌克·胡普克斯，埃戈尔·拉科姆金，伊哈布·阿尔巴达维，伊琳娜·洛巴诺娃，艾米丽·迪南，埃里克·迈克尔·史密斯，菲利普·拉德诺维奇，弗兰克·张，加布里埃尔·辛奈夫，嘉布丽尔·李，乔治亚·刘易斯·安德森，格雷姆·奈尔，格雷戈尔·米亚隆，庚·庞，吉列姆·库库雷尔，海莉·阮，汉娜·科雷瓦尔，胡旭，雨果·图弗龙，伊利扬·扎罗夫，伊马诺尔·阿里埃塔·伊巴拉，伊莎贝尔·M·克卢曼，伊山·米斯拉，伊万·埃夫季莫夫，杰德·科佩，李在元，简·格费特，雅娜·弗拉内斯，杰森·朴，杰·马哈迪奥卡，吉特·沙阿，耶尔梅尔·范德林德，詹妮弗·比洛克，珍妮·洪，珍雅·李，杰里米·傅，纪建峰，黄建宇，刘佳文，王杰，余洁，乔安娜·比顿，乔·斯皮萨克，张钟秀，Joseph Rocca，约书亚·约翰斯顿，约书亚·萨克斯，贾俊腾，卡利扬·瓦苏登·阿尔瓦拉，卡尔蒂克耶·乌帕西尼，凯特·普拉维亚克，李柯，肯尼斯·希菲尔德，凯文·斯通，等等。Llama 3模型群。*CoRR*，abs/2407.21783，2024年。

Simin Fan、Matteo Pagliardini 和 Martin Jaggi。DoGE：具有泛化估计的领域重加权。*arXiv preprint arXiv:2310.15393*，2023年。

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 等人。我们完成 MMLU 吗？*CoRR*，abs/2406.04127，2024。

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, 和 Sida I. Wang。CRUXEval：一个用于代码推理、理解和执行的基准测试。*arXiv preprint arXiv:2401.03065*，2024年。

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi 等。DeepSeek-R1：通过强化学习激励大规模语言模型的推理能力。*arXiv preprint arXiv:2501.12948*，2025年。

云和，金迪，王超奇，毕 Chloe，Mandyam Karishma，张贺佳，朱晨，李宁，徐腾宇，吕宏江，等。Multi-IF：基准测试多轮和多语言指令跟随的LLMs。*arXiv preprint arXiv:2410.15553*，2024。

丹·亨德里克斯、科林·伯恩斯、史蒂文·巴萨特、安迪·邹、马纳塔斯·马泽伊卡、唐·宋和雅各布·斯坦哈特。衡量大规模多任务语言理解。在 *ICLR*。OpenReview.net，2021a。

丹·亨德里克斯、科林·伯恩斯、索拉夫·卡达瓦斯、阿库尔·阿罗拉、史蒂文·巴萨特、埃里克·唐、唐·宋和雅各布·斯坦哈特。使用 MATH 数据集衡量数学问题解决能力。在 *NeurIPS Datasets and Benchmarks*，2021b。

谢正平、孙思萌、克里曼·塞缪尔、阿查赖·尚坦努、雷克什·迪马、贾飞、张阳、金斯堡·鲍里斯。RULER：你的长上下文语言模型的实际上下文大小是多少？*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.

黄玉珍，白玉卓，朱志豪，张俊雷，张静涵，苏唐军，刘俊腾，吕传成，张亦凯，雷佳怡，傅耀，孙茂松，和何俊贤。C-Eval：面向基础模型的多层次多学科中文评估套件。发表于 *NeurIPS*，2023年。惠彬源，杨建，崔泽宇，杨佳熙，刘大恒，张磊，刘天宇，张佳俊，余博文，卢克明等。Qwen2.5-Coder技术报告。*CoRR*，abs/2409.12186，2024年。 Jain Naman，韩金，Gu Alex，李文丁，闫范佳，张天俊，王思达，Armando Solar-Lezama，Sen Koushik，Ion Stoica。LiveCodeBench：面向代码的大型语言模型的整体和无污染评估。*CoRR*，abs/2403.07974，2024年。江子轩，顾佳奇，朱汉青，David Z. Pan。Pre-RMSNorm 和 Pre-CRMSNorm 转换器：等价且高效的预-LN 转换器。*CoRR*，abs/2305.14858，2023年。Nathan Lambert，Jacob Morrison，Valentina Pyatkin，黄胜义，Hamish Ivison，Faeze Brahman，Lester James V. Miranda，刘艾丽莎，Nouha Dziri，Lyu Shane，谷玉玲，Malik Saumya，Victoria Graf，Jen D. Hwang，杨江江，Ronan Le Bras，Oyvind Tafjord，Chris Wilhelm，Luca Soldaini，Noah A. Smith，王一中，Pradeep Dasigi，Hannaneh Hajishirzi。T ülu 3：推动开源语言模型后训练的前沿。*CoRR*，abs/2411.15124，2024年。李天乐，蒋伟林，Evan Frick，Lisa Dunlap，吴天浩，朱邦华，Joseph E. Gonzalez，Ion Stoica。从众包数据到高质量基准：Arena-Hard 和 BenchBuilder 流水线。*CoRR*，abs/2406.11939，2024年。Hunter Lightman，Vineet Kosaraju，Yura Burda，Harri Edwards，Baker Baker，Teddy Lee，Jan Leike，John Schulman，Ilya Sutskever，Karl Cobbe。让我们一步步验证。*CoRR*，abs/2305.20050，2023年。林裕辰，Ronan Le Bras，Kyle Richardson，Ashish Sabharwal，Radha Poovendran，Peter Clark，Yejin Choi。ZebraLogic：关于LLMs在逻辑推理中的扩展极限。*CoRR*，abs/2502.01100，2025年。刘爱欣，冯贝，薛冰，王冰轩，吴博超，卢成达，赵成刚，邓成奇，张晨宇，阮冲等。DeepSeek-V3技术报告。*arXiv preprint arXiv:2412.19437*，2024年。刘佳伟，夏春秋Steven，王宇瑶，张灵茗。你的代码由ChatGPT生成真的正确吗？对大型语言模型进行代码生成的严格评估。发表于 *NeurIPS*，2023年。刘骞，郑晓森，Niklas Muennighoff，曾广涛，窦龙旭，庞天瑜，姜晶，林敏。RegMix：用于语言模型预训练的数据混合回归。*arXiv preprint arXiv:2407.01492*，2024年。刘晓，雷玄宇，王胜远，黄越，冯卓尔，文博思，程佳乐，柯沛，许一凡，谭永乐，张晓涵，孙立超，王宏宁，张静，黄敏烈，董玉晓，Jie Tang。AlignBench：中文大模型对齐基准测试。*CoRR*，abs/2311.18743，2023年。Meta-AI。Llama 4 群：开启原生多模态AI创新新时代，2025年。网址 `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`。OpenAI。Hello GPT-4o，2024年。网址 `https://openai.com/index/hello-gpt-4o/`。OpenAI。多语言大规模多任务语言理解，2024年。网址 `https://huggingface.co /datasets/openai/MMMLU`。OpenAI。学习推理与LLMs，2024年。网址 `https://openai.com/index/learning-to-reaso n-with-llms/`。OpenAI。介绍openai o3 和 o4-mini，2025年。网址 `https://openai.com/index/introducing-o 3-and-o4-mini/`。Samuel J. Paech。创造性写作v3，2024年。网址 `https://eqbench.com/creative writing.html`。彭博文，Jeffrey Quesnelle，范宏鲁，Enrico Shippole。YaRN：高效扩展大型语言模型上下文窗口的方法。*CoRR*，abs/2309.00071，2023年。邱子涵，黄泽宇，郑博，温凯越，王泽坤，门瑞，Ivan Titov，刘大恒，周景仁，林俊阳。细节中的魔鬼：实现负载平衡损失以训练专业化专家混合模型。*CoRR*，abs/2501.11873，2025年。

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.

尚浩然全，杨佳熙，余博文，郑博，刘大恒，杨安，任轩程，高博飞，苗毅博，冯云龙，王泽坤，杨健，崔泽宇，范阳，张一昌，惠彬源，林俊阳。CodeElo：使用人类可比 Elo 评级对比竞赛级别的 LLM 代码生成。*CoRR*，abs/2501.01257，2025。

Qwen 团队。QwQ：深入反思未知的边界，2024年11月。链接 https://qwenlm.github.io/blog/qwq-32b-preview/。

Qwen 团队。QwQ-32B：拥抱强化学习的力量，2025年3月。网址 https://qwenlm.github.io/blog/qwq-32b/。

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael 和 Samuel R. Bowman。GPQA：一个研究生级别的 Google 级别问答基准。*CoRR*，abs/2311.12022，2023。

A安吉丽卡·罗马诺乌、内贾尔·福鲁坦、安娜·索特尼科娃、泽明·陈、斯瑞·哈沙·内拉图鲁、希瓦利卡·辛格、里沙布·马赫什瓦里、米科尔·奥尔托马雷、穆罕默德·A哈加格、斯涅格哈·A、阿方索·阿迈乌莱斯、阿兹里尔·哈菲兹·阿米鲁丁、维拉特·阿里亚布米、达尼洛·博伊科、迈克尔·张、珍妮·钦、加勒·科恩、阿迪蒂亚·库马尔·达尔米亚、亚伯拉罕·迪雷斯、沙拉德·杜瓦尔、达尼尔·泽纳利乌、丹尼尔·费尔南多·埃拉佐·弗洛雷斯、法比安·法雷斯坦、约瑟夫·马文·帝国、谢耶克·宾·伊斯兰、佩尔图·伊索塔洛、马拉尔·贾巴里希维亚里、博耶·F·卡尔松、埃尔达尔·哈利洛夫、克里斯托弗·克拉姆、法吉里·科托、多米尼克·克热明斯基、加布里埃尔·阿德里亚诺·德·梅洛、西亚莱尔·蒙塔里奥尔、南一阳、乔尔·尼克劳斯、叶卡捷琳娜·诺维科娃、约翰·萨米尔·奥班多·塞隆、德比吉特·保罗、埃丝特·普洛格尔、杰比什·普尔贝、斯瓦蒂·拉贾瓦尔、塞尔万·苏尼莎·拉维、莎拉·瑞德尔、罗尚·桑托什、德里什蒂·沙尔玛、马尔贾纳·普里夫蒂·斯肯杜利、阿尔西亚·索尔塔尼·莫哈尔、巴尔迪亚·索尔塔尼·莫哈尔、兰·塔米尔、阿尤什·库马尔·塔伦、阿兹米恩·图希克·瓦西、塞努卡·奥文·韦拉辛格、塞尔汉·伊尔马兹、迈克·张、伊马诺尔·施拉格、马尔齐耶·法达伊、莎拉·胡克、安托万·博塞卢特。包括：用区域知识评估多语种语言理解能力。*CoRR*，abs/2411.19799，2024。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。使用子词单元进行罕见词的神经机器翻译。在 *ACL (1)*。计算机语言学协会，2016年。

邵志宏，王沛怡，朱启浩，许润新，宋俊孝，张明川，Y. K. Li，Y. Wu，郭大雅。DeepSeekMath：推动开放语言模型中数学推理的极限。*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 和 Jason Wei。语言模型是多语言的链式思考推理器。在 *ICLR*。OpenReview.net，2023。

Guijin Son、Jiwoo Hong、Hyunwoo Ko 和 James Thorne。数学推理中测试时缩放的语言普遍性。*CoRR*，abs/2502.17407，2025。

苏建林，穆尔塔达·H·M·艾哈迈德，卢瑜，潘胜锋，文博，刘云峰。Roformer：带有旋转位置嵌入的增强型Transformer。*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 和 Jason Wei。具有挑战性的 BIG-Bench 任务以及链式思维是否能解决它们。收录于 *ACL (Findings)*，第 13003–13051 页。计算语言学协会，2023年。

Gemma团队，Aishwarya Kamath，Johan Ferret，Shreya Pathak，Nino Vieillard，Ramona Merhej，Sarah Perrin，Tatiana Matejovicova，Alexandre Ramé，Morgane Rivière 等。Gemma 3技术报告。*arXiv preprint arXiv:2503.19786*，2025年。

Changhan Wang, Kyunghyun Cho, 和 Jiatao Gu。基于字节级子词的神经机器翻译。在 *AAAI*，第 9154–9160 页。AAAI 出版社，2020。

王一鸣，张沛，唐嘉龙，韦浩然，杨宝松，王锐，孙晨曙，孙飞彤，张吉然，吴俊轩，藏启乾，张一昌，黄飞，林俊扬，黄飞，周靖然。PolyMath：在多语言环境中评估数学推理能力，2025。

王昱博，马学光，张戈，倪元胜，Abhranil Chandra，郭世光，任伟明，Aaran Arulraj，何轩，江子彦，李天乐，Max Ku，Wang Kai，Zhuang Alex，范荣祺，岳翔，陈文虎。MMLU-Pro：一个更稳健、更具挑战性的多任务语言理解基准。*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.

科林·怀特、塞缪尔·杜利、曼利·罗伯茨、阿尔卡·帕尔、本杰明·费尔、悉达多·贾因、拉维德·施瓦茨-齐夫、尼尔·贾因、哈立德·赛弗拉、悉达多·奈杜、秦梅·赫德、扬·勒康、汤姆·戈德斯坦、威利·奈斯旺格和迈克·戈德布鲁姆。LiveBench：一个具有挑战性且无污染的LLM基准测试。*CoRR*，abs/2406.19314，2024。

吴昀宁，梅嘉豪，严明，李晨亮，赖少鹏，任宇然，王子佳，张济，吴梦悦，金勤，黄飞。WritingBench：一个全面的生成写作基准。*CoRR*，abs/2503.05244，2025。

xAI。Grok 3 测试版——推理代理的时代，2025年。网址 `https://x.ai/news/grok-3`。

桑迈克尔·谢、黄平、董轩怡、杜楠、刘汉霄、卢一峰、梁珮西、黎国维、马腾宇和余伟。Doremi：优化数据混合加快语言模型预训练。*Advances in Neural Information Processing Systems*，36：69798–69818，2023。

文汉雄、刘靖宇、伊戈尔·莫利博格、张贺佳、普拉杰瓦尔·巴尔加瓦、侯锐、路易斯·马丁、拉希·荣塔、卡尔蒂克·阿比纳夫·桑卡拉拉曼、巴拉斯·奥古兹、马迪安·哈布萨、韩方、雅沙尔·梅达德、沙兰·纳兰、克希蒂兹·马利克、安吉拉·范、斯鲁蒂·博萨莱、谢尔盖·爱杜诺夫、迈克·刘易斯、辛农·王和郝马。基础模型的有效长上下文扩展。*CoRR*，abs/2309.16039，2023。

范佳燕、环志毛、纪成杰、张天俊、Shishir G. Patil、Ion Stoica 和 Joseph E. Gonzalez。伯克利函数调用排行榜。`https://gorilla.cs.berkeley.edu/blogs/8 ber keley function calling leaderboard.html`，2024。

安阳、杨宝松、惠彬源、郑博、于博文、周昌、李成鹏、李成远、刘岱恒、黄飞、董冠廷、韦浩然、林欢、唐佳龙、王佳琳、杨健、涂建宏、张建伟、马建新、杨建新、徐金、周景仁、白金泽、何金正、林俊阳、邓凯、陆克明、陈柯勤、杨科新、李梅、薛明峰、倪娜、张佩、王鹏、彭如、门锐、郜瑞泽、林润基、王世杰、白帅、谭思楠、朱天航、李天浩、刘天宇、葛文彬、邓晓东、周晓欢、任兴章、张新宇、魏西平、任宣成、刘雪晶、范阳、姚阳、张义昌、万宇、楚云飞、刘玉琼、崔泽宇、张振儒、郭志芳、范志浩。Qwen2技术报告。*CoRR*，abs/2407.10671，2024a。

安阳，杨宝松，张贝辰，惠彬源，郑博，余博文，李成远，刘大恒，黄飞，魏浩然 等。Qwen2.5技术报告。*arXiv preprint arXiv:2412.15115*，2024b。

安阳，贝辰张，彬源惠，博飞高，博文于，成鹏李，达一恒刘，建宏涂，景仁周，俊阳林等。Qwen2.5-Math技术报告：迈向数学专家模型的自我提升。*CoRR*，abs/2409.12122，2024c。

Yidan Zhang, Boyi Deng, Yu Wan, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin 和 Jingren Zhou。P-MMEval：一个用于一致评估LLMs的并行多语言多任务基准。*CoRR*，abs/2411.09116，2024。

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, 和 Le Hou。大型语言模型的指令遵循评估。*CoRR*，abs/2311.07911，2023。

秦竹、费黄、润宇彭、克明卢、博文于、钦远成、席鹏邱、轩靖黄、和俊阳林。AutoLogi：用于评估大型语言模型推理能力的逻辑谜题自动生成。*CoRR*，abs/2502.16906，2025。