Qwen2.5 Technical Report

Qwen Team

https://huggingface.co/Qwen
https://modelscope.cn/organization/qwen
https://github.com/QwenLM/Qwen2.5

Abstract

In this report, we introduce Qwen2.5, a comprehensive series of large language models (LLMs) designed to meet diverse needs. Compared to previous iterations, Qwen 2.5 has been significantly improved during both the pre-training and post-training stages. In terms of pre-training, we have scaled the high-quality pre-training datasets from the previous 7 trillion tokens to 18 trillion tokens. This provides a strong foundation for common sense, expert knowledge, and reasoning capabilities. In terms of post-training, we implement intricate supervised finetuning with over 1 million samples, as well as multistage reinforcement learning, including offline learning DPO and online learning GRPO. Post-training techniques significantly enhance human preference, and notably improve long text generation, structural data analysis, and instruction following.

To handle diverse and varied use cases effectively, we present Qwen2.5 LLM series in rich configurations. The open-weight offerings include base models and instruction-tuned models in sizes of 0.5B, 1.5B, 3B, 7B, 14B, 32B, and 72B parameters. Quantized versions of the instruction-tuned models are also provided. Over 100 models can be accessed from Hugging Face Hub, ModelScope, and Kaggle. In addition, for hosted solutions, the proprietary models currently include two mixture-of-experts (MoE) variants: Qwen2.5-Turbo and Qwen2.5-Plus, both available from Alibaba Cloud Model Studio.

Qwen2.5 has demonstrated top-tier performance on a wide range of benchmarks evaluating language understanding, reasoning, mathematics, coding, human preference alignment, etc. Specifically, the open-weight flagship Qwen2.5-72B-Instruct outperforms a number of open and proprietary models and demonstrates competitive performance to the state-of-the-art open-weight model, Llama-3-405B-Instruct, which is around 5 times larger. Qwen2.5-Turbo and Qwen2.5-Plus offer superior cost-effectiveness while performing competitively against GPT-4o-mini and GPT-4o respectively. Additionally, as the foundation, Qwen2.5 models have been instrumental in training specialized models such as Qwen2.5-Math (Yang et al., 2024b), Qwen2.5-Coder (Hui et al., 2024), QwQ (Qwen Team, 2024d), and multimodal models.

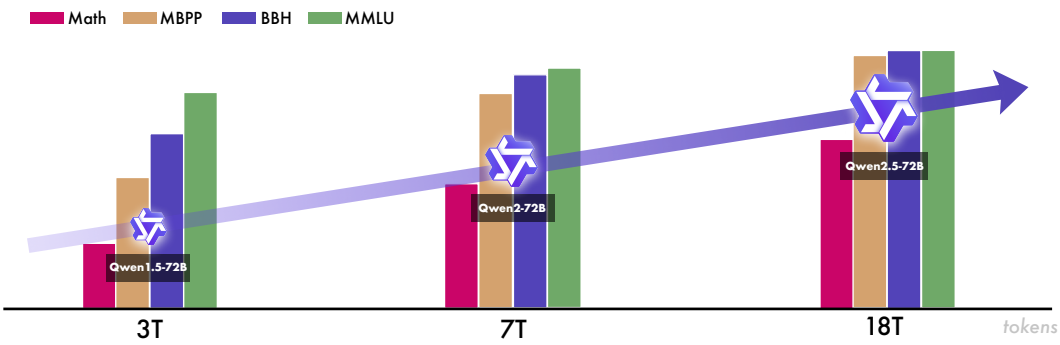

Figure 1: In the iterative development of the Qwen series, data scaling has played a crucial role. Qwen 2.5, which leverages 18 trillion tokens for pre-training, has demonstrated the most advanced capabilities within the Qwen series, especially in terms of domain expertise, underscoring the importance of scale together with mixture in enhancing the model's capabilities.

QWEN2.5技术报告

QWEN团队

https://huggingface.co/qwen https://modelscope.cn/organization/qwen https://github.com/qwenlm/qwen2.5

抽象的

在本报告中，我们介绍了QWEN2.5，这是一系列旨在满足各种需求的大型语言模型（LLMS）。与以前的迭代相比，QWEN 2.5在训练前和训练后阶段都得到了显着改善。在预训练方面，我们将高质量的预训练数据集从前7万亿代币缩放到18万亿代币。这为常识，专家知识和推理能力提供了坚实的基础。在培训后，我们使用超过100万个样本以及多阶段的加强学习，包括离线学习DPO和在线学习GRPO，并实施复杂的监督填充。训练后技术可显着提高人类的偏好，并特别改善长文本生成，结构数据分析和随后的教学。

为了有效地处理多样化和多样化的用例，我们以丰富的配置为QWEN2.5 LLM系列。开放量的产品包括基本模型和指令调整的模型，尺寸为0.5B，1.5B，3B，7B，14B，32B和72B参数。还提供了指令调整模型的量化版本。可以从拥抱面轮，Modelscope和Kaggle访问100多个型号。此外，对于托管解决方案，专有模型目前包括两种混合物（MOE）变体：QWEN2.5- TURBO和QWEN2.5-PLUS，均可从Alibaba Cloud Model Studio获得。

Qwen2.5 has demonstrated top-tier performance on a wide range of benchmarks eval- uating language understanding, reasoning, mathematics, coding, human preference alignment, etc. Specifically, the open-weight flagship Qwen2.5-72B-Instruct outperforms a number of open and proprietary models and demonstrates competitive performance to the state-of-the-art open-weight model, Llama-3-405B-Instruct, which is大约大约5倍。 QWEN2.5-TURBO和QWEN2.5-Plus提供了卓越的成本效益，同时对GPT-4O-Mini和GPT-4O进行了竞争性。此外，作为基础，QWEN2.5模型在训练专业模型中发挥了重要作用，例如QWEN2.5-MATH（Yang等，2024b），Qwen2.5-Coder（Hui等人，2024年），QWQ（QWQ（Qwen Team，2024d）和多模型模型。

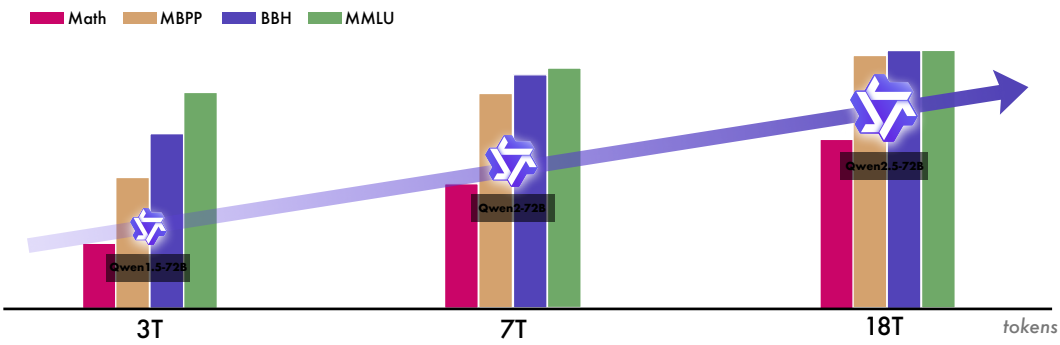

图1：在QWEN系列的迭代发展中，数据扩展起着至关重要的作用。 QWEN 2.5利用18万亿代币进行预训练，已经证明了QWEN系列中最先进的功能，尤其是在域专业知识方面，强调了规模的重要性以及混合物在增强模型功能方面的重要性。

1 Introduction

The sparks of artificial general intelligence (AGI) are increasingly visible through the fast development of large foundation models, notably large language models (LLMs) (Brown et al., 2020; OpenAI, 2023; 2024a; Gemini Team, 2024; Anthropic, 2023a;b; 2024; Bai et al., 2023; Yang et al., 2024a; Touvron et al., 2023a;b; Dubey et al., 2024). The continuous advancement in model and data scaling, combined with the paradigm of large-scale pre-training followed by high-quality supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), has enabled large language models (LLMs) to develop emergent capabilities in language understanding, generation, and reasoning. Building on this foundation, recent breakthroughs in inference time scaling, particularly demonstrated by o1 (OpenAI, 2024b), have enhanced LLMs' capacity for deep thinking through step-by-step reasoning and reflection. These developments have elevated the potential of language models, suggesting they may achieve significant breakthroughs in scientific exploration as they continue to demonstrate emergent capabilities indicative of more general artificial intelligence.

Besides the fast development of model capabilities, the recent two years have witnessed a burst of open (open-weight) large language models in the LLM community, for example, the Llama series (Touvron et al., 2023a;b; Dubey et al., 2024), Mistral series (Jiang et al., 2023a; 2024a), and our Qwen series (Bai et al., 2023; Yang et al., 2024a; Qwen Team, 2024a; Hui et al., 2024; Qwen Team, 2024c; Yang et al., 2024b). The open-weight models have democratized the access of large language models to common users and developers, enabling broader research participation, fostering innovation through community collaboration, and accelerating the development of AI applications across diverse domains.

Recently, we release the details of our latest version of the Qwen series, Qwen2.5. In terms of the open-weight part, we release pre-trained and instruction-tuned models of 7 sizes, including 0.5B, 1.5B, 3B, 7B, 14B, 32B, and 72B, and we provide not only the original models in bfloat16 precision but also the quantized models in different precisions. Specifically, the flagship model Qwen2.5-72B-Instruct demonstrates competitive performance against the state-of-the-art open-weight model, Llama-3-405B-Instruct, which is around 5 times larger. Additionally, we also release the proprietary models of Mixture-of-Experts (MoE, Lepikhin et al., 2020; Fedus et al., 2022; Zoph et al., 2022), namely Qwen2.5-Turbo and Qwen2.5-Plus[1], which performs competitively against GPT-4o-mini and GPT-4o respectively.

In this technical report, we introduce Qwen2.5, the result of our continuous endeavor to create better LLMs. Below, we show the key features of the latest version of Qwen:

- **Better in Size**: Compared with Qwen2, in addition to 0.5B, 1.5B, 7B, and 72B models, Qwen2.5 brings back the 3B, 14B, and 32B models, which are more cost-effective for resource-limited scenarios and are under-represented in the current field of open foundation models. Qwen2.5-Turbo and Qwen2.5-Plus offer a great balance among accuracy, latency, and cost.
- **Better in Data**: The pre-training and post-training data have been improved significantly. The pre-training data increased from 7 trillion tokens to 18 trillion tokens, with focus on knowledge, coding, and mathematics. The pre-training is staged to allow transitions among different mixtures. The post-training data amounts to 1 million examples, across the stage of supervised finetuning (SFT, Ouyang et al., 2022), direct preference optimization (DPO, Rafailov et al., 2023), and group relative policy optimization (GRPO, Shao et al., 2024).
- **Better in Use**: Several key limitations of Qwen2 in use have been eliminated, including larger generation length (from 2K tokens to 8K tokens), better support for structured input and output, (e.g., tables and JSON), and easier tool use. In addition, Qwen2.5-Turbo supports a context length of up to 1 million tokens.

2 Architecture & Tokenizer

Basically, the Qwen2.5 series include dense models for opensource, namely Qwen2.5-0.5B / 1.5B / 3B / 7B / 14B / 32B / 72B, and MoE models for API service, namely Qwen2.5-Turbo and Qwen2.5-Plus. Below, we provide details about the architecture of models.

For dense models, we maintain the Transformer-based decoder architecture (Vaswani et al., 2017; Radford et al., 2018) as Qwen2 (Yang et al., 2024a). The architecture incorporates several key components: Grouped Query Attention (GQA, Ainslie et al., 2023) for efficient KV cache utilization, SwiGLU activation function (Dauphin et al., 2017) for non-linear activation, Rotary Positional Embeddings (RoPE, Su

[1]Qwen2.5-Turbo is identified as `qwen-turbo-2024-11-01` and Qwen2.5-Plus is identified as `qwen-plus-2024-xx-xx` (to be released) in the API.

1 简介

通过大型基础模型，尤其是大型语言模型（LLM）的快速发展，通用人工智能（AGI）的火花越来越明显（Brown et al., 2020; OpenAI, 2023; 2024a; Gemini Team, 2024; Anthropic, 2023a;b; 2024; Bai et al., 2023; Yang et al., 2024a; Touvron）等人，2023a；b；Dubey 等人，2024）。模型和数据扩展的不断进步，结合大规模预训练的范式，然后是高质量的监督微调（SFT）和来自人类反馈的强化学习（RLHF）（Ouyang et al., 2022），使得大型语言模型（LLM）能够发展语言理解、生成和推理方面的新兴能力。在此基础上，最近在推理时间缩放方面取得的突破，特别是 o1 (OpenAI, 2024b) 所证明的突破，增强了法学硕士通过逐步推理和反思进行深度思考的能力。这些发展提高了语言模型的潜力，表明它们可能在科学探索中取得重大突破，因为它们继续展示表明更通用人工智能的新兴能力。

除了模型能力的快速发展之外，近两年LLM社区还出现了一系列开放（open-weight）大型语言模型，例如Llama系列（Touvron et al., 2023a;b; Dubey et al., 2024）、Mistral系列（Jiang et al., 2023a; 2024a）和我们的Qwen系列（Bai et al., 2023; Yang）等人，2024a；Qwen 团队，2024a；Hui 等人，2024；Yang 等人，2024b）。开放权重模型使普通用户和开发人员能够民主地访问大型语言模型，从而实现更广泛的研究参与，通过社区协作促进创新，并加速跨不同领域的人工智能应用程序的开发。

近日，我们发布了Qwen系列最新版本Qwen2.5的详细信息。在开放权重部分，我们发布了7种尺寸的预训练和指令调整模型，包括0.5B、1.5B、3B、7B、14B、32B和72B，我们不仅提供bfloat16精度的原始模型，还提供不同精度的量化模型。具体来说，旗舰型号 Qwen2.5-72B-Instruct 表现出与最先进的开放式重量模型 Llama-3-405B-Instruct 的竞争性能，Llama-3-405B-Instruct 的尺寸大约是后者的 5 倍。此外，我们还发布了 Mixture-of-Experts 的专有模型（MoE, Lepikhin et al., 2020；Fedus et al., 2022；Zoph et al., 2022），即 Qwen2.5-Turbo 和 Qwen2.5-Plus[1]，其性能分别与 GPT-4o-mini 和 GPT-4o 竞争。

在这份技术报告中，我们介绍了Qwen2.5，这是我们不断努力创建更好的法学硕士的成果。下面，我们展示最新版本Qwen的主要功能：

- 尺寸更优：与Qwen2相比，除了0.5B、1.5B、7B、72B模型外，Qwen2.5还带回了3B、14B、32B模型，这些模型对于资源有限的场景更具性价比，在当前开放基础模型领域代表性不足。 Qwen2.5- Turbo 和 Qwen2.5-Plus 在准确性、延迟和成本之间提供了良好的平衡。
- 数据更好：训练前和训练后数据都有显着改善。预训练数据从7万亿个token增加到18万亿个token，重点是知识、编码和数学。预训练是分阶段进行的，以允许不同混合物之间的过渡。训练后数据达到 100 万个样本，涵盖监督微调（SFT，Ouyang et al.，2022）、直接偏好优化（DPO，Rafailov et al.，2023）和群体相对策略优化（GRPO，Shao et al.，2024）阶段。
- 更好的使用：Qwen2 在使用中的几个关键限制已经被消除，包括更大的生成长度（从 2K 令牌到 8K 令牌）、更好地支持结构化输入和输出（例如表和 JSON）以及更容易的工具使用。此外，Qwen2.5-Turbo 支持高达 100 万个令牌的上下文长度。

2 架构和分词器

基本上，Qwen2.5系列包括用于开源的密集模型，即Qwen2.5-0.5B / 1.5B / 3B / 7B / 14B / 32B / 72B，以及用于API服务的MoE模型，即Qwen2.5-Turbo和Qwen2.5-Plus。下面，我们提供有关模型架构的详细信息。

对于密集模型，我们将基于 Transformer 的解码器架构（Vaswani 等人，2017；Radford 等人，2018）维护为 Qwen2（Yang 等人，2024a）。该架构包含几个关键组件：用于高效 KV 缓存利用的分组查询注意（GQA，Ainslie 等人，2023）、用于非线性激活的 SwiGLU 激活函数（Dauphin 等人，2017）、旋转位置嵌入（RoPE，Su）

[1]Qwen2.5-Turbo is identified as `qwen-turbo-2024-11-01` and Qwen2.5-Plus is identified as `qwen-plus-2024-xx-xx` (to be released) in the API.

Table 1: Model architecture and license of Qwen2.5 open-weight models.

Models	Layers	Heads (Q / KV)	Tie Embedding	Context / Generation Length	License
0.5B	24	14 / 2	Yes	32K / 8K	Apache 2.0
1.5B	28	12 / 2	Yes	32K / 8K	Apache 2.0
3B	36	16 / 2	Yes	32K / 8K	Qwen Research
7B	28	28 / 4	No	128K / 8K	Apache 2.0
14B	48	40 / 8	No	128K / 8K	Apache 2.0
32B	64	40 / 8	No	128K / 8K	Apache 2.0
72B	80	64 / 8	No	128K / 8K	Qwen

et al., 2024) for encoding position information, QKV bias (Su, 2023) in the attention mechanism and RMSNorm (Jiang et al., 2023b) with pre-normalization to ensure stable training.

Building upon the dense model architectures, we extend it to MoE model architectures. This is achieved by replacing standard feed-forward network (FFN) layers with specialized MoE layers, where each layer comprises multiple FFN experts and a routing mechanism that dispatches tokens to the top-K experts. Following the approaches demonstrated in Qwen1.5-MoE (Yang et al., 2024a), we implement fine-grained expert segmentation (Dai et al., 2024) and shared experts routing (Rajbhandari et al., 2022; Dai et al., 2024). These architectural innovations have yielded substantial improvements in model performance across downstream tasks.

For tokenization, we 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 of 151,643 regular tokens. We have expanded the set of control tokens from 3 to 22 compared to previous Qwen versions, adding two new tokens for tool functionality and allocating the remainder for other model capabilities. This expansion establishes a unified vocabulary across all Qwen2.5 models, enhancing consistency and reducing potential compatibility issues.

3 Pre-training

Our language model pre-training process consists of several key components. First, we carefully curate high-quality training data through sophisticated filtering and scoring mechanisms, combined with strategic data mixture. Second, we conduct extensive research on hyperparameter optimization to effectively train models at various scales. Finally, we incorporate specialized long-context pre-training to enhance the model's ability to process and understand extended sequences. Below, we detail our approaches to data preparation, hyperparameter selection, and long-context training.

3.1 Pre-training Data

Qwen2.5 demonstrates significant enhancements in pre-training data quality compared to its predecessor Qwen2. These improvements stem from several key aspects:

(1) **Better data filtering**. High-quality pre-training data is crucial for model performance, making data quality assessment and filtering a critical component of our pipeline. We leverage Qwen2-Instruct models as data quality filters that perform comprehensive, multi-dimensional analysis to evaluate and score training samples. The filtering method represents a significant advancement over our previous approach used for Qwen2, as it benefits from Qwen2's expanded pre-training on a larger multilingual corpus. The enhanced capabilities enable more nuanced quality assessment, resulting in both improved retention of high-quality training data and more effective filtering of low-quality samples across multiple languages.

(2) **Better math and code data**. During the pre-training phase of Qwen2.5, we incorporate training data from Qwen2.5-Math (Yang et al., 2024b) and Qwen2.5-Coder (Hui et al., 2024). This data integration strategy proves highly effective, as these specialized datasets are instrumental in achieving state-of-the-art performance on mathematical and coding tasks. By leveraging these high-quality domain-specific datasets during pre-training, Qwen2.5 inherits strong capabilities in both mathematical reasoning and code generation.

(3) **Better synthetic data**. To generate high-quality synthetic data, particularly in mathematics, code, and knowledge domains, we leverage both Qwen2-72B-Instruct (Yang et al., 2024a) and Qwen2-Math-72B-Instruct (Qwen Team, 2024c). The quality of this synthesized data is further enhanced through rigorous filtering using our proprietary general reward model and the specialized Qwen2-Math-RM-72B (Qwen Team, 2024c) model.

表1：Qwen2.5开放权重模型的模型架构和许可。

Models	Layers	Heads (Q / KV)	Tie Embedding	Context / Generation Length	License
0.5B	24	14 / 2	Yes	32K / 8K	Apache 2.0
1.5B	28	12 / 2	Yes	32K / 8K	Apache 2.0
3B	36	16 / 2	Yes	32K / 8K	Qwen Research
7B	28	28 / 4	No	128K / 8K	Apache 2.0
14B	48	40 / 8	No	128K / 8K	Apache 2.0
32B	64	40 / 8	No	128K / 8K	Apache 2.0
72B	80	64 / 8	No	128K / 8K	Qwen

et al., 2024）用于编码位置信息，注意力机制中的 QKV 偏差（Su, 2023）和带有预归一化的 RMSNorm（Jiang et al., 2023b）以确保稳定的训练。

在密集模型架构的基础上，我们将其扩展到 MoE 模型架构。这是通过用专门的 MoE 层替换标准前馈网络 (FFN) 层来实现的，其中每层都包含多个 FFN 专家和一个将令牌分派给前 K 个专家的路由机制。遵循 Qwen1.5-MoE（Yang 等人，2024a）中演示的方法，我们实现了细粒度专家分割（Dai 等人，2024）和共享专家路由（Rajbhandari 等人，2022；Dai 等人，2024）。这些架构创新极大地提高了下游任务的模型性能。

对于标记化，我们利用 Qwen 的标记生成器（Bai et al., 2023），它实现了字节级字节对编码（BBPE, Brown et al., 2020；Wang et al., 2020；Sennrich et al., 2016），词汇表包含 151,643 个常规标记。与之前的 Qwen 版本相比，我们将控制令牌集从 3 个扩展至 22 个，为工具功能添加了两个新令牌，并将其余令牌分配给其他模型功能。此扩展在所有 Qwen2.5 模型中建立了统一的词汇表，增强了一致性并减少了潜在的兼容性问题。

3预训练

我们的语言模型预训练过程由几个关键组成部分组成。首先，我们通过复杂的过滤和评分机制，结合战略数据混合，精心策划高质量的训练数据。其次，我们对超参数优化进行了广泛的研究，以有效地训练各种规模的模型。最后，我们结合了专门的长上下文预训练，以增强模型处理和理解扩展序列的能力。下面，我们详细介绍了数据准备、超参数选择和长上下文训练的方法。

3.1预训练数据

与前身 Qwen2 相比，Qwen2.5 展示了预训练数据质量的显着增强。这些改进源于几个关键方面：

(1)更好的数据过滤。高质量的预训练数据对于模型性能至关重要，数据质量评估和过滤是我们流程的关键组成部分。我们利用 Qwen2-Instruct 模型作为数据质量过滤器，执行全面的多维分析来评估和评分训练样本。该过滤方法比我们之前用于 Qwen2 的方法取得了重大进步，因为它受益于 Qwen2 在更大的多语言语料库上进行的扩展预训练。增强的功能可实现更细致的质量评估，从而提高高质量训练数据的保留率，并更有效地过滤多种语言的低质量样本。

(2) 更好的数学和代码数据。在 Qwen2.5 的预训练阶段，我们合并了来自 Qwen2.5-Math (Yang et al., 2024b) 和 Qwen2.5-Coder (Hui et al., 2024) 的训练数据。这种数据集成策略被证明是非常有效的，因为这些专门的数据集有助于在数学和编码任务上实现最先进的性能。通过在预训练期间利用这些高质量的特定领域数据集，Qwen2.5继承了数学推理和代码生成方面的强大能力。

(3) 更好的综合数据。为了生成高质量的合成数据，特别是在数学、代码和知识领域，我们利用 Qwen2-72B-Instruct（Yang 等人，2024a）和 Qwen2-Math-72B-Instruct（Qwen Team，2024c）。通过使用我们专有的一般奖励模型和专门的 Qwen2-Math-RM-72B（Qwen Team，2024c）模型进行严格过滤，进一步提高了合成数据的质量。

(4) **Better data mixture**. To optimize the pre-training data distribution, we employ Qwen2-Instruct models to classify and balance content across different domains. Our analysis revealed that domains like e-commerce, social media, and entertainment are significantly overrepresented in web-scale data, often containing repetitive, template-based, or machine-generated content. Conversely, domains such as technology, science, and academic research, while containing higher-quality information, are traditionally underrepresented. Through strategic down-sampling of overrepresented domains and up-sampling of high-value domains, we ensure a more balanced and information-rich training dataset that better serves our model's learning objectives.

Building on these techniques, we have developed a larger and higher-quality pre-training dataset, expanding from the 7 trillion tokens used in Qwen2 (Yang et al., 2024a) to **18 trillion** tokens.

3.2 Scaling Law for Hyper-parameters

We develop scaling laws for hyper-parameter based on the pre-training data of Qwen2.5 (Hoffmann et al., 2022; Kaplan et al., 2020). While previous studies (Dubey et al., 2024; Almazrouei et al., 2023; Hoffmann et al., 2022) primarily used scaling laws to determine optimal model sizes given compute budgets, we leverage them to identify optimal hyperparameters across model architectures. Specifically, our scaling laws help determine key training parameters like batch size B and learning rate μ for both dense models and MoE models of varying sizes.

Through extensive experimentation, we systematically study the relationship between model architecture and optimal training hyper-parameters. Specifically, we analyze how the optimal learning rate μ_{opt} and batch size B_{opt} vary with model size N and pre-training data size D. Our experiments cover a comprehensive range of architectures, including dense models with 44M to 14B parameters and MoE models with 44M to 1B activated parameters, trained on datasets ranging from 0.8B to 600B tokens. Using these optimal hyper-parameter predictions, we then model the final loss as a function of model architecture and training data scale.

Additionally, we leverage scaling laws to predict and compare the performance of MoE models with varying parameter counts against their dense counterparts. This analysis guides our hyper-parameter configuration for MoE models, enabling us to achieve performance parity with specific dense model variants (such as Qwen2.5-72B and Qwen2.5-14B) through careful tuning of both activated and total parameters.

3.3 Long-context Pre-training

For optimal training efficiency, Qwen2.5 employs a two-phase pre-training approach: an initial phase with a 4,096-token context length, followed by an extension phase for longer sequences. Following the strategy used in Qwen2, we extend the context length from 4,096 to 32,768 tokens during the final pre-training stage for all model variants except Qwen2.5-Turbo. Concurrently, we increase the base frequency of RoPEfrom 10,000 to 1,000,000 using the ABF technique (Xiong et al., 2023).

For Qwen2.5-Turbo, we implement a progressive context length expansion strategy during training, advancing through four stages: 32,768 tokens, 65,536 tokens, 131,072 tokens, and ultimately 262,144 tokens, with a RoPE base frequency of 10,000,000. At each stage, we carefully curate the training data to include 40% sequences at the current maximum length and 60% shorter sequences. This progressive training methodology enables smooth adaptation to increasing context lengths while maintaining the model's ability to effectively process and generalize across sequences of varying lengths.

To enhance our models' ability to process longer sequences during inference, we implement two key strategies: YARN (Peng et al., 2023) and Dual Chunk Attention (DCA, An et al., 2024). Through these innovations, we achieve a four-fold increase in sequence length capacity, enabling Qwen2.5-Turbo to handle up to **1 million** tokens and other models to process up to 131,072 tokens. Notably, these approaches not only improve the modeling of long sequences by reducing perplexity but also maintain the models' strong performance on shorter sequences, ensuring consistent quality across varying input lengths.

4 Post-training

Qwen 2.5 introduces two significant advancements in its post-training design compared to Qwen 2:

(1) **Expanded Supervised Fine-tuning Data Coverage:** The supervised fine-tuning process leverages a massive dataset comprising millions of high-quality examples. This expansion specifically addresses key areas where the previous model showed limitations, such as long-sequence

(4)更好的数据混合。为了优化预训练数据分布，我们采用 Qwen2-Instruct 模型来分类和平衡不同领域的内容。我们的分析显示，电子商务、社交媒体和娱乐等领域在网络规模数据中所占比例明显过高，通常包含重复的、基于模板或机器生成的内容。相反，技术、科学和学术研究等领域虽然包含更高质量的信息，但传统上代表性不足。通过对代表性过高的领域进行战略性下采样和对高价值领域进行上采样，我们确保了更加平衡和信息丰富的训练数据集，更好地服务于我们模型的学习目标。

基于这些技术，我们开发了一个更大、更高质量的预训练数据集，从 Qwen2（Yang 等人，2024a）中使用的 7 万亿个令牌扩展到 18 万亿个令牌。

3.2 超参数的缩放定律

我们根据 Qwen2.5 的预训练数据制定超参数的缩放定律（Hoffmann et al., 2022；Kaplan et al., 2020）。虽然之前的研究（Dubey 等人，2024；Almazrouei 等人，2023；Hoffmann 等人，2022）主要使用缩放定律来确定给定计算预算的最佳模型大小，但我们利用它们来识别跨模型架构的最佳超参数。具体来说，我们的缩放定律有助于确定不同大小的密集模型和 MoE 模型的关键训练参数，例如批量大小 B 和学习率 μ。

通过大量的实验，我们系统地研究了模型架构和最优训练超参数之间的关系。具体来说，我们分析了最佳学习率 μ_{opt} 和批量大小 B_{opt} 如何随模型大小 N 和预训练数据大小 D 变化。我们的实验涵盖了全面的架构，包括具有 44M 到 14B 参数的密集模型和具有 44M 到 1B 激活参数的 MoE 模型，在 0.8B 到 600B 令牌的数据集上进行训练。使用这些最佳超参数预测，我们将最终损失建模为模型架构和训练数据规模的函数。

此外，我们利用缩放定律来预测和比较具有不同参数数量的 MoE 模型与密集模型的性能。该分析指导我们对 MoE 模型的超参数配置，使我们能够通过仔细调整激活参数和总参数来实现与特定密集模型变体（例如 Qwen2.5-72B 和 Qwen2.5-14B）的性能相当。

3.3 长上下文预训练

为了获得最佳训练效率，Qwen2.5 采用两阶段预训练方法：初始阶段具有 4,096 个令牌上下文长度，随后是更长序列的扩展阶段。遵循 Qwen2 中使用的策略，在最后的预训练阶段，我们将除 Qwen2.5-Turbo 之外的所有模型变体的上下文长度从 4,096 个标记扩展到 32,768 个标记。同时，我们使用 ABF 技术将 RoPE 的基频从 10,000 增加到 1,000,000（Xiong 等人，2023）。

对于 Qwen2.5-Turbo，我们在训练期间实施渐进式上下文长度扩展策略，通过四个阶段推进：32,768 个令牌、65,536 个令牌、131,072 个令牌，最终 262,144 个令牌，RoPE 基频为 10,000,000。在每个阶段，我们都会仔细整理训练数据，以包含当前最大长度的 40% 序列和较短的 60% 序列。这种渐进式训练方法能够平滑适应不断增加的上下文长度，同时保持模型有效处理和泛化不同长度序列的能力。

为了增强模型在推理过程中处理较长序列的能力，我们实施了两种关键策略：YARN（Peng 等人，2023）和双块注意力（DCA，An 等人，2024）。通过这些创新，我们实现了序列长度容量的四倍增长，使 Qwen2.5-Turbo 能够处理多达 100 万个令牌，其他模型能够处理多达 131,072 个令牌。值得注意的是，这些方法不仅通过减少复杂性来改进长序列的建模，而且还保持了模型在较短序列上的强大性能，确保不同输入长度的质量一致。

4训练后

与 Qwen 2 相比，Qwen 2.5 在训练后设计上引入了两项重大改进：

(1) 扩大监督微调数据覆盖范围：监督微调过程利用包含数百万个高质量示例的海量数据集。这种扩展专门解决了先前模型显示出局限性的关键领域，例如长序列

generation, mathematical problem-solving, coding, instruction-following, structured data understanding, logical reasoning, cross-lingual transfer, and robust system instruction.

(2) **Two-stage Reinforcement Learning:** The reinforcement learning (RL) process in Qwen 2.5 is divided into two distinct stages: Offline RL and Online RL.

- *Offline RL:* This stage focuses on developing capabilities that are challenging for the reward model to evaluate, such as reasoning, factuality, and instruction-following. Through meticulous construction and validation of training data, we ensure that the Offline RL signals are both learnable and reliable (Xiang et al., 2024), enabling the model to acquire those complex skills effectively.
- *Online RL:* The Online RL phase leverages the reward model's ability to detect nuances in output quality, including truthfulness, helpfulness, conciseness, relevance, harmlessness and debiasing. It enables the model to generate responses that are precise, coherent, and well-structured while maintaining safety and readability. As a result, the model's outputs consistently meet human quality standards and expectations.

4.1 Supervised Fine-tuning

In this section, we detail the key enhancements made during the SFT phase of Qwen2.5, focusing on several critical areas:

(1) **Long-sequence Generation:** Qwen2.5 is capable of generating high-quality content with an output context length of up to 8,192 tokens, a significant advancement over the typical post-training response length, which often remains under 2,000 tokens. To address this gap, we develop long-response datasets (Quan et al., 2024). We employ back-translation techniques to generate queries for long-text data from pre-training corpora, impose output length constraints, and use Qwen2 to filter out low-quality paired data.

(2) **Mathematics:** We introduce the chain-of-thought data of Qwen2.5-Math (Yang et al., 2024b), which encompasses a diverse range of query sources, including public datasets, K-12 problem collections, and synthetic problems. To ensure high-quality reasoning, we employ rejection sampling (Yuan et al., 2023) along with reward modeling and annotated answers for guidance, producing step-by-step reasoning process.

(3) **Coding:** To enhance coding capabilities, we incorporate the instruction tuning data of Qwen2.5-Coder (Hui et al., 2024). We use multiple language-specific agents into a collaborative framework, generating diverse and high-quality instruction pairs across nearly 40 programming languages. We expand our instruction dataset by synthesizing new examples from code-related Q&A websites and gathering algorithmic code snippets from GitHub. A comprehensive multilingual sandbox is used to perform static code checking and validate code snippets through automated unit testing, ensuring code quality and correctness (Dou et al., 2024; Yang et al., 2024c).

(4) **Instruction-following:** To ensure high-quality instruction-following data, we implement a rigorous code-based validation framework. In this approach, LLMs generate both instructions and corresponding verification code, along with comprehensive unit tests for cross-validation. Through execution feedback-based rejection sampling, we carefully curate the training data used for Supervised Fine-Tuning, thereby guaranteeing the model's faithful adherence to intended instructions (Dong et al., 2024).

(5) **Structured Data Understanding:** We develop a comprehensive structured understanding dataset that encompasses both traditional tasks, such as tabular question-answering, fact verification, error correction, and structural understanding, as well as complex tasks involving structured and semi-structured data. By incorporating reasoning chains into the model's responses, we significantly enhance its ability to infer information from structured data, thereby improving its performance across these diverse tasks. This approach not only broadens the scope of the dataset but also deepens the model's capacity to reason and derive meaningful insights from complex data structures.

(6) **Logical Reasoning:** To enhance the model's logical reasoning capabilities, we introduce a diverse set of 70,000 new queries spanning various domains. These queries encompass multiple-choice questions, true / false questions, and open-ended questions. The model is trained to approach problems systematically, employing a range of reasoning methods such as deductive reasoning, inductive generalization, analogical reasoning, causal reasoning, and statistical reasoning. Through iterative refinement, we systematically filter out data containing incorrect answers or flawed reasoning processes. This process progressively strengthens the model's ability to reason logically and accurately, ensuring robust performance across different types of reasoning tasks.

生成、数学问题解决、编码、指令跟踪、结构化数据理解、逻辑推理、跨语言迁移和强大的系统指令。

（2）两阶段强化学习：Qwen 2.5中的强化学习（RL）过程分为两个不同的阶段：离线强化学习和在线强化学习。

- *Offline RL:* 此阶段的重点是开发对奖励模型评估具有挑战性的能力，例如推理、事实性和遵循指令。通过精心构建和验证训练数据，我们确保离线强化学习信号既可学习又可靠（Xiang et al., 2024），使模型能够有效地获得这些复杂的技能。

- *Online RL:* 在线强化学习阶段利用奖励模型检测输出质量细微差别的能力，包括真实性、有用性、简洁性、相关性、无害性和去偏差。它使模型能够生成精确、连贯且结构良好的响应，同时保持安全性和可读性。因此，该模型的输出始终满足人类质量标准和期望。

4.1 有监督微调

在本节中，我们将详细介绍 Qwen2.5 SFT 阶段所做的关键增强，重点关注几个关键领域：

(1) 长序列生成：Qwen2.5 能够生成输出上下文长度高达 8,192 个令牌的高质量内容，这比典型的训练后响应长度（通常保持在 2,000 个令牌以下）有显着进步。为了弥补这一差距，我们开发了长响应数据集（Quan 等人，2024）。我们采用反向翻译技术从预训练语料库中生成长文本数据的查询，施加输出长度限制，并使用 Qwen2 过滤掉低质量的配对数据。 （2）数学：我们引入了Qwen2.5-Math（Yang et al., 2024b）的思想链数据，它涵盖了各种查询源，包括公共数据集、K-12问题集合和综合问题。为了确保高质量的推理，我们采用拒绝抽样（Yuan et al., 2023）以及奖励建模和带注释的答案作为指导，产生逐步的推理过程。 （3）编码：为了增强编码能力，我们纳入了Qwen2.5-Coder的指令调优数据（Hui et al., 2024）。我们在协作框架中使用多种特定于语言的代理，生成跨近 40 种编程语言的多样化且高质量的指令对。我们通过综合来自代码相关问答网站的新示例并从 GitHub 收集算法代码片段来扩展我们的指令数据集。使用全面的多语言沙箱进行静态代码检查，并通过自动化单元测试验证代码片段，确保代码质量和正确性（Dou et al., 2024；Yang et al., 2024c）。 (4) 指令跟踪：为了确保高质量的指令跟踪数据，我们实施了严格的基于代码的验证框架。在这种方法中，法学硕士生成指令和相应的验证代码，以及用于交叉验证的全面单元测试。通过基于执行反馈的拒绝采样，我们精心策划用于监督微调的训练数据，从而保证模型忠实地遵守预期指令（Dong et al., 2024）。 （5）结构化数据理解：我们开发了一个全面的结构化理解数据集，既涵盖表格问答、事实验证、纠错和结构理解等传统任务，也涵盖涉及结构化和半结构化数据的复杂任务。通过将推理链纳入模型的响应中，我们显着增强了其从结构化数据推断信息的能力，从而提高了其在这些不同任务中的性能。这种方法不仅扩大了数据集的范围，而且加深了模型的推理能力并从复杂的数据结构中得出有意义的见解。 (6) 逻辑推理：为了增强模型的逻辑推理能力，我们引入了一组跨越各个领域的 70,000 个新查询。这些查询包括多项选择题、对/错问题和开放式问题。该模型经过训练可以系统地处理问题，采用一系列推理方法，例如演绎推理、归纳概括、类比推理、因果推理和统计推理。通过迭代细化，我们系统地过滤掉包含错误答案或有缺陷的推理过程的数据。这一过程逐步增强模型逻辑、准确推理的能力，确保在不同类型的推理任务中具有稳健的性能。

(7) **Cross-Lingual Transfer:** To facilitate the transfer of the model's general capabilities across languages, we employ a translation model to convert instructions from high-resource languages into various low-resource languages, thereby generating corresponding response candidates. To ensure the accuracy and consistency of these responses, we evaluate the semantic alignment between each multilingual response and its original counterpart. This process preserves the logical structure and stylistic nuances of the original responses, thereby maintaining their integrity and coherence across different languages.

(8) **Robust System Instruction:** We construct hundreds of general system prompts to improve the diversity of system prompts in post-training, ensuring consistency between system prompts and conversations. Evaluations with different system prompts show that the model maintains good performance (Lu et al., 2024b) and reduced variance, indicating improved robustness.

(9) **Response Filtering:** To evaluate the quality of responses, we employ multiple automatic annotation methods, including a dedicated critic model and a multi-agent collaborative scoring system. Responses are subjected to rigorous assessment, and only those deem flawless by all scoring systems are retained. This comprehensive approach ensures that our outputs maintain the highest quality standards.

Ultimately, we construct a dataset of over 1 million SFT examples. The model is fine-tuned for two epochs with a sequence length of 32,768 tokens. To optimize learning, the learning rate is gradually decreased from 7×10^{-6} to 7×10^{-7}. To address overfitting, we apply a weight decay of 0.1, and gradient norms are clipped at a maximum value of 1.0.

4.2 Offline Reinforcement Learning

Compared to Online Reinforcement Learning (RL), Offline RL enables the pre-preparation of training signals, which is particularly advantageous for tasks where standard answers exist but are challenging to evaluate using reward models. In this study, we focus on objective query domains such as mathematics, coding, instruction following, and logical reasoning, where obtaining accurate evaluations can be complex. In the previous phase, we extensively employ strategies like execution feedback and answer matching to ensure the quality of responses. For the current phase, we reuse that pipeline, employing the SFT model to resample responses for a new set of queries. Responses that pass our quality checks are used as positive examples, while those that fail are treated as negative examples for Direct Preference Optimization (DPO) training (Rafailov et al., 2023). To further enhance the reliability and accuracy of the training signals, we make use of both human and automated review processes (Cao et al., 2024). This dual approach ensures that the training data is not only learnable but also aligned with human expectations. Ultimately, we construct a dataset consisting of approximately 150,000 training pairs. The model is then trained for one epoch using the Online Merging Optimizer (Lu et al., 2024a), with a learning rate of 7×10^{-7}.

4.3 Online Reinforcement Learning

To develop a robust reward model for online RL, we adhere to a set of carefully defined labeling criteria. Those criteria ensure that the responses generated by the model are not only high-quality but also aligned with ethical and user-centric standards (Wang et al., 2024a). The specific guidelines for data labeling are as follows:

- **Truthfulness:** Responses must be grounded in factual accuracy, faithfully reflecting the provided context and instructions. The model should avoid generating information that is false or unsupported by the given data.
- **Helpfulness:** The model's output should be genuinely useful, addressing the user's query effectively while providing content that is positive, engaging, educational, and relevant. It should follow the given instructions precisely and offer value to the user.
- **Conciseness:** Responses should be succinct and to the point, avoiding unnecessary verbosity. The goal is to convey information clearly and efficiently without overwhelming the user with excessive detail.
- **Relevance:** All parts of the response should be directly related to the user's query, dialogue history, and the assistant's context. The model should tailor its output to ensure it is perfectly aligned with the user's needs and expectations.
- **Harmlessness:** The model must prioritize user safety by avoiding any content that could lead to illegal, immoral, or harmful behavior. It should promote ethical conduct and responsible communication at all times.

（7）跨语言迁移：为了促进模型的通用能力跨语言迁移，我们采用翻译模型将指令从高资源语言转换为各种低资源语言，从而生成相应的响应候选。为了确保这些响应的准确性和一致性，我们评估每个多语言响应与其原始对应响应之间的语义对齐。此过程保留了原始响应的逻辑结构和风格上的细微差别，从而保持了不同语言之间的完整性和连贯性。（8）强大的系统指令：我们构建了数百个通用系统提示，以提高训练后系统提示的多样性，确保系统提示和对话之间的一致性。不同系统提示下的评估表明，该模型保持了良好的性能（Lu et al., 2024b），并且方差减少，表明鲁棒性有所提高。（9）响应过滤：为了评估响应的质量，我们采用了多种自动注释方法，包括专用的评论家模型和多智能体协作评分系统。回答要经过严格的评估，只有那些被所有评分系统认为完美的回答才会被保留。这种全面的方法可确保我们的产品保持最高的质量标准。

最终，我们构建了一个包含超过 100 万个 SFT 示例的数据集。该模型针对序列长度为 32,768 个令牌的两个时期进行了微调。为了优化学习，学习率从 7×10^{-6} 逐渐降低到 7×10^{-7}。为了解决过度拟合问题，我们应用 0.1 的权重衰减，并将梯度范数限制为最大值 1.0。

4.2 离线强化学习

与在线强化学习（RL）相比，离线强化学习可以预先准备训练信号，这对于存在标准答案但难以使用奖励模型进行评估的任务特别有利。在本研究中，我们重点关注数学、编码、指令遵循和逻辑推理等客观查询领域，在这些领域获得准确的评估可能很复杂。在前一阶段，我们广泛采用执行反馈、答案匹配等策略来保证响应的质量。在当前阶段，我们重用该管道，采用 SFT 模型对一组新查询的响应进行重新采样。通过质量检查的响应将被用作正面示例，而那些未通过质量检查的响应将被视为直接偏好优化 (DPO) 训练的反面示例（Rafailov 等人，2023）。为了进一步提高训练信号的可靠性和准确性，我们利用人工和自动审查流程（Cao 等人，2024）。这种双重方法确保训练数据不仅是可学习的，而且符合人类的期望。最终，我们构建了一个由大约 150,000 个训练对组成的数据集。然后使用在线合并优化器（Lu et al., 2024a）对模型进行一个 epoch 的训练，学习率为 7×10^{-7}。

4.3 在线强化学习

为了开发一个强大的在线强化学习奖励模型，我们遵循一套精心定义的标签标准。这些标准确保模型生成的响应不仅是高质量的，而且符合道德和以用户为中心的标准（Wang 等人，2024a）。数据标注的具体准则如下：

- 真实性：回答必须基于事实准确性，忠实反映所提供的背景和说明。模型应避免生成错误的或不被给定数据支持的信息。
- 有用性：模型的输出应该真正有用，有效解决用户的查询，同时提供积极、有吸引力、有教育意义和相关的内容。它应该准确遵循给定的说明并为用户提供价值。
- 简洁：回答应该简洁明了，避免不必要的冗长。目标是清晰有效地传达信息，而又不会因过多的细节而让用户感到不知所措。
- 相关性：响应的所有部分都应与用户的查询、对话历史记录和助手的上下文直接相关。该模型应该调整其输出，以确保它完全符合用户的需求和期望。
- 无害性：模型必须优先考虑用户安全，避免任何可能导致非法、不道德或有害行为的内容。它应始终促进道德行为和负责任的沟通。

- **Debiasing:** The model should produce responses that are free from bias, including but not limited to gender, race, nationality, and politics. It should treat all topics equally and fairly, adhering to widely accepted moral and ethical standards.

The queries utilized to train the reward model are drawn from two distinct datasets: publicly available open-source data and a proprietary query set characterized by higher complexity. Responses are generated from checkpoints of the Qwen models, which have been fine-tuned using different methods—SFT, DPO, and RL—at various stages of training. To introduce diversity, those responses are sampled at different temperature settings. Preference pairs are created through both human and automated labeling processes, and the training data for DPO is also integrated into this dataset.

In our online reinforcement learning (RL) framework, we employ Group Relative Policy Optimization (GRPO, Shao et al., 2024). The query set utilized for training the reward model is identical to the one used in the RL training phase. The sequence in which queries are processed during training is determined by the variance of their response scores, as evaluated by the reward model. Specifically, queries with higher variance in response scores are prioritized to ensure more effective learning. We sample 8 responses for each query. All models are trained with a 2048 global batch size and 2048 samples in each episode, considering a pair of queries and responses as a sample.

4.4 Long Context Fine-tuning

To further extend the context length of Qwen2.5-Turbo, we introduce longer SFT examples during post-training, enabling it to better align with human preference in long queries.

In the SFT phase, we employ a two-stage approach. In the first stage, the model is fine-tuned exclusively using short instructions, each containing up to 32,768 tokens. This stage uses the same data and training steps as those employed for the other Qwen2.5 models, ensuring strong performance on short tasks. In the second stage, the fine-tuning process combines both short instructions (up to 32,768 tokens) and long instructions (up to 262,144 tokens). This hybrid approach effectively enhances the model's instruction-following ability in long context tasks while maintaining its performance on short tasks.

During the RL stage, we use a training strategy similar to that used for the other Qwen2.5 models, focusing solely on short instructions. This design choice is driven by two primary considerations: first, RL training is computationally expensive for long context tasks; second, there is currently a scarcity of reward models that provide suitable reward signals for long context tasks. Additionally, we find that adopting RL on short instructions alone can still significantly enhance the model's alignment with human preferences in long context tasks.

5 Evaluation

The base models produced by pre-training and the instruction-tuned models produced by post-training are evaluated accordingly with a comprehensive evaluation suite, including both commonly-used open benchmarks and skill-oriented in-house datasets. The evaluation suite is designed to be primarily automatic with minimal human interaction.

To prevent test data leakage, we exclude potentially contaminated data using n-gram matching when constructing the pre-training and post-training datasets. Following the criteria used in Qwen2, a training sequence $\mathbf{s}_t$ is removed from the training data if there exists a test sequence $\mathbf{s}_e$ such that the length of the longest common subsequence (LCS) between tokenized $\mathbf{s}_t$ and $\mathbf{s}_e$ satisfies both $|\text{LCS}(\mathbf{s}_t, \mathbf{s}_e)| \geq 13$ and $|\text{LCS}(\mathbf{s}_t, \mathbf{s}_e)| \geq 0.6 \times \min(|\mathbf{s}_t|, |\mathbf{s}_e|)$.

5.1 Base Models

We conduct comprehensive evaluations of the base language models of the Qwen2.5 series. The evaluation of base models primarily emphasizes their performance in natural language understanding, general question answering, coding, mathematics, scientific knowledge, reasoning, and multilingual capabilities.

The evaluation datasets include:

General Tasks MMLU (Hendrycks et al., 2021a) (5-shot), MMLU-Pro (Wang et al., 2024b) (5-shot), MMLU-redux (Gema et al., 2024) (5-shot), BBH (Suzgun et al., 2023) (3-shot), ARC-C (Clark et al., 2018) (25-shot), TruthfulQA (Lin et al., 2022a) (0-shot), Winogrande (Sakaguchi et al., 2021) (5-shot), HellaSwag (Zellers et al., 2019) (10-shot).

- 去偏见：模型应该产生没有偏见的反应，包括但不限于性别、种族、国籍和政治。它应该平等和公平地对待所有主题，遵守广泛接受的道德和伦理标准。

用于训练奖励模型的查询来自两个不同的数据集：公开可用的开源数据和具有较高复杂性的专有查询集。响应是从 Qwen 模型的检查点生成的，这些模型在训练的各个阶段使用不同的方法（SFT、DPO 和 RL）进行了微调。为了引入多样性，这些响应在不同的温度设置下进行采样。偏好对是通过人工和自动标记流程创建的，DPO 的训练数据也集成到该数据集中。

在我们的在线强化学习（RL）框架中，我们采用了组相对策略优化（GRPO，Shao 等人，2024）。用于训练奖励模型的查询集与 RL 训练阶段使用的查询集相同。训练期间处理查询的顺序由响应分数的方差决定，并由奖励模型评估。具体来说，优先考虑响应分数差异较大的查询，以确保更有效的学习。我们为每个查询抽取 8 个响应样本。所有模型均使用 2048 个全局批量大小和每集中 2048 个样本进行训练，并将一对查询和响应作为样本。

4.4 长上下文微调

为了进一步扩展 Qwen2.5-Turbo 的上下文长度，我们在训练后引入了更长的 SFT 示例，使其能够更好地符合人类在长查询中的偏好。

在 SFT 阶段，我们采用两阶段方法。在第一阶段，模型仅使用短指令进行微调，每个指令最多包含 32,768 个令牌。此阶段使用与其他 Qwen2.5 模型相同的数据和训练步骤，确保在短任务上具有强大的性能。在第二阶段，微调过程结合了短指令（最多 32,768 个令牌）和长指令（最多 262,144 个令牌）。这种混合方法有效地增强了模型在长上下文任务中的指令跟踪能力，同时保持其在短任务上的性能。

在 RL 阶段，我们使用与其他 Qwen2.5 模型类似的训练策略，仅关注简短指令。这种设计选择是由两个主要考虑因素驱动的：首先，对于长上下文任务，强化学习训练的计算成本很高；其次，目前缺乏为长上下文任务提供合适奖励信号的奖励模型。此外，我们发现仅在短指令上采用强化学习仍然可以显着增强模型在长上下文任务中与人类偏好的一致性。

5 评价

通过综合评估套件对预训练生成的基础模型和训练后生成的指令调整模型进行相应的评估，包括常用的开放基准和面向技能的内部数据集。该评估套件被设计为主要是自动的，以最少的人机交互。

为了防止测试数据泄漏，我们在构建训练前和训练后数据集时使用 n-gram 匹配排除潜在污染的数据。遵循 Qwen2 中使用的标准，如果存在测试序列 s_e，使得标记化的 s_t 和 s_e 之间的最长公共子序列 (LCS) 的长度满足 $|\text{LCS}(s_t, s_e)| \geq 13$ 和 $|\text{LCS}(s_t, s_e)| \geq 0.6 \times \text{分钟}(|s_t|, |s_e|)$。

5.1 基础模型

我们对Qwen2.5系列的基础语言模型进行综合评估。基础模型的评估主要强调其在自然语言理解、一般问题回答、编码、数学、科学知识、推理和多语言能力方面的表现。

评估数据集包括：

一般任务 MMLU (Hendrycks et al., 2021a) (5-shot)、MMLU-Pro (Wang et al., 2024b) (5-shot)、MMLU-redux (Gema et al., 2024) (5-shot)、BBH (Suzgun et al., 2023) (3-shot)、ARC-C (Clark et al., 2018）（25 个镜头）、TruthfulQA（Lin 等人，2022a）（0 个镜头）、Winogrande（Sakaguchi 等人，2021）（5 个镜头）、HellaSwag（Zellers 等人，2019）（10 个镜头）。

Table 2: **Performance of the 70B+ base models and Qwen2.5-Plus.**

Datasets	Llama-3-70B	Mixtral-8x22B	Llama-3-405B	Qwen2-72B	Qwen2.5-72B	Qwen2.5-Plus
General Tasks						
MMLU	79.5	77.8	85.2	84.2	**86.1**	85.4
MMLU-Pro	52.8	51.6	61.6	55.7	58.1	**64.0**
MMLU-redux	75.0	72.9	-	80.5	**83.9**	82.8
BBH	81.0	78.9	85.9	82.4	**86.3**	85.8
ARC-C	68.8	70.7	-	68.9	**72.4**	70.9
TruthfulQA	45.6	51.0	-	54.8	**60.4**	55.3
WindoGrande	85.3	85.0	**86.7**	85.1	83.9	85.5
HellaSwag	88.0	88.7	-	87.3	87.6	**89.2**
Mathematics & Science Tasks						
GPQA	36.3	34.3	-	37.4	**45.9**	43.9
TheoremQA	32.3	35.9	-	42.8	42.4	**48.5**
MATH	42.5	41.7	53.8	50.9	62.1	**64.4**
MMLU-stem	73.7	71.7	-	79.6	**82.7**	81.2
GSM8K	77.6	83.7	89.0	89.0	91.5	**93.0**
Coding Tasks						
HumanEval	48.2	46.3	**61.0**	64.6	59.1	59.1
HumanEval+	42.1	40.2	-	**56.1**	51.2	52.4
MBPP	70.4	71.7	73.0	76.9	**84.7**	79.7
MBPP+	58.4	58.1	-	63.9	**69.2**	66.9
MultiPL-E	46.3	46.7	-	59.6	60.5	**61.0**
Multilingual Tasks						
Multi-Exam	70.0	63.5	-	76.6	**78.7**	78.5
Multi-Understanding	79.9	77.7	-	80.7	**89.6**	89.2
Multi-Mathematics	67.1	62.9	-	76.0	76.7	**82.4**
Multi-Translation	38.0	23.3	-	37.8	39.0	**40.4**

Mathematics & Science Tasks GPQA (Rein et al., 2023) (5-shot), Theorem QA (Chen et al., 2023a) (5-shot), GSM8K (Cobbe et al., 2021) (4-shot), MATH (Hendrycks et al., 2021b) (4-shot).

Coding Tasks HumanEval (Chen et al., 2021) (0-shot), HumanEval+ (Liu et al., 2023)(0-shot), MBPP (Austin et al., 2021) (0-shot), MBPP+ (Liu et al., 2023) (0-shot), MultiPL-E (Cassano et al., 2023) (0-shot) (Python, C++, JAVA, PHP, TypeScript, C#, Bash, JavaScript).

Multilingual Tasks We group them into four categories: (a) Exam: M3Exam (5-shot, we only choose examples that require no image), IndoMMLU (Koto et al., 2023) (3-shot), ruMMLU (Fenogenova et al., 2024) (5-shot), and translated MMLU (Chen et al., 2023b) (5-shot on Arabic, Spanish, French, Portuguese, German, Italian, Japanese, and Korean); (b) Understanding: BELEBELE (Bandarkar et al., 2023) (5-shot), XCOPA (Ponti et al., 2020) (5-shot), XWinograd (Muennighoff et al., 2023) (5-shot), XStoryCloze (Lin et al., 2022b) (0-shot) and PAWS-X (Yang et al., 2019) (5-shot); (c) Mathematics: MGSM (Goyal et al., 2022) (8-shot CoT); and (d) Translation: Flores-101 (Goyal et al., 2022) (5-shot).

For base models, we compare Qwen2.5 models with Qwen2 models and other leading open-weight models in terms of scales of parameters.

Qwen2.5-72B & Qwen2.5-Plus We compare the base models of Qwen2.5-72B and Qwen2.5-Plus to other leading open-weight base models: Llama3-70B (Dubey et al., 2024), Llama3-405B (Dubey et al., 2024), Mixtrail-8x22B (Jiang et al., 2024a), and our previous 72B version, the Qwen2-72B (Yang et al., 2024a). The Qwen2.5-72B base model significantly outperforms its peers in the same category across a wide range of tasks. It achieves results comparable to Llama-3-405B while utilizing only one-fifth of the parameters. Furthermore, when compared to its predecessor, Qwen2-72B, the Qwen2.5-72B shows marked improvements in nearly all benchmark evaluations, particularly excelling in general tasks, mathematics, and coding challenges. With significantly lower training and inference costs, Qwen2.5-Plus achieves very competitive performance results compared to Qwen2.5-72B and Llama3-405B, outperforming other baseline models on the Hellaswag, TheoremQA, MATH, GSM8K, MultiPL-E, Multi-Mathematics, and Multi-Translation. Moreover, Qwen2.5-Plus achieves 64.0 on MMLU-Pro, which is 5.9 points higher than Qwen2.5-72B.

Qwen2.5-14B/32B & Qwen2.5-Turbo The evaluation of the Qwen2.5-Turbo, Qwen2.5-14B, and 32B models is compared against baselines of similar sizes. These baselines include Yi-1.5-34B (Young et al.,

表 2：70B+ 基本型号和 Qwen2.5-Plus 的性能。

Datasets	Llama-3-70B	Mixtral-8x22B	Llama-3-405B	Qwen2-72B	Qwen2.5-72B	Qwen2.5-Plus
General Tasks						
MMLU	79.5	77.8	85.2	84.2	**86.1**	85.4
MMLU-Pro	52.8	51.6	61.6	55.7	58.1	**64.0**
MMLU-redux	75.0	72.9	-	80.5	**83.9**	82.8
BBH	81.0	78.9	85.9	82.4	**86.3**	85.8
ARC-C	68.8	70.7	-	68.9	**72.4**	70.9
TruthfulQA	45.6	51.0	-	54.8	**60.4**	55.3
WindoGrande	85.3	85.0	**86.7**	85.1	83.9	85.5
HellaSwag	88.0	88.7	-	87.3	87.6	**89.2**
Mathematics & Science Tasks						
GPQA	36.3	34.3	-	37.4	**45.9**	43.9
TheoremQA	32.3	35.9	-	42.8	42.4	**48.5**
MATH	42.5	41.7	53.8	50.9	62.1	**64.4**
MMLU-stem	73.7	71.7	-	79.6	**82.7**	81.2
GSM8K	77.6	83.7	89.0	89.0	91.5	**93.0**
Coding Tasks						
HumanEval	48.2	46.3	**61.0**	64.6	59.1	59.1
HumanEval+	42.1	40.2	-	**56.1**	51.2	52.4
MBPP	70.4	71.7	73.0	76.9	**84.7**	79.7
MBPP+	58.4	58.1	-	63.9	**69.2**	66.9
MultiPL-E	46.3	46.7	-	59.6	60.5	**61.0**
Multilingual Tasks						
Multi-Exam	70.0	63.5	-	76.6	**78.7**	78.5
Multi-Understanding	79.9	77.7	-	80.7	**89.6**	89.2
Multi-Mathematics	67.1	62.9	-	76.0	76.7	**82.4**
Multi-Translation	38.0	23.3	-	37.8	39.0	**40.4**

数学和科学任务 GPQA（Rein 等人，2023）（5 个镜头）、定理 QA（Chen 等人，2023a）（5 个镜头）、GSM8K（Cobbe 等人，2021）（4 个镜头）、MATH（Hendrycks 等人，2021b）（4 个镜头）。

编码任务 HumanEval (Chen et al., 2021) (0-shot)、HumanEval+ (Liu et al., 2023)(0-shot)、MBPP (Austin et al., 2021) (0-shot)、MBPP+ (Liu et al., 2023) (0-shot)、MultiPL-E (Cassano et al.) al., 2023）（0-shot）（Python、C++、JAVA、PHP、TypeScript、C#、Bash、JavaScript）。

多语言任务我们将它们分为四类：（a）考试：M3Exam（5-shot，我们只选择不需要图像的示例）、Indo MMLU（Koto et al., 2023）（3-shot）、ruMMLU（Fenogenova et al., 2024）（5-shot）和翻译的 MMLU（Chen et al., 2023b）（5-shot 包括阿拉伯语、西班牙语、法语、葡萄牙语、德语、意大利语、日语和韩语）； (b) 理解：BELEBELE（Bandarkar 等人，2023）（5 次）、XCOPA（Ponti 等人，2020）（5 次）、XWinograd（Muennighoff 等人，2023）（5 次）、XStoryCloze（Lin 等人，2022b）（0 次）和 PAWS-X（Yang 等人，2019）（5 次）； (c) 数学：MGSM（Goyal 等人，2022）（8-shot CoT）； (d) 翻译：Flores-101（Goyal 等人，2022）（5 次）。

对于基础模型，我们将 Qwen2.5 模型与 Qwen2 模型和其他领先的开放权重模型在参数规模方面进行比较。

Qwen2.5-72B 和 Qwen2.5-Plus 我们将 Qwen2.5-72B 和 Qwen2.5-Plus 的基础模型与其他领先的开放权重基础模型进行比较：Llama3-70B（Dubey 等人，2024）、Llama3-405B（Dubey 等人，2024）、Mixtrail-8x22B（Jiang 等人，2024a），以及我们之前的 72B 版本 Qwen2-72B（Yang 等人，2024a）。Qwen2.5-72B 基础模型在广泛的任务中显着优于同类同类产品。它仅使用五分之一的参数即可实现与 Llama-3-405B 相当的结果。此外，与前身 Qwen2-72B 相比，Qwen2.5-72B 在几乎所有基准评估中都显示出显着改进，特别是在一般任务、数学和编码挑战方面表现出色。与 Qwen2.5-72B 和 Llama3-405B 相比，Qwen2.5-Plus 的训练和推理成本显着降低，取得了非常有竞争力的性能结果，在 Hellaswag、TheoremQA、MATH、GSM8K、MultiPL-E、Multi-Mathematics 和 Multi-Translation 上优于其他基线模型。而且，Qwen2.5-Plus在MMLU-Pro上取得了64.0的成绩，比Qwen2.5-72B高出5.9分。

Qwen2.5-14B/32B 和 Qwen2.5-Turbo Qwen2.5-Turbo、Qwen2.5-14B 和 32B 模型的评估与相似尺寸的基线进行比较。这些基线包括 Yi-1.5-34B（Young 等人，

Table 3: **Performance of the 14B-30B+ base models and Qwen2.5-Turbo.**

Datasets	Qwen1.5-32B	Gemma2-27B	Yi-1.5-34B	Qwen2.5-Turbo	Qwen2.5-14B	Qwen2.5-32B
General Tasks						
MMLU	74.3	75.2	77.2	79.5	79.7	**83.3**
MMLU-pro	44.1	49.1	48.3	**55.6**	51.2	55.1
MMLU-redux	69.0	-	74.1	77.1	76.6	**82.0**
BBH	66.8	74.9	76.4	76.1	78.2	**84.5**
ARC-C	63.6	**71.4**	65.6	67.8	67.3	70.4
TruthfulQA	57.4	40.1	53.9	56.3	**58.4**	57.8
Winogrande	81.5	59.7	**84.9**	81.1	81.0	82.0
Hellaswag	85.0	**86.4**	85.9	85.0	84.3	85.2
Mathematics & Science Tasks						
GPQA	30.8	34.9	37.4	41.4	32.8	**48.0**
Theoremqa	28.8	35.8	40.0	42.1	43.0	**44.1**
MATH	36.1	42.7	41.7	55.6	55.6	**57.7**
MMLU-stem	66.5	71.0	72.6	77.0	76.4	**80.9**
GSM8K	78.5	81.1	81.7	88.3	90.2	**92.9**
Coding Tasks						
HumanEval	43.3	54.9	46.3	57.3	56.7	**58.5**
HumanEval+	40.2	46.3	40.2	51.2	51.2	**52.4**
MBPP	64.2	75.7	65.5	76.2	76.7	**84.5**
MBPP+	53.9	60.2	55.4	63.0	63.2	**67.2**
MultiPL-E	38.5	48.0	39.5	53.9	53.5	**59.4**
Multilingual Tasks						
Multi-Exam	61.6	65.8	58.3	70.3	70.6	**75.4**
Multi-Understanding	76.5	82.2	73.9	85.3	85.9	**88.4**
Multi-Mathematics	56.1	61.6	49.3	71.3	68.5	**73.7**
Multi-Translation	33.5	**38.7**	30.0	36.8	36.2	37.3

Table 4: **Performance of the 7B+ base models.**

Datasets	Mistral-7B	Llama3-8B	Gemma2-9B	Qwen2-7B	Qwen2.5-7B
General Tasks					
MMLU	64.2	66.6	71.3	70.3	**74.2**
MMLU-pro	30.9	35.4	44.7	40.1	**45.0**
MMLU-redux	58.1	61.6	67.9	68.1	**71.1**
BBH	56.1	57.7	68.2	62.3	**70.4**
ARC-C	60.0	59.3	**68.2**	60.6	63.7
TruthfulQA	42.2	44.0	45.3	54.2	**56.4**
Winogrande	78.4	77.4	**79.5**	77.0	75.9
HellaSwag	**83.3**	82.1	81.9	80.7	80.2
Mathematics & Science Tasks					
GPQA	24.7	25.8	32.8	30.8	**36.4**
TheoremQA	19.2	22.1	28.9	29.6	**36.0**
MATH	10.2	20.5	37.7	43.5	**49.8**
MMLU-stem	50.1	55.3	65.1	64.2	**72.3**
GSM8K	36.2	55.3	70.7	80.2	**85.4**
Coding Tasks					
HumanEval	29.3	33.5	37.8	51.2	**57.9**
HumanEval+	24.4	29.3	30.5	43.3	**50.6**
MBPP	51.1	53.9	62.2	64.2	**74.9**
MBPP+	40.9	44.4	50.6	51.9	**62.9**
MultiPL-E	29.4	22.6	34.9	41.0	**50.3**
Multilingual Tasks					
Multi-Exam	47.1	52.3	**61.2**	59.2	59.4
Multi-Understanding	63.3	68.6	78.3	72.0	**79.3**
Multi-Mathematics	26.3	36.3	53.0	57.5	**57.8**
Multi-Translation	23.3	31.9	**36.5**	31.5	32.4

表 3：14B-30B+ 基本型号和 Qwen2.5-Turbo 的性能。

Datasets	Qwen1.5-32B	Gemma2-27B	Yi-1.5-34B	Qwen2.5-Turbo	Qwen2.5-14B	Qwen2.5-32B
General Tasks						
MMLU	74.3	75.2	77.2	79.5	79.7	**83.3**
MMLU-pro	44.1	49.1	48.3	**55.6**	51.2	55.1
MMLU-redux	69.0	-	74.1	77.1	76.6	**82.0**
BBH	66.8	74.9	76.4	76.1	78.2	**84.5**
ARC-C	63.6	**71.4**	65.6	67.8	67.3	70.4
TruthfulQA	57.4	40.1	53.9	56.3	**58.4**	57.8
Winogrande	81.5	59.7	**84.9**	81.1	81.0	82.0
Hellaswag	85.0	**86.4**	85.9	85.0	84.3	85.2
Mathematics & Science Tasks						
GPQA	30.8	34.9	37.4	41.4	32.8	**48.0**
Theoremqa	28.8	35.8	40.0	42.1	43.0	**44.1**
MATH	36.1	42.7	41.7	55.6	55.6	**57.7**
MMLU-stem	66.5	71.0	72.6	77.0	76.4	**80.9**
GSM8K	78.5	81.1	81.7	88.3	90.2	**92.9**
Coding Tasks						
HumanEval	43.3	54.9	46.3	57.3	56.7	**58.5**
HumanEval+	40.2	46.3	40.2	51.2	51.2	**52.4**
MBPP	64.2	75.7	65.5	76.2	76.7	**84.5**
MBPP+	53.9	60.2	55.4	63.0	63.2	**67.2**
MultiPL-E	38.5	48.0	39.5	53.9	53.5	**59.4**
Multilingual Tasks						
Multi-Exam	61.6	65.8	58.3	70.3	70.6	**75.4**
Multi-Understanding	76.5	82.2	73.9	85.3	85.9	**88.4**
Multi-Mathematics	56.1	61.6	49.3	71.3	68.5	**73.7**
Multi-Translation	33.5	**38.7**	30.0	36.8	36.2	37.3

表 4：7B+ 基本模型的性能。

Datasets	Mistral-7B	Llama3-8B	Gemma2-9B	Qwen2-7B	Qwen2.5-7B
General Tasks					
MMLU	64.2	66.6	71.3	70.3	**74.2**
MMLU-pro	30.9	35.4	44.7	40.1	**45.0**
MMLU-redux	58.1	61.6	67.9	68.1	**71.1**
BBH	56.1	57.7	68.2	62.3	**70.4**
ARC-C	60.0	59.3	**68.2**	60.6	63.7
TruthfulQA	42.2	44.0	45.3	54.2	**56.4**
Winogrande	78.4	77.4	**79.5**	77.0	75.9
HellaSwag	**83.3**	82.1	81.9	80.7	80.2
Mathematics & Science Tasks					
GPQA	24.7	25.8	32.8	30.8	**36.4**
TheoremQA	19.2	22.1	28.9	29.6	**36.0**
MATH	10.2	20.5	37.7	43.5	**49.8**
MMLU-stem	50.1	55.3	65.1	64.2	**72.3**
GSM8K	36.2	55.3	70.7	80.2	**85.4**
Coding Tasks					
HumanEval	29.3	33.5	37.8	51.2	**57.9**
HumanEval+	24.4	29.3	30.5	43.3	**50.6**
MBPP	51.1	53.9	62.2	64.2	**74.9**
MBPP+	40.9	44.4	50.6	51.9	**62.9**
MultiPL-E	29.4	22.6	34.9	41.0	**50.3**
Multilingual Tasks					
Multi-Exam	47.1	52.3	**61.2**	59.2	59.4
Multi-Understanding	63.3	68.6	78.3	72.0	**79.3**
Multi-Mathematics	26.3	36.3	53.0	57.5	**57.8**
Multi-Translation	23.3	31.9	**36.5**	31.5	32.4

Table 5: **Performance of the smaller base models.**

Datasets	Qwen2-0.5B	Qwen2.5-0.5B	Qwen2-1.5B	Qwen2.5-1.5B	Gemma2-2.6B	Qwen2.5-3B
General Tasks						
MMLU	44.3	47.5	55.9	60.9	52.2	**65.6**
MMLU-pro	14.7	15.7	21.6	28.5	23.0	**34.6**
MMLU-redux	40.7	45.1	51.8	58.5	50.9	**63.7**
BBH	18.2	20.3	36.5	45.1	41.9	**56.3**
ARC-C	31.0	35.6	43.7	54.7	55.7	**56.5**
TruthfulQA	39.7	40.2	45.9	46.6	36.2	**48.9**
Winogrande	56.9	56.3	65.0	65.0	**71.5**	71.1
Hellaswag	49.1	52.1	67.0	67.9	**74.6**	**74.6**
Mathematics & Science Tasks						
GPQA	**29.8**	24.8	20.7	24.2	25.3	26.3
TheoremQA	9.6	16.0	14.8	22.1	15.9	**27.4**
MATH	11.2	19.5	21.6	35.0	18.3	**42.6**
MMLU-STEM	27.5	39.8	42.7	54.8	45.8	**62.5**
GSM8K	36.4	41.6	46.9	68.5	30.3	**79.1**
Coding Tasks						
HumanEval	22.6	30.5	34.8	37.2	19.5	**42.1**
HumanEval+	18.9	26.8	29.9	32.9	15.9	**36.0**
MBPP	33.1	39.3	46.9	**60.2**	42.1	57.1
MBPP+	27.6	33.8	37.6	**49.6**	33.6	49.4
MultiPL-E	16.3	18.9	27.9	33.1	17.6	**41.2**
Multilingual Tasks						
Multi-Exam	29.4	30.8	43.1	47.9	38.1	**54.6**
Multi-Understanding	40.4	41.0	50.7	65.1	46.8	**76.6**
Multi-Mathematics	7.8	13.5	21.3	37.5	18.2	**48.9**
Multi-Translation	14.1	15.3	23.8	25.0	26.9	**29.3**

2024), Gemma2-27B (Gemma Team et al., 2024), and Qwen1.5-32B (Qwen Team, 2024b). The results are shown in Table 3. The Qwen2.5-14B model demonstrates a solid performance across various tasks, particularly excelling in general tasks like MMLU and BBH, where it achieves scores of 79.7 and 78.2, outcompeting competitors of larger sizes. Meanwhile, Qwen2.5-32B, in particular, showcases exceptional capabilities, often surpassing larger models of similar model sizes. Notably, it outperforms its predecessor Qwen1.5-32B significantly, especially in challenging areas such as mathematics and coding, with notable scores of 57.7 in MATH and 84.5 in MBPP. For Qwen2.5-Turbo, although its training cost and inference cost are significantly smaller than those of Qwen2.5-14B, it achieves comparable results, where its MMLU-Pro score is even better than that of Qwen2.5-32B.

Qwen2.5-7B For 7B-level models, we focus on comparing Qwen2.5-7B with other leading 7B+ models, including Mistral-7B (Jiang et al., 2023a), Llama3-8B (Dubey et al., 2024), Gemma2-9B (Gemma Team et al., 2024), and our predecessor, Qwen2-7B (Yang et al., 2024a). The results can be found in Table 4. Note that the non-embedding parameters of Qwen2-7B and Qwen2.5-7B are only 6.5B, while that of Gemma2-9B is 8.2B. The Qwen2.5-7B model surpasses its predecessors and counterparts in numerous benchmarks, despite having fewer non-embedding parameters. It demonstrates significant improvements across various tasks, achieving 74.2 on general benchmarks like MMLU (Hendrycks et al., 2021a), 49.8 on math challenges such as MATH (Hendrycks et al., 2021b), and 57.9 on coding tasks like HumanEval (Chen et al., 2021).

Qwen2.5-0.5B/1.5B/3B For edge-side models, we compare Qwen2.5-0.5B, 1.5B, and 3B against established baselines: Qwen2-0.5B/1.5B (Yang et al., 2024a) and Gemma2-2.6B (Gemma Team et al., 2024). The results are given in Table 5. Qwen2.5-0.5B, 1.5B, and 3B continue to maintain strong performance across nearly all benchmarks. Notably, the Qwen2.5-0.5B model outperforms the Gemma2-2.6B on various math and coding tasks.

5.2 Instruction-tuned Model

To critically evaluate instruction-tuned models, we adopt a multifaceted approach. Foundational skills and human preferences are assessed using open datasets and benchmarks. Additionally, our detailed in-house evaluations delve deeper into the models'competencies in key areas and multilingualism. A particular focus is placed on assessing long-context capability. The subsequent sections outline the evaluation methods and present the results.

表 5：较小基本型号的性能。

Datasets	Qwen2-0.5B	Qwen2.5-0.5B	Qwen2-1.5B	Qwen2.5-1.5B	Gemma2-2.6B	Qwen2.5-3B
			General Tasks			
MMLU	44.3	47.5	55.9	60.9	52.2	**65.6**
MMLU-pro	14.7	15.7	21.6	28.5	23.0	**34.6**
MMLU-redux	40.7	45.1	51.8	58.5	50.9	**63.7**
BBH	18.2	20.3	36.5	45.1	41.9	**56.3**
ARC-C	31.0	35.6	43.7	54.7	55.7	**56.5**
TruthfulQA	39.7	40.2	45.9	46.6	36.2	**48.9**
Winogrande	56.9	56.3	65.0	65.0	**71.5**	71.1
Hellaswag	49.1	52.1	67.0	67.9	**74.6**	**74.6**
			Mathematics & Science Tasks			
GPQA	**29.8**	24.8	20.7	24.2	25.3	26.3
TheoremQA	9.6	16.0	14.8	22.1	15.9	**27.4**
MATH	11.2	19.5	21.6	35.0	18.3	**42.6**
MMLU-STEM	27.5	39.8	42.7	54.8	45.8	**62.5**
GSM8K	36.4	41.6	46.9	68.5	30.3	**79.1**
			Coding Tasks			
HumanEval	22.6	30.5	34.8	37.2	19.5	**42.1**
HumanEval+	18.9	26.8	29.9	32.9	15.9	**36.0**
MBPP	33.1	39.3	46.9	**60.2**	42.1	57.1
MBPP+	27.6	33.8	37.6	**49.6**	33.6	49.4
MultiPL-E	16.3	18.9	27.9	33.1	17.6	**41.2**
			Multilingual Tasks			
Multi-Exam	29.4	30.8	43.1	47.9	38.1	**54.6**
Multi-Understanding	40.4	41.0	50.7	65.1	46.8	**76.6**
Multi-Mathematics	7.8	13.5	21.3	37.5	18.2	**48.9**
Multi-Translation	14.1	15.3	23.8	25.0	26.9	**29.3**

2024）、Gemma2-27B（Gemma Team 等人，2024）和 Qwen1.5-32B（Qwen Team，2024b）。结果如表 3 所示。Qwen2.5-14B 模型在各种任务中表现出稳定的性能，特别是在 MMLU 和 BBH 等一般任务中表现出色，得分为 79.7 和 78.2，超过了更大尺寸的竞争对手。与此同时，Qwen2.5-32B 尤其展示了卓越的功能，常常超越类似型号尺寸的较大型号。值得注意的是，它的表现显着优于其前身 Qwen1.5-32B，尤其是在数学和编码等具有挑战性的领域，数学成绩高达 57.7，MBPP 成绩高达 84.5。对于Qwen2.5-Turbo，虽然其训练成本和推理成本明显小于Qwen2.5-14B，但取得了相当的结果，其MMLU-Pro得分甚至优于Qwen2.5-32B。

Qwen2.5-7B 对于 7B 级模型，我们重点将 Qwen2.5-7B 与其他领先的 7B+ 模型进行比较，包括 Mistral-7B (Jiang et al., 2023a)、Llama3-8B (Dubey et al., 2024)、Gemma2-9B (Gemma Team et al., 2024) 和我们的前身 Qwen2-7B （Yang 等人，2024a）。结果见表4。注意，Qwen2-7B和Qwen2.5-7B的非嵌入参数仅为6.5B，而Gemma2-9B的非嵌入参数为8.2B。尽管非嵌入参数较少，但 Qwen2.5-7B 模型在许多基准测试中都超越了其前辈和同类模型。它在各种任务上都取得了显着的改进，在 MMLU（Hendrycks 等人，2021a）等一般基准测试中取得了 74.2 分，在 MATH（Hendrycks 等人，2021b）等数学挑战中取得了 49.8 分，在 HumanEval（Chen 等人，2021 年）等编码任务中取得了 57.9 分。

Qwen2.5-0.5B/1.5B/3B 对于边缘模型，我们将 Qwen2.5-0.5B、1.5B 和 3B 与既定基线进行比较：Qwen2-0.5B/1.5B（Yang 等人，2024a）和 Gemma2-2.6B（Gemma Team 等人，2024）。结果如表 5 所示。Qwen2.5-0.5B、1.5B 和 3B 在几乎所有基准测试中继续保持强劲的性能。值得注意的是，Qwen2.5-0.5B 模型在各种数学和编码任务上优于 Gemma2-2.6B。

5.2 指令调整模型

为了批判性地评估指令调整模型，我们采用了多方面的方法。使用开放数据集和基准评估基础技能和人类偏好。此外，我们详细的内部评估深入探讨了模型在关键领域和多语言能力的能力。特别关注评估长上下文能力。随后的部分概述了评估方法并介绍了结果。

Table 6: **Performance of the 70B+ Instruct models and Qwen2.5-Plus.**

Datasets	Llama-3.1-70B	Llama-3.1-405B	Qwen2-72B	Qwen2.5-72B	Qwen2.5-Plus
General Tasks					
MMLU-Pro	66.4	**73.3**	64.4	71.1	72.5
MMLU-redux	83.0	86.2	81.6	**86.8**	86.3
LiveBench 0831	46.6	53.2	41.5	52.3	**54.6**
Mathematics & Science Tasks					
GPQA	46.7	**51.1**	42.4	49.0	49.7
MATH	68.0	73.8	69.0	83.1	**84.7**
GSM8K	95.1	**96.8**	93.2	95.8	96.0
Coding Tasks					
HumanEval	80.5	**89.0**	86.0	86.6	87.8
MBPP	84.2	84.5	80.2	**88.2**	85.5
MultiPL-E	68.2	73.5	69.2	75.1	**77.0**
LiveCodeBench	32.1	41.6	32.2	**55.5**	51.4
Alignment Tasks					
IFEval	83.6	86.0	77.6	84.1	**86.3**
Arena-Hard	55.7	69.3	48.1	81.2	**81.4**
MTbench	8.79	9.08	9.12	**9.35**	9.30

Table 7: **Performance of the 14B-30B+ instruction-tuned models and Qwen2.5-Turbo.**

Datasets	Qwen2-57BA14B	Gemma2-27B	GPT4o-mini	Qwen2.5-Turbo	Qwen2.5-14B	Qwen2.5-32B
General Tasks						
MMLU-Pro	52.8	55.5	63.1	64.5	63.7	**69.0**
MMLU-redux	72.6	75.7	81.5	81.7	80.0	**83.9**
LiveBench 0831	31.1	39.6	43.3	42.3	44.4	**50.7**
Mathematics & Science Tasks						
GPQA	34.3	38.4	40.2	42.3	45.5	**49.5**
MATH	49.1	54.4	70.2	81.1	80.0	**83.1**
GSM8K	85.3	90.4	93.2	93.8	94.8	**95.9**
Coding Tasks						
HumanEval	79.9	78.7	**88.4**	86.6	83.5	**88.4**
MBPP	70.9	81.0	**85.7**	82.8	82.0	84.0
MultiPL-E	66.4	67.4	75.0	73.7	72.8	**75.4**
LiveCodeBench	22.5	-	40.7	37.8	42.6	**51.2**
Alignment Tasks						
IFEval	59.9	77.1	80.4	76.3	**81.0**	79.5
Arena-Hard	17.8	57.5	**74.9**	67.1	68.3	74.5
MTbench	8.55	9.10	-	8.81	8.88	**9.20**

5.2.1 Open Benchmark Evaluation

To comprehensively evaluate the quality of instruction-tuned models, we compile automatic and human evaluation to assess the capabilities and human preference. For the evaluation of basic capabilities, we apply similar datasets in the pre-trained model evaluation, which target on natural language understanding, coding, mathematics, and reasoning. Specifically, we evaluate on MMLU-Pro, MMLU-redux and LiveBench 0831 (White et al., 2024) for general evaluation, GPQA, GSM8K and MATH for science and mathematics, HumanEval, MBPP, MultiPL-E and LiveCodeBench 2305-2409 (Jain et al., 2024) for coding, IFEval (Zhou et al., 2023)[2] for instruction following. Additionally, we assess the performance of human preference alignment and instruction following by evaluating on benchmarks including MT-Bench (Zheng et al., 2023) and Arena-Hard (Li et al., 2024).

Qwen2.5-72B-Instruct & Qwen2.5-Plus As shown in Table 6, we compare Qwen2.5-72B-Instruct and Qwen2.5-Plus to other leading open-weight instrution-tuned models: Llama3.1-70B-Instruct (Dubey

[2]For simplicity, we report the results of the subset *strict-prompt*.

表 6：70B+ Instruct 型号和 Qwen2.5-Plus 的性能。

Datasets	Llama-3.1-70B	Llama-3.1-405B	Qwen2-72B	Qwen2.5-72B	Qwen2.5-Plus
General Tasks					
MMLU-Pro	66.4	**73.3**	64.4	71.1	72.5
MMLU-redux	83.0	86.2	81.6	**86.8**	86.3
LiveBench 0831	46.6	53.2	41.5	52.3	**54.6**
Mathematics & Science Tasks					
GPQA	46.7	**51.1**	42.4	49.0	49.7
MATH	68.0	73.8	69.0	83.1	**84.7**
GSM8K	95.1	**96.8**	93.2	95.8	96.0
Coding Tasks					
HumanEval	80.5	**89.0**	86.0	86.6	87.8
MBPP	84.2	84.5	80.2	**88.2**	85.5
MultiPL-E	68.2	73.5	69.2	75.1	**77.0**
LiveCodeBench	32.1	41.6	32.2	**55.5**	51.4
Alignment Tasks					
IFEval	83.6	86.0	77.6	84.1	**86.3**
Arena-Hard	55.7	69.3	48.1	81.2	**81.4**
MTbench	8.79	9.08	9.12	**9.35**	9.30

表 7：14B-30B+ 指令调整模型和 Qwen2.5-Turbo 的性能。

Datasets	Qwen2-57BA14B	Gemma2-27B	GPT4o-mini	Qwen2.5-Turbo	Qwen2.5-14B	Qwen2.5-32B
General Tasks						
MMLU-Pro	52.8	55.5	63.1	64.5	63.7	**69.0**
MMLU-redux	72.6	75.7	81.5	81.7	80.0	**83.9**
LiveBench 0831	31.1	39.6	43.3	42.3	44.4	**50.7**
Mathematics & Science Tasks						
GPQA	34.3	38.4	40.2	42.3	45.5	**49.5**
MATH	49.1	54.4	70.2	81.1	80.0	**83.1**
GSM8K	85.3	90.4	93.2	93.8	94.8	**95.9**
Coding Tasks						
HumanEval	79.9	78.7	**88.4**	86.6	83.5	**88.4**
MBPP	70.9	81.0	**85.7**	82.8	82.0	84.0
MultiPL-E	66.4	67.4	75.0	73.7	72.8	**75.4**
LiveCodeBench	22.5	-	40.7	37.8	42.6	**51.2**
Alignment Tasks						
IFEval	59.9	77.1	80.4	76.3	**81.0**	79.5
Arena-Hard	17.8	57.5	**74.9**	67.1	68.3	74.5
MTbench	8.55	9.10	-	8.81	8.88	**9.20**

5.2.1 开放基准评估

为了全面评估指令调整模型的质量，我们编写了自动和人工评估来评估能力和人类偏好。对于基本能力的评估，我们在预训练模型评估中应用了类似的数据集，针对自然语言理解、编码、数学和推理。具体来说，我们在 MMLU-Pro、MMLU-redux 和 LiveBench 0831 (White et al., 2024) 上进行一般评估，在 GPQA、GSM8K 和 MATH 上进行科学和数学评估，在 HumanEval、MBPP、MultiPL-E 和 LiveCodeBench 2305-2409 (Jain et al., 2024) 上进行编码评估，在 IFEval (Zhou et al., 2024) 上进行评估。 2023)[2] 用于遵循说明。此外，我们通过评估 MT-Bench（Zheng 等人，2023）和 Arena-Hard（Li 等人，2024）等基准来评估人类偏好调整和指令遵循的表现。

Qwen2.5-72B-Instruct 和 Qwen2.5-Plus 如表 6 所示，我们将 Qwen2.5-72B-Instruct 和 Qwen2.5-Plus 与其他领先的开放重量指令调整模型进行比较：Llama3.1-70B-Instruct (Dubey

[2]For simplicity, we report the results of the subset *strict-prompt*.

Table 8: **Performance of the 7B+ instruction-tuned models.**

Datasets	Gemma2-9B	Llama3.1-8B	Qwen2-7B	Qwen2.5-7B
General Tasks				
MMLU-Pro	52.1	48.3	44.1	**56.3**
MMLU-redux	72.8	67.2	67.3	**75.4**
LiveBench 0831	30.6	26.7	29.2	**35.9**
Mathematics & Science Tasks				
GPQA	32.8	32.8	34.3	**36.4**
MATH	44.3	51.9	52.9	**75.5**
GSM8K	76.7	84.5	85.7	**91.6**
Coding Tasks				
HumanEval	68.9	72.6	79.9	**84.8**
MBPP	74.9	69.6	67.2	**79.2**
MultiPL-E	53.4	50.7	59.1	**70.4**
LiveCodeBench	18.9	8.3	23.9	**28.7**
Alignment Tasks				
IFEval	70.1	**75.9**	54.7	71.2
Arena-Hard	41.6	27.8	25.0	**52.0**
MTbench	8.49	8.23	8.26	**8.75**

Table 9: **Performance comparison of 2B-4B instruction-tuned models.**

Datasets	Gemma2-2B	Phi3.5-Mini	MiniCPM3-4B	Qwen2.5-3B
Non-Emb Params	2.0B	3.6B	4.0B	2.8B
General Tasks				
MMLU-Pro	26.7	**47.5**	43.0	43.7
MMLU-redux	51.9	**67.7**	59.9	64.4
LiveBench 0831	20.1	27.4	**27.6**	26.8
Mathematics & Science Tasks				
GPQA	29.3	27.2	**31.3**	30.3
MATH	26.6	48.5	46.6	**65.9**
GSM8K	63.2	86.2	81.1	**86.7**
Coding Tasks				
HumanEval	68.9	72.6	**74.4**	**74.4**
MBPP	**74.9**	63.2	72.5	72.7
MultiPL-E	30.5	47.2	49.1	**60.2**
LiveCodeBench	5.8	15.8	**23.8**	19.9
Alignment Tasks				
IFEval	51.0	52.1	**68.4**	58.2

et al., 2024), Llama3.1-405B-Instruct (Dubey et al., 2024), and our previous 72B version, Qwen2-72B-Instruct (Yang et al., 2024a). The Qwen2.5-72B-Instruct model delivers exceptional performance, even surpassing the larger Llama-3.1-405B-Instruct in several critical benchmarks including MMLU-redux, MATH, MBPP, MultiPL-E, LiveCodeBench, Arena-Hard and MTBench. Moreover, Qwen2.5-Plus outperforms Qwen2.5-72B-Instruct on 9 out of 13 benchmarks.

Qwen2.5-14B/32B-Instruct & Qwen2.5-Turbo The performance of the Qwen2.5-Turbo, Qwen2.5-14B-Instruct, and Qwen2.5-32B-Instruct models is evaluated and compared against baselines of similar sizes. The baselines include GPT4o-mini, Gemma2-27B-IT (Gemma Team et al., 2024), and Qwen2-57BA14B-Instruct (Yang et al., 2024a). The results are summarized in Table 7. The Qwen2.5-32B-Instruct model exhibits superior performance across most tasks when compared to other models of similar size. Notably, our open-weight Qwen2.5-14B-Instruct model delivers competitive results across all benchmarks, rivaling those of GPT-4o-mini. Despite its significantly lower training and inference costs, the Qwen2.5-Turbo model outperforms Qwen2.5-14B-Instruct on eight out of ten benchmarks. This demonstrates that Qwen2.5-Turbo achieves remarkable efficiency and effectiveness, making it a compelling choice for resource-constrained environments.

表 8：7B+ 指令调整模型的性能。

Datasets	Gemma2-9B	Llama3.1-8B	Qwen2-7B	Qwen2.5-7B
General Tasks				
MMLU-Pro	52.1	48.3	44.1	**56.3**
MMLU-redux	72.8	67.2	67.3	**75.4**
LiveBench 0831	30.6	26.7	29.2	**35.9**
Mathematics & Science Tasks				
GPQA	32.8	32.8	34.3	**36.4**
MATH	44.3	51.9	52.9	**75.5**
GSM8K	76.7	84.5	85.7	**91.6**
Coding Tasks				
HumanEval	68.9	72.6	79.9	**84.8**
MBPP	74.9	69.6	67.2	**79.2**
MultiPL-E	53.4	50.7	59.1	**70.4**
LiveCodeBench	18.9	8.3	23.9	**28.7**
Alignment Tasks				
IFEval	70.1	**75.9**	54.7	71.2
Arena-Hard	41.6	27.8	25.0	**52.0**
MTbench	8.49	8.23	8.26	**8.75**

表 9：2B-4B 指令调整模型的性能比较。

Datasets	Gemma2-2B	Phi3.5-Mini	MiniCPM3-4B	Qwen2.5-3B
Non-Emb Params	2.0B	3.6B	4.0B	2.8B
General Tasks				
MMLU-Pro	26.7	**47.5**	43.0	43.7
MMLU-redux	51.9	**67.7**	59.9	64.4
LiveBench 0831	20.1	27.4	**27.6**	26.8
Mathematics & Science Tasks				
GPQA	29.3	27.2	**31.3**	30.3
MATH	26.6	48.5	46.6	**65.9**
GSM8K	63.2	86.2	81.1	**86.7**
Coding Tasks				
HumanEval	68.9	72.6	**74.4**	**74.4**
MBPP	**74.9**	63.2	72.5	72.7
MultiPL-E	30.5	47.2	49.1	**60.2**
LiveCodeBench	5.8	15.8	**23.8**	19.9
Alignment Tasks				
IFEval	51.0	52.1	**68.4**	58.2

等人，2024），Llama3.1-405B-Instruct（Dubey 等人，2024），以及我们之前的 72B 版本 Qwen2-72B-Instruct（Yang 等人，2024a）。 Qwen2.5-72B-Instruct 模型具有卓越的性能，甚至在多个关键基准测试中超越了较大的 Llama-3.1-405B-Instruct，包括 MMLU-redux、MATH、MBPP、MultiPL-E、LiveCodeBench、Arena-Hard 和 MTBench。此外，Qwen2.5-Plus 在 13 个基准测试中的 9 个上优于 Qwen2.5-72B-Instruct。

Qwen2.5-14B/32B-Instruct 和 Qwen2.5-Turbo 对 Qwen2.5-Turbo、Qwen2.5-14B-Instruct 和 Qwen2.5-32B-Instruct 模型的性能进行了评估，并与类似大小的基线进行比较。基线包括 GPT4o-mini、Gemma2-27B-IT（Gemma Team 等人，2024）和 Qwen2-57BA14B-Instruct（Yang 等人，2024a）。结果总结在表 7 中。与类似大小的其他模型相比，Qwen2.5-32B-Instruct 模型在大多数任务中表现出卓越的性能。值得注意的是，我们的开放重量 Qwen2.5-14B-Instruct 模型在所有基准测试中都提供了具有竞争力的结果，可与 GPT-4o-mini 相媲美。尽管训练和推理成本显着降低，但 Qwen2.5-Turbo 模型在十分之八的基准测试中优于 Qwen2.5-14B-Instruct。这表明Qwen2.5-Turbo实现了显着的效率和效果，使其成为资源有限环境中令人信服的选择。

Table 10: **Performance comparison of 0.5B-1.5B instruction-tuned models.**

Datasets	Qwen2-0.5B	Qwen2.5-0.5B	Qwen2-1.5B	Qwen2.5-1.5B
General Tasks				
MMLU-Pro	14.4	**15.0**	22.9	**32.4**
MMLU-redux	12.9	**24.1**	41.2	**50.7**
LiveBench	7.4	**12.6**	12.4	**18.8**
Mathematics & Science Tasks				
GPQA	23.7	**29.8**	21.2	**29.8**
MATH	13.9	**34.4**	25.3	**55.2**
GSM8K	40.1	**49.6**	61.6	**73.2**
Coding Tasks				
HumanEval	31.1	**35.4**	42.1	**61.6**
MBPP	39.7	**49.6**	44.2	**63.2**
MultiPL-E	20.8	**28.5**	38.5	**50.4**
LiveCodeBench	1.6	**5.1**	4.5	**14.8**
Alignment Tasks				
IFEval	14.6	**27.9**	29.0	**42.5**

Table 11: **Performance Comparison on our in-house English automatic evaluation benchmark.**

Models	IF	Knowledge	Comprehension	Coding	Math	Reasoning
Proprietary LLMs						
GPT-4o-2024-08-06	83.28	68.08	76.51	58.05	52.36	66.45
GPT-4o-2024-11-20	80.06	65.25	79.07	60.19	49.74	67.07
Claude3.5-sonnet-2024-10-22	84.22	74.61	79.02	67.17	48.67	70.20
Qwen2 Series						
Qwen2-0.5B-Instruct	18.33	18.59	30.64	5.42	13.16	32.03
Qwen2-1.5B-Instruct	29.42	29.23	45.81	17.02	20.34	38.86
Qwen2-7B-Instruct	50.47	44.79	58.04	43.04	38.31	50.25
Qwen2-72B-Instruct	76.08	59.49	72.19	48.95	48.07	60.33
Llama-3.1 Series						
Llama-3.1-70B-Instruct	81.33	63.42	69.29	55.96	48.00	63.18
Llama-3.1-405B-Instruct	83.33	67.10	75.55	58.14	47.09	64.74
Qwen2.5 Series						
Qwen2.5-0.5B-Instruct	33.35	30.29	29.78	15.41	26.29	36.13
Qwen2.5-1.5B-Instruct	40.25	41.19	47.69	26.19	40.99	42.23
Qwen2.5-3B-Instruct	60.60	46.11	57.98	41.43	49.38	49.80
Qwen2.5-7B-Instruct	70.01	52.74	62.69	48.41	56.93	54.69
Qwen2.5-14B-Instruct	74.17	59.78	69.11	52.68	59.68	62.51
Qwen2.5-Turbo	72.76	58.56	68.70	54.48	57.77	61.06
Qwen2.5-32B-Instruct	76.79	64.08	71.28	58.90	60.97	65.49
Qwen2.5-72B-Instruct	82.65	66.09	74.43	60.41	59.73	65.90
Qwen2.5-Plus	83.18	68.41	79.35	59.58	62.52	66.92

Other Instruction-tuned Models As illustrated in Table 8, the Qwen2.5-7B-Instruct model significantly outperforms its competitors, Gemma2-9B-IT and Llama3.1-8B-Instruct, across all tasks except IFEval. Notably, Qwen2.5-7B-Instruct exhibits clear advantages in mathematics (MATH: 75.5) and coding (HumanEval: 84.8). For the edge-side instruction models, the Qwen2.5-3B-Instruct model, despite having fewer parameters than both the Phi3.5-mini-instruct (Abdin et al., 2024) and MiniCPM3-4B-Instruct (Hu et al., 2024) models, surpasses them in mathematics and coding tasks, as shown in Table 9. Additionally, it delivers competitive results in language understanding. The Qwen2.5-1.5B-Instruct and Qwen2.5-0.5B-Instruct models have also seen substantial performance improvements over their previous versions, as detailed in Table 10. These enhancements make them particularly well-suited for edge-side applications in highly resource-constrained environments.

表 10：0.5B-1.5B 指令调整模型的性能比较。

Datasets	Qwen2-0.5B	Qwen2.5-0.5B	Qwen2-1.5B	Qwen2.5-1.5B
General Tasks				
MMLU-Pro	14.4	**15.0**	22.9	**32.4**
MMLU-redux	12.9	**24.1**	41.2	**50.7**
LiveBench	7.4	**12.6**	12.4	**18.8**
Mathematics & Science Tasks				
GPQA	23.7	**29.8**	21.2	**29.8**
MATH	13.9	**34.4**	25.3	**55.2**
GSM8K	40.1	**49.6**	61.6	**73.2**
Coding Tasks				
HumanEval	31.1	**35.4**	42.1	**61.6**
MBPP	39.7	**49.6**	44.2	**63.2**
MultiPL-E	20.8	**28.5**	38.5	**50.4**
LiveCodeBench	1.6	**5.1**	4.5	**14.8**
Alignment Tasks				
IFEval	14.6	**27.9**	29.0	**42.5**

表 11：我们内部英语自动评估基准的性能比较。

Models	IF	Knowledge	Comprehension	Coding	Math	Reasoning
Proprietary LLMs						
GPT-4o-2024-08-06	83.28	68.08	76.51	58.05	52.36	66.45
GPT-4o-2024-11-20	80.06	65.25	79.07	60.19	49.74	67.07
Claude3.5-sonnet-2024-10-22	84.22	74.61	79.02	67.17	48.67	70.20
Qwen2 Series						
Qwen2-0.5B-Instruct	18.33	18.59	30.64	5.42	13.16	32.03
Qwen2-1.5B-Instruct	29.42	29.23	45.81	17.02	20.34	38.86
Qwen2-7B-Instruct	50.47	44.79	58.04	43.04	38.31	50.25
Qwen2-72B-Instruct	76.08	59.49	72.19	48.95	48.07	60.33
Llama-3.1 Series						
Llama-3.1-70B-Instruct	81.33	63.42	69.29	55.96	48.00	63.18
Llama-3.1-405B-Instruct	83.33	67.10	75.55	58.14	47.09	64.74
Qwen2.5 Series						
Qwen2.5-0.5B-Instruct	33.35	30.29	29.78	15.41	26.29	36.13
Qwen2.5-1.5B-Instruct	40.25	41.19	47.69	26.19	40.99	42.23
Qwen2.5-3B-Instruct	60.60	46.11	57.98	41.43	49.38	49.80
Qwen2.5-7B-Instruct	70.01	52.74	62.69	48.41	56.93	54.69
Qwen2.5-14B-Instruct	74.17	59.78	69.11	52.68	59.68	62.51
Qwen2.5-Turbo	72.76	58.56	68.70	54.48	57.77	61.06
Qwen2.5-32B-Instruct	76.79	64.08	71.28	58.90	60.97	65.49
Qwen2.5-72B-Instruct	82.65	66.09	74.43	60.41	59.73	65.90
Qwen2.5-Plus	83.18	68.41	79.35	59.58	62.52	66.92

其他指令调整模型 如表 8 所示，Qwen2.5-7B-Instruct 模型在除 IFEval 之外的所有任务中均显着优于其竞争对手 Gemma2-9B-IT 和 Llama3.1-8B-Instruct。值得注意的是，Qwen2.5-7B-Instruct 在数学（MATH：75.5）和编码（HumanEval：84.8）方面表现出明显的优势。对于边缘指令模型，Qwen2.5-3B-Instruct 模型尽管比 Phi3.5-mini-instruct (Abdin et al., 2024) 和 MiniCPM3-4B-Instruct (Hu et al., 2024) 模型的参数少，但在数学和编码任务中超越了它们，如表 9 所示。此外，它在语言理解方面提供了有竞争力的结果。 Qwen2.5-1.5B-Instruct 和 Qwen2.5-0.5B-Instruct 模型也比以前的版本有了显着的性能改进，如表 10 所示。这些增强功能使它们特别适合资源高度受限环境中的边缘应用。

Table 12: **Performance Comparison on our in-house Chinese automatic evaluation benchmark.**

Models	IF	Knowledge	Comprehension	Coding	Math	Reasoning
Proprietary LLMs						
GPT-4o-2024-08-06	42.50	68.55	80.11	61.53	61.74	56.88
GPT-4o-2024-11-20	42.71	71.29	83.04	62.39	66.04	62.04
Claude3.5-sonnet-2024-10-22	49.25	72.09	82.16	66.00	63.71	66.60
Qwen2 Series						
Qwen2-0.5B-Instruct	4.69	40.43	39.13	9.85	14.07	32.73
Qwen2-1.5B-Instruct	6.81	51.54	46.89	14.14	24.57	35.19
Qwen2-7B-Instruct	16.83	65.95	60.30	37.05	50.52	44.96
Qwen2-72B-Instruct	31.98	74.96	75.49	41.57	65.55	58.19
Llama-3.1 Series						
Llama-3.1-70B-Instruct	28.96	57.41	67.24	54.82	41.18	52.42
Llama-3.1-405B-Instruct	30.39	63.79	72.27	60.73	46.05	55.88
Qwen2.5 Series						
Qwen2.5-0.5B-Instruct	6.12	39.13	42.97	9.60	24.03	33.72
Qwen2.5-1.5B-Instruct	7.38	48.68	49.69	22.96	37.30	39.17
Qwen2.5-3B-Instruct	16.50	57.18	62.55	29.88	51.64	39.57
Qwen2.5-7B-Instruct	26.64	65.77	67.55	39.56	61.06	49.70
Qwen2.5-14B-Instruct	26.87	70.28	76.96	49.78	67.01	56.41
Qwen2.5-Turbo	32.94	72.93	74.37	51.92	66.08	53.30
Qwen2.5-32B-Instruct	32.64	74.70	79.46	54.45	67.86	60.19
Qwen2.5-72B-Instruct	37.22	75.86	78.85	56.71	68.39	63.02
Qwen2.5-Plus	46.15	72.07	82.64	58.48	69.96	62.98

5.2.2 In-house Automatic Evaluation

Despite the availability of several open benchmark datasets for evaluation, we believe that these are insufficient to fully capture the capabilities of LLMs. To address this, we have developed a series of in-house datasets designed to assess various aspects of model performance, including knowledge understanding, text generation, coding, and more. These evaluations are conducted in both Chinese and English. In addition, we have specifically evaluated the multilingual performance of instruction-tuned models. The results are summarized in Table 11 for English, Table 12 for Chinese, Table 13 for multilingualism of 70B+ Instruct models, and Table 14 for 7B-14B models, respectively.

English & Chinese Evaluation We compare the performance of Qwen2.5-Instruct models against several leading language models, including GPT-4, Claude3.5-sonnet, Qwen2, and Llama-3.1, across both English and Chinese languages. Our analysis focuses on model size and its impact on performance, as well as how our latest Qwen2.5 series compares to previous iterations and competing models. For smaller models, we observe that the Qwen2.5-0.5B model achieves performance that is on par with or even surpasses the Qwen2-1.5B model. This indicates that the Qwen2.5 series has optimized parameter usage, enabling mid-sized models to achieve similar performance levels to larger models from the previous generation. The Qwen2.5-3B model demonstrates performance that is comparable to the Qwen2-7B model. Notably, the Qwen2.5-32B model exhibits a remarkable improvement over the Qwen2-72B model. Our flagship model, Qwen2.5-72B, further narrows the gap between Qwen and state-of-the-art models like GPT-4 and Claude3.5-sonnet. In particular, Qwen2.5-72B matches or exceeds the performance of Llama-3.1-405B in all metrics except for instruction following. This achievement underscores the competitiveness of Qwen2.5-72B in a wide range of language processing tasks, while also identifying areas for future improvement. Qwen2.5-Plus addresses the previous shortcomings in Chinese instruction following and further enhances its advantages in other areas.

Multilingual Evaluation To comprehensively evaluate the multilingual capabilities of instruction-tuned models, we followed P-MMEval (Zhang et al., 2024) and extended several benchmarks as follows: (1) IFEval (Multilingual): We expanded the IFEval benchmark, originally in English, to include multilingual examples. To ensure language neutrality, we removed instances that contained language-specific content (e.g., "start with letter A"). (2) Knowledge Utilization: to assess the knowledge utilization abilities of the Qwen2.5 series models across multiple languages, we employed five MMLU-like benchmarks (multiple-choice format). These benchmarks include: AMMLU (Arabic), JMMLU (Japanese), KMMLU (Korean), IndoMMLU (Indonesian), and TurkishMMLU (Turkish). Additionally, we evaluated the models' performance on the translated version of the MMLU benchmark (okapi_MMLU), which has been adapted

表 12：我们内部中文自动评估基准的性能比较。

Models	IF	Knowledge	Comprehension	Coding	Math	Reasoning
Proprietary LLMs						
GPT-4o-2024-08-06	42.50	68.55	80.11	61.53	61.74	56.88
GPT-4o-2024-11-20	42.71	71.29	83.04	62.39	66.04	62.04
Claude3.5-sonnet-2024-10-22	49.25	72.09	82.16	66.00	63.71	66.60
Qwen2 Series						
Qwen2-0.5B-Instruct	4.69	40.43	39.13	9.85	14.07	32.73
Qwen2-1.5B-Instruct	6.81	51.54	46.89	14.14	24.57	35.19
Qwen2-7B-Instruct	16.83	65.95	60.30	37.05	50.52	44.96
Qwen2-72B-Instruct	31.98	74.96	75.49	41.57	65.55	58.19
Llama-3.1 Series						
Llama-3.1-70B-Instruct	28.96	57.41	67.24	54.82	41.18	52.42
Llama-3.1-405B-Instruct	30.39	63.79	72.27	60.73	46.05	55.88
Qwen2.5 Series						
Qwen2.5-0.5B-Instruct	6.12	39.13	42.97	9.60	24.03	33.72
Qwen2.5-1.5B-Instruct	7.38	48.68	49.69	22.96	37.30	39.17
Qwen2.5-3B-Instruct	16.50	57.18	62.55	29.88	51.64	39.57
Qwen2.5-7B-Instruct	26.64	65.77	67.55	39.56	61.06	49.70
Qwen2.5-14B-Instruct	26.87	70.28	76.96	49.78	67.01	56.41
Qwen2.5-Turbo	32.94	72.93	74.37	51.92	66.08	53.30
Qwen2.5-32B-Instruct	32.64	74.70	79.46	54.45	67.86	60.19
Qwen2.5-72B-Instruct	37.22	75.86	78.85	56.71	68.39	63.02
Qwen2.5-Plus	46.15	72.07	82.64	58.48	69.96	62.98

5.2.2 内部自动评估

尽管有几个开放的基准数据集可供评估，但我们认为这些不足以充分体现法学硕士的能力。为了解决这个问题，我们开发了一系列内部数据集，旨在评估模型性能的各个方面，包括知识理解、文本生成、编码等。这些评估以中文和英文进行。此外，我们还专门评估了指令调整模型的多语言性能。结果分别总结于表 11（英语）、表 12（中文）、表 13（70B+ Instruct 模型的多语言性）和表 14（7B-14B 模型）。

英语和中文评估 我们将 Qwen2.5-Instruct 模型与几种领先语言模型（包括 GPT-4、Claude3.5-sonnet、Qwen2 和 Llama-3.1）在英语和中文上的性能进行比较。我们的分析重点是模型大小及其对性能的影响，以及我们最新的 Qwen2.5 系列与之前的迭代和竞争模型的比较。对于较小的模型，我们观察到 Qwen2.5-0.5B 模型的性能与 Qwen2-1.5B 模型相当甚至超过。这表明Qwen2.5系列优化了参数使用，使中型机型能够达到与上一代大型机型相似的性能水平。 Qwen2.5-3B 型号的性能与 Qwen2-7B 型号相当。值得注意的是，Qwen2.5-32B 模型比 Qwen2-72B 模型表现出显着的改进。我们的旗舰型号 Qwen2.5-72B 进一步缩小了 Qwen 与 GPT-4 和 Claude3.5-sonnet 等最先进型号之间的差距。特别是，Qwen2.5-72B 在除指令跟踪之外的所有指标中均达到或超过 Llama-3.1-405B 的性能。这一成就强调了 Qwen2.5-72B 在广泛的语言处理任务中的竞争力，同时也确定了未来需要改进的领域。 Qwen2.5-Plus解决了之前中文指令跟随方面的不足，并进一步增强了其在其他方面的优势。

多语言评估为了全面评估指令调整模型的多语言能力，我们遵循 P-MMEval（Zhang et al., 2024）并扩展了几个基准，如下所示：（1）IFEval（多语言）：我们扩展了最初为英语的 IFEval 基准，以包含多语言示例。为了确保语言中立，我们删除了包含特定语言内容的实例（例如“以字母 A 开头”）。（2）知识利用：为了评估Qwen2.5系列模型跨多种语言的知识利用能力，我们采用了五个类似MMLU的基准（多项选择格式）。这些基准包括：AMMLU（阿拉伯语）、JMMLU（日语）、KMMLU（韩语）、IndoMMLU（印度尼西亚语）和土耳其语MMLU（土耳其语）。此外，我们还在 MMLU 基准测试的翻译版本（okapi MMLU）上评估了模型的性能，该基准已进行了调整

Table 13: **Performance of the 70B+ Instruct models on Multilingual Tasks.**

Datasets	Qwen2-72B	Llama3.1-70B	Qwen2.5-32B	Mistral-Large	GPT4o-mini	Qwen2.5-72B
Instruction Following						
IFEval (multilingual)	79.69	80.47	82.68	82.69	85.03	**86.98**
Knowledge						
AMMLU (Arabic)	68.85	70.08	70.44	69.24	69.73	**72.44**
JMMLU (Japanese)	77.37	73.89	76.55	75.77	73.74	**80.56**
KMMLU (Korean)	57.04	53.23	60.75	56.42	56.77	**61.96**
IndoMMLU (Indonesian)	66.31	67.50	66.42	63.21	67.75	**69.25**
TurkishMMLU (Turkish)	69.22	66.89	72.41	64.78	71.19	**76.12**
okapi MMLU (translated)	77.84	76.49	77.16	78.37	73.44	**79.97**
Math Reasoning						
MGSM8K (extended)	82.72	73.31	87.15	**89.01**	87.36	88.16
Cultural Nuances						
BLEnD	25.90	30.49	27.88	33.47	**35.91**	32.48

Table 14: **Performance of the 7B-14B Instruct models on Multilingual Tasks.**

Datasets	Qwen2-7B	Llama3.1-8B	Qwen2.5-7B	Gemma2-9B	Qwen2.5-14B
Instruction Following					
IFEval (multilingual)	51.43	60.68	74.87	**77.47**	77.08
Knowledge					
AMMLU (Arabic)	54.87	54.28	59.78	60.26	**66.81**
JMMLU (Japanese)	57.71	53.26	61.88	64.59	**72.78**
KMMLU (Korean)	43.96	42.28	46.59	46.24	**59.71**
IndoMMLU (Indonesian)	54.05	53.92	56.42	61.73	**65.09**
TurkishMMLU (Turkish)	49.27	45.61	54.28	55.44	**66.85**
okapi MMLU (translated)	60.47	55.18	66.98	46.72	**72.12**
Math Reasoning					
MGSM8K (extended)	56.13	66.05	66.11	78.37	**82.27**
Cultural Nuances					
BLEnD	22.49	19.47	23.66	**28.31**	26.99

into multiple languages from its original English form. (3) MGSM8K (Extended): Building upon the original MGSM8K benchmark, we extended the language support to include Arabic (ar), Korean (ko), Portuguese (pt), and Vietnamese (vi). (4) Cultural Nuances: To evaluate the models' ability to capture cultural nuances, we utilized the BLEnD benchmark (Myung et al., 2024). This benchmark is specifically designed to test LLMs on their understanding of cultural subtleties.

Qwen2.5 exhibits competitive performance in instruction following, multilingual knowledge, and mathematical reasoning, aligning well with models of comparable size. Although it shows notable improvements in capturing cultural nuances relative to its predecessor, Qwen2, there remains potential for further refinement in this domain.

5.2.3 Reward Model

The reward model serves as the cornerstone for guiding RL processes, and thus we conduct a separate evaluation of the reward model used in the Qwen2.5 series. Our assessment benchmarks encompass Reward Bench (Lambert et al., 2024), RMB (Zhou et al., 2024), PPE (Frick et al., 2024b), and an internally collected out-of-domain Chinese human preference benchmark (Human-Preference-Chinese) to provide a comprehensive analysis. For comparison, we included baseline models such as Nemotron-4-340B-Reward (Adler et al., 2024), Llama-3.1-Nemotron-70B-Reward (Wang et al., 2024c), and Athene-RM-70B (Frick et al., 2024a). The results are shown in Table 15. Overall, our findings indicate that Llama-3.1-Nemotron-70B-Reward excels on the Reward Bench, while Athene-RM-70B performs best on the RMB benchmark. The Qwen2.5-RM-72B, leads in both the PPE and Human-Preference-Chinese evaluations, ranking second only to Athene-RM-70B on the RMB and achieving a performance level comparable to

表 13：70B+ Instruct 模型在多语言任务上的性能。

Datasets	Qwen2-72B	Llama3.1-70B	Qwen2.5-32B	Mistral-Large	GPT4o-mini	Qwen2.5-72B
Instruction Following						
IFEval (multilingual)	79.69	80.47	82.68	82.69	85.03	**86.98**
Knowledge						
AMMLU (Arabic)	68.85	70.08	70.44	69.24	69.73	**72.44**
JMMLU (Japanese)	77.37	73.89	76.55	75.77	73.74	**80.56**
KMMLU (Korean)	57.04	53.23	60.75	56.42	56.77	**61.96**
IndoMMLU (Indonesian)	66.31	67.50	66.42	63.21	67.75	**69.25**
TurkishMMLU (Turkish)	69.22	66.89	72.41	64.78	71.19	**76.12**
okapi MMLU (translated)	77.84	76.49	77.16	78.37	73.44	**79.97**
Math Reasoning						
MGSM8K (extended)	82.72	73.31	87.15	**89.01**	87.36	88.16
Cultural Nuances						
BLEnD	25.90	30.49	27.88	33.47	**35.91**	32.48

表 14：7B-14B Instruct 模型在多语言任务上的性能。

Datasets	Qwen2-7B	Llama3.1-8B	Qwen2.5-7B	Gemma2-9B	Qwen2.5-14B
Instruction Following					
IFEval (multilingual)	51.43	60.68	74.87	**77.47**	77.08
Knowledge					
AMMLU (Arabic)	54.87	54.28	59.78	60.26	**66.81**
JMMLU (Japanese)	57.71	53.26	61.88	64.59	**72.78**
KMMLU (Korean)	43.96	42.28	46.59	46.24	**59.71**
IndoMMLU (Indonesian)	54.05	53.92	56.42	61.73	**65.09**
TurkishMMLU (Turkish)	49.27	45.61	54.28	55.44	**66.85**
okapi MMLU (translated)	60.47	55.18	66.98	46.72	**72.12**
Math Reasoning					
MGSM8K (extended)	56.13	66.05	66.11	78.37	**82.27**
Cultural Nuances					
BLEnD	22.49	19.47	23.66	**28.31**	26.99

从最初的英语形式翻译成多种语言。 (3) MGSM8K（扩展）：在原始 MGSM8K 基准的基础上，我们扩展了语言支持，包括阿拉伯语 (ar)、韩语 (ko)、葡萄牙语 (pt) 和越南语 (vi)。 (4) 文化细微差别：为了评估模型捕获文化细微差别的能力，我们使用了 BLenD 基准（Myung 等人，2024）。该基准是专门为测试法学硕士对文化微妙之处的理解而设计的。

Qwen2.5 在指令遵循、多语言知识和数学推理方面表现出竞争性的表现，与同等规模的模型非常吻合。尽管相对于其前身 Qwen2，它在捕捉文化细微差别方面显示出显着的进步，但该领域仍有进一步完善的潜力。

5.2.3 奖励模型

奖励模型是指导强化学习过程的基石，因此我们对Qwen2.5系列中使用的奖励模型进行了单独的评估。我们的评估基准包括Reward Bench（Lambert et al., 2024）、RMB（Zhou et al., 2024）、PPE（Frick et al., 2024b）以及内部收集的域外华人人类偏好基准（Human-Preference-Chinese）来提供全面的分析。为了进行比较，我们纳入了基线模型，例如 Nemotron-4-340B-Reward（Adler 等人，2024）、Llama-3.1-Nemotron-70B-Reward（Wang 等人，2024c）和 Athene-RM-70B（Frick 等人，2024a）。结果如表 15 所示。总体而言，我们的研究结果表明 Llama-3.1-Nemotron-70B-Reward 在奖励基准上表现出色，而 Athene-RM-70B 在人民币基准上表现最佳。 Qwen2.5-RM-72B在PPE和人类偏好中文评估中均领先，在人民币上仅次于Athene-RM-70B，达到了与

Table 15: **Performance comparison across multiple RM benchmarks.**

Metric	Nemotron-4-340B-Reward	Llama-3.1-Nemotron-70B-Reward	Athene-RM -70B	Qwen2.5-RM -72B
		Reward Bench		
Chat	95.80	97.50	**98.32**	97.21
Chat Hard	**87.10**	85.70	70.61	78.73
Safety	91.50	**95.10**	92.10	92.71
Reasoning	93.60	**98.10**	92.19	97.65
Score	92.00	**94.10**	88.32	91.59
		RMB		
Helpfulness (BoN)	48.85	61.02	**67.24**	65.72
Helpfulness (Pairwise)	68.70	75.28	**80.82**	78.83
Harmlessness (BoN)	50.92	52.00	**67.02**	56.35
Harmlessness (Pairwise)	70.84	69.96	**80.83**	73.94
Overall	59.83	64.57	**73.98**	68.71
		PPE		
Human Preference	59.28	64.32	**66.48**	64.80
IFEval	62.66	63.40	62.15	**67.97**
GPQA	56.56	59.14	59.26	**59.80**
MATH	65.12	69.73	79.14	**81.48**
MBPP-Plus	49.15	55.62	**67.97**	64.34
MMLU-Pro	69.69	70.20	**76.95**	75.66
Objective-Avg	60.64	63.62	69.09	**69.85**
		Human-Preference-Chinese		
Accuracy	50.46	59.95	61.11	**61.27**

Nemotron-4-340B-Reward on the Reward Bench, albeit slightly behind Llama-3.1-Nemotron-70B-Reward.

Due to the lack of evaluation methods for reward models, current reward models are typically evaluated using Reward Bench. However, our evaluation results from multiple RM benchmarks suggest that over-optimization on a specific benchmark may trigger Goodhart's law (Hoskin, 1996), resulting in degraded performance on other benchmarks and potentially impacting downstream alignment performance. This highlights the need for comprehensive evaluation of reward models across diverse benchmarks rather than relying solely on a single benchmark.

More importantly, through iterative experimentation, we have also come to recognize a critical limitation: current reward model evaluation benchmarks do not accurately predict the performance of the RL models trained under their guidance. In other words, a higher score on RM benchmarks does not necessarily correlate with superior performance of the resulting RL model. This insight underscores the need for further research into more predictive evaluation methods for reward models.

5.2.4 Long Context Capabilities

We utilize three benchmarks to evaluate long context capabilities of Qwen2.5 models: RULER (Hsieh et al., 2024), LV-Eval (Yuan et al., 2024), and Longbench-Chat (Bai et al., 2024). In LV-Eval, we adopt keyword recall as the reported score to mitigate the high rate of false negatives present in the original metrics.

The results are shown in Table 16 and Table 17. We can observe that the Qwen2.5 models, after equipping length extrapolation techniques (i.e., DCA + YARN), have demonstrated strong long context processing capabilities on the three datasets. Among them, Qwen2.5-72B-Instruct has shown the strongest performance across all context lengths, significantly outperforming existing open-weight long-context models as well as the proprietary models like GPT-4o-mini and GPT-4.

Furthermore, as shown in Figure 2, Qwen2.5-Turbo achieves 100% accuracy in the 1M-token passkey retrieval task, demonstrating its exceptional ability to capture detailed information from ultra-long contexts. We develop a sparse attention mechanism based on Minference (Jiang et al., 2024b) to significantly enhance inference speed, which is critical for user experience when processing long contexts. For sequences of 1M tokens, this approach reduces the computational load of the attention mechanism by 12.5 times. Figure 3 illustrates the time to first token (TTFT) of Qwen2.5-Turbo across various hardware configurations, where our method achieves a 3.2 to 4.3 times speedup.

表 15：多个 RM 基准的性能比较。

Metric	Nemotron-4-340B-Reward	Llama-3.1-Nemotron-70B-Reward	Athene-RM-70B	Qwen2.5-RM-72B
Reward Bench				
Chat	95.80	97.50	**98.32**	97.21
Chat Hard	**87.10**	85.70	70.61	78.73
Safety	91.50	**95.10**	92.10	92.71
Reasoning	93.60	**98.10**	92.19	97.65
Score	92.00	**94.10**	88.32	91.59
RMB				
Helpfulness (BoN)	48.85	61.02	**67.24**	65.72
Helpfulness (Pairwise)	68.70	75.28	**80.82**	78.83
Harmlessness (BoN)	50.92	52.00	**67.02**	56.35
Harmlessness (Pairwise)	70.84	69.96	**80.83**	73.94
Overall	59.83	64.57	**73.98**	68.71
PPE				
Human Preference	59.28	64.32	**66.48**	64.80
IFEval	62.66	63.40	62.15	**67.97**
GPQA	56.56	59.14	59.26	**59.80**
MATH	65.12	69.73	79.14	**81.48**
MBPP-Plus	49.15	55.62	**67.97**	64.34
MMLU-Pro	69.69	70.20	**76.95**	75.66
Objective-Avg	60.64	63.62	69.09	**69.85**
Human-Preference-Chinese				
Accuracy	50.46	59.95	61.11	**61.27**

Nemotron-4-340B-Reward 位于奖励台上，尽管略落后于 Llama-3.1-Nemotron-70B-Reward。

由于缺乏奖励模型的评估方法，目前的奖励模型通常使用Reward Bench进行评估。然而，我们对多个 RM 基准的评估结果表明，对特定基准的过度优化可能会触发 Goodhart 定律（Hoskin，1996），导致其他基准的性能下降，并可能影响下游对齐性能。这凸显了需要对不同基准的奖励模型进行综合评估，而不是仅仅依赖单一基准。

更重要的是，通过迭代实验，我们也认识到一个关键的局限性：当前的奖励模型评估基准无法准确预测在其指导下训练的强化学习模型的性能。换句话说，RM 基准得分较高并不一定与最终的 RL 模型的卓越性能相关。这一见解强调需要进一步研究奖励模型的更具预测性的评估方法。

5.2.4 长上下文能力

我们利用三个基准来评估 Qwen2.5 模型的长上下文能力：RULER (Hsieh et al., 2024)、LV-Eval (Yuan et al., 2024) 和 Longbench-Chat (Bai et al., 2024)。在 LV-Eval 中，我们采用关键字召回作为报告分数，以减轻原始指标中存在的高漏报率。

结果如表16和表17所示。我们可以观察到，Qwen2.5模型在配备长度外推技术（即DCA + YARN）后，在三个数据集上表现出了强大的长上下文处理能力。其中，Qwen2.5-72B-Instruct 在所有上下文长度上都表现出了最强的性能，显着优于现有的开放权重长上下文模型以及 GPT-4o-mini 和 GPT-4 等专有模型。

此外，如图 2 所示，Qwen2.5-Turbo 在 1M 令牌密钥检索任务中实现了 100% 的准确率，展示了其从超长上下文中捕获详细信息的卓越能力。我们开发了一种基于 Minference 的稀疏注意力机制（Jiang et al., 2024b），以显着提高推理速度，这对于处理长上下文时的用户体验至关重要。对于 1M 个 token 的序列，这种方法将注意力机制的计算负载减少了 12.5 倍。图 3 显示了 Qwen2.5-Turbo 在各种硬件配置上的首次令牌时间 (TTFT)，其中我们的方法实现了 3.2 至 4.3 倍的加速。

Table 16: **Performance of Qwen2.5 Models on RULER.** *YARN+DCA* does not change the model behavior within 32K tokens.

Model	Claimed Length	RULER						
		Avg.	4K	8K	16K	32K	64K	128K
GLM4-9b-Chat-1M	1M	89.9	94.7	92.8	92.1	89.9	86.7	83.1
Llama-3-8B-Instruct-Gradient-1048k	1M	88.3	95.5	93.8	91.6	87.4	84.7	77.0
Llama-3.1-70B-Instruct	128K	89.6	96.5	95.8	95.4	94.8	88.4	66.6
GPT-4o-mini	128K	87.3	95.0	92.9	92.7	90.2	87.6	65.8
GPT-4	128K	91.6	96.6	96.3	95.2	93.2	87.0	81.2
Qwen2.5-7B-Instruct	128K	85.4	96.7	95.1	93.7	89.4	82.3	55.1
w/o DCA + YARN		80.1	96.7	95.1	93.7	89.4	74.5	31.4
Qwen2.5-14B-Instruct	128K	91.4	97.7	96.8	95.9	93.4	86.7	78.1
w/o DCA + YARN		86.5	97.7	96.8	95.9	93.4	82.3	53.0
Qwen2.5-32B-Instruct	128K	92.9	96.9	97.1	95.5	95.5	90.3	82.0
w/o DCA + YARN		88.0	96.9	97.1	95.5	95.5	85.3	57.7
Qwen2.5-72B-Instruct	128K	**95.1**	**97.7**	**97.2**	**97.7**	**96.5**	**93.0**	**88.4**
w/o DCA + YARN		90.8	97.7	97.2	97.7	96.5	88.5	67.0
Qwen2.5-Turbo	1M	93.1	97.5	95.7	95.5	94.8	90.8	84.5

Table 17: **Performance of Qwen2.5 Models on LV-Eval and LongBench-Chat.** *YARN+DCA* does not change the model behavior within 32k tokens.

Model	Claimed Length	LV-Eval					LongBench-Chat
		16k	32k	64k	128k	256k	
GLM4-9B-Chat-1M	1M	46.4	43.2	42.9	40.4	37.0	7.82
Llama-3-8B-Instruct-Gradient-1048k	1M	31.7	31.8	28.8	26.3	21.1	6.20
Llama-3.1-70B-Instruct	128k	48.6	47.4	42.9	26.2	N/A	6.80
GPT-4o-mini	128k	52.9	48.1	46.0	40.7	N/A	8.48
Qwen2.5-7B-Instruct	128k	55.9	49.7	48.0	41.1	36.9	7.42
w/o DCA + YARN		55.9	49.7	33.1	13.6	0.5	-
Qwen2.5-14B-Instruct	128k	53.0	50.8	46.8	43.6	39.4	8.04
w/o DCA + YARN		53.0	50.8	37.0	18.4	0.8	-
Qwen2.5-32B-Instruct	128k	56.0	53.6	48.8	45.3	41.0	8.70
w/o DCA + YARN		56.0	53.6	40.1	20.5	0.7	-
Qwen2.5-72B-Instruct	128k	**60.4**	**57.5**	**53.9**	**50.9**	**45.2**	**8.72**
w/o DCA + YARN		60.4	57.5	47.4	27.0	2.4	-
Qwen2.5-Turbo	1M	53.4	50.0	45.4	43.9	38.0	8.34

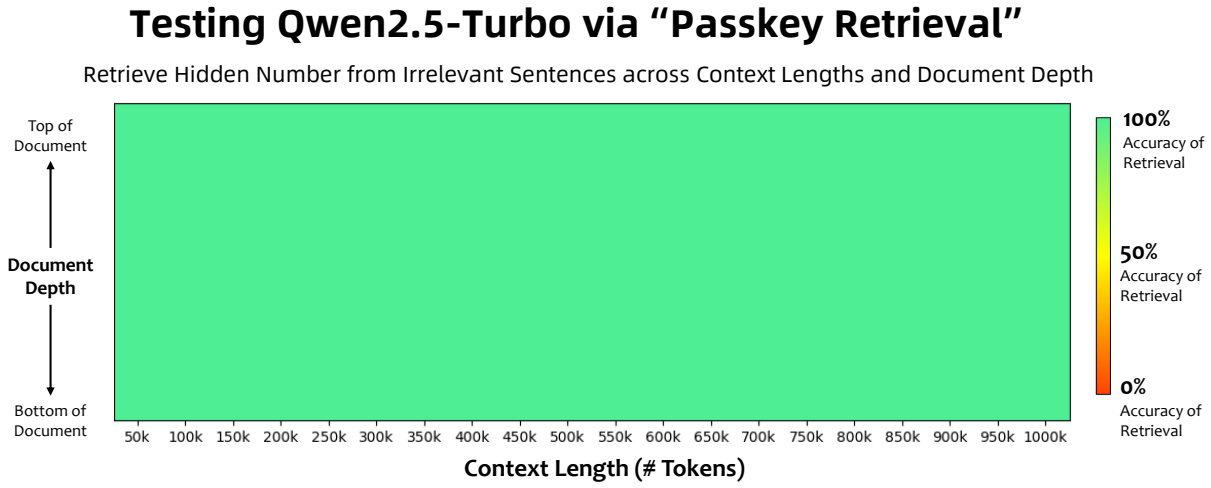

Figure 2: **Performance of Qwen2.5-Turbo on Passkey Retrieval Task with 1M Token Lengths.**

表 16：Qwen2.5 模型在 RULER 上的性能。 *YARN+DCA* 不会更改 32K 令牌内的模型行为。

Model	Claimed Length	RULER						
		Avg.	4K	8K	16K	32K	64K	128K
GLM4-9b-Chat-1M	1M	89.9	94.7	92.8	92.1	89.9	86.7	83.1
Llama-3-8B-Instruct-Gradient-1048k	1M	88.3	95.5	93.8	91.6	87.4	84.7	77.0
Llama-3.1-70B-Instruct	128K	89.6	96.5	95.8	95.4	94.8	88.4	66.6
GPT-4o-mini	128K	87.3	95.0	92.9	92.7	90.2	87.6	65.8
GPT-4	128K	91.6	96.6	96.3	95.2	93.2	87.0	81.2
Qwen2.5-7B-Instruct	128K	85.4	96.7	95.1	93.7	89.4	82.3	55.1
w/o DCA + YARN		80.1	96.7	95.1	93.7	89.4	74.5	31.4
Qwen2.5-14B-Instruct	128K	91.4	97.7	96.8	95.9	93.4	86.7	78.1
w/o DCA + YARN		86.5	97.7	96.8	95.9	93.4	82.3	53.0
Qwen2.5-32B-Instruct	128K	92.9	96.9	97.1	95.5	95.5	90.3	82.0
w/o DCA + YARN		88.0	96.9	97.1	95.5	95.5	85.3	57.7
Qwen2.5-72B-Instruct	128K	**95.1**	**97.7**	**97.2**	**97.7**	**96.5**	**93.0**	**88.4**
w/o DCA + YARN		90.8	97.7	97.2	97.7	96.5	88.5	67.0
Qwen2.5-Turbo	1M	93.1	97.5	95.7	95.5	94.8	90.8	84.5

表 17：Qwen2.5 模型在 LV-Eval 和 LongBench-Chat 上的性能。 *YARN+DCA* 不会更改 32k 令牌内的模型行为。

Model	Claimed Length	LV-Eval					LongBench-Chat
		16k	32k	64k	128k	256k	
GLM4-9B-Chat-1M	1M	46.4	43.2	42.9	40.4	37.0	7.82
Llama-3-8B-Instruct-Gradient-1048k	1M	31.7	31.8	28.8	26.3	21.1	6.20
Llama-3.1-70B-Instruct	128k	48.6	47.4	42.9	26.2	N/A	6.80
GPT-4o-mini	128k	52.9	48.1	46.0	40.7	N/A	8.48
Qwen2.5-7B-Instruct	128k	55.9	49.7	48.0	41.1	36.9	7.42
w/o DCA + YARN		55.9	49.7	33.1	13.6	0.5	-
Qwen2.5-14B-Instruct	128k	53.0	50.8	46.8	43.6	39.4	8.04
w/o DCA + YARN		53.0	50.8	37.0	18.4	0.8	-
Qwen2.5-32B-Instruct	128k	56.0	53.6	48.8	45.3	41.0	8.70
w/o DCA + YARN		56.0	53.6	40.1	20.5	0.7	-
Qwen2.5-72B-Instruct	128k	**60.4**	**57.5**	**53.9**	**50.9**	**45.2**	**8.72**
w/o DCA + YARN		60.4	57.5	47.4	27.0	2.4	-
Qwen2.5-Turbo	1M	53.4	50.0	45.4	43.9	38.0	8.34

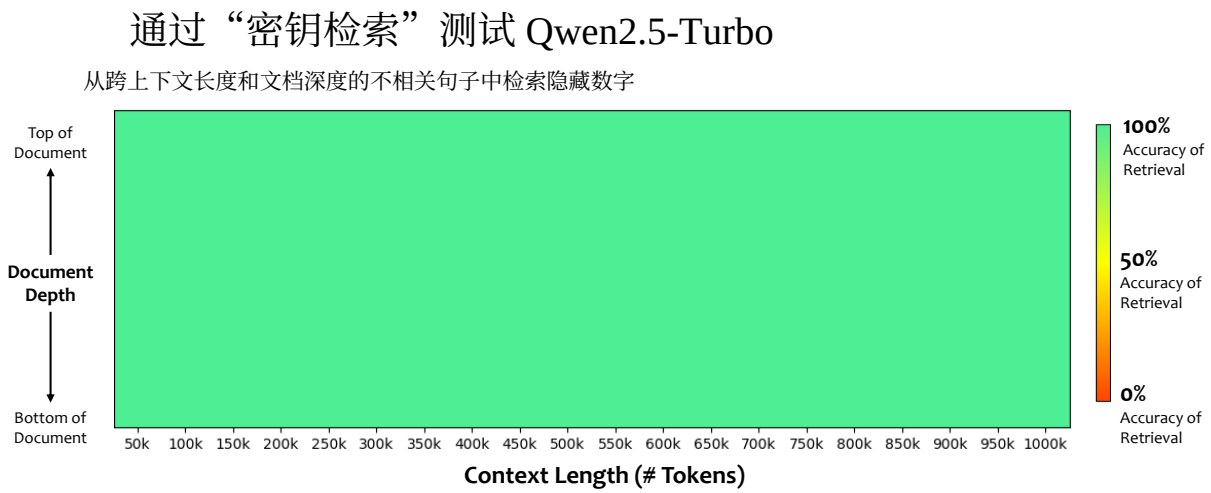

图 2：Qwen2.5-Turbo 在 1M 令牌长度的密钥检索任务上的性能。

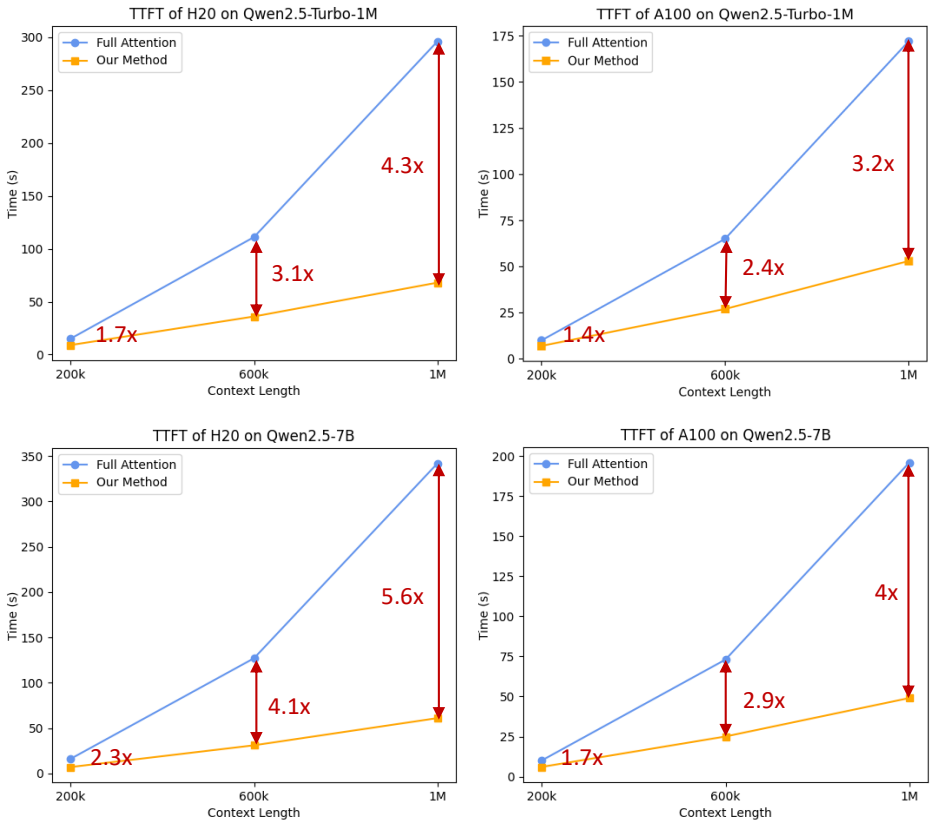

Figure 3: **TTFT (Time To First Token) of Qwen2.5-Turbo and Qwen2.5-7B with Full Attention and Our Method.**

6 Conclusion

Qwen2.5 represents a significant advancement in large language models (LLMs), with enhanced pre-training on 18 trillion tokens and sophisticated post-training techniques, including supervised fine-tuning and multi-stage reinforcement learning. These improvements boost human preference alignment, long text generation, and structural data analysis, making Qwen2.5 highly effective for instruction-following tasks. Available in various configurations, Qwen2.5 offers both open-weight from 0.5B to 72B parameters and proprietary models including cost-effective MoE variants like Qwen2.5-Turbo and Qwen2.5-Plus. Empirical evaluations show that Qwen2.5-72B-Instruct matches the performance of the state-of-the-art Llama-3-405B-Instruct, despite being six times smaller. Qwen2.5 also serves as a foundation for specialized models, demonstrating its versatility for domain-specific applications. We believe that Qwen2.5's robust performance, flexible architecture, and broad availability make it a valuable resource for both academic research and industrial applications, positioning it as a key player of future innovations.

In the future, we will focus on advancing robust foundational models. First, we will iteratively refine both base and instruction-tuned large language models (LLMs) by incorporating broader, more diverse, higher-quality data. Second, we will also continue to develop multimodal models. Our goal is to integrate various modalities into a unified framework. This will facilitate seamless, end-to-end information processing across textual, visual, and auditory domains. Third, we are committed to enhancing the reasoning capabilities of our models. This will be achieved through strategic scaling of inference compute resources. These efforts aim to push the boundaries of current technological limitations and contribute to the broader field of artificial intelligence.

7 Authors

Core Contributors: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia,

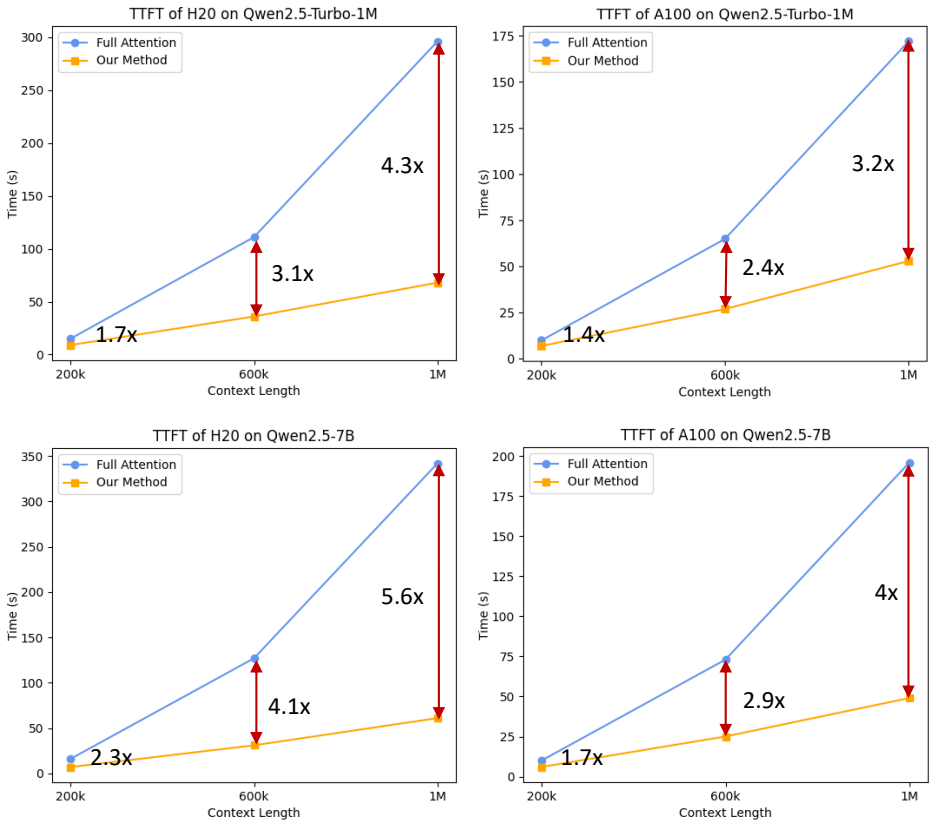

图 3：充分注意的 Qwen2.5-Turbo 和 Qwen2.5-7B 的 TTFT（首次令牌时间）和我们的方法。

6结论

Qwen2.5 代表了大型语言模型 (LLM) 的重大进步，增强了 18 万亿个令牌的预训练和复杂的后训练技术，包括监督微调和多阶段强化学习。这些改进促进了人类偏好对齐、长文本生成和结构数据分析，使得 Qwen2.5 对于指令跟踪任务非常有效。 Qwen2.5 有多种配置可供选择，提供 0.5B 至 72B 参数的开放式重量和专有模型，包括 Qwen2.5-Turbo 和 Qwen2.5-Plus 等经济高效的 MoE 变体。实证评估表明，Qwen2.5-72B-Instruct 的性能与最先进的 Llama-3-405B-Instruct 相当，尽管其尺寸只有其六倍。 Qwen2.5 还可以作为专用模型的基础，展示其针对特定领域应用程序的多功能性。我们相信，Qwen2.5 强大的性能、灵活的架构和广泛的可用性使其成为学术研究和工业应用的宝贵资源，使其成为未来创新的关键参与者。

未来，我们将专注于推进强大的基础模型。首先，我们将通过整合更广泛、更多样化、更高质量的数据，迭代地完善基础和指令调整的大语言模型（LLM）。其次，我们还将继续开发多式联运模式。我们的目标是将各种模式整合到一个统一的框架中。这将促进跨文本、视觉和听觉领域的无缝、端到端信息处理。第三，我们致力于增强模型的推理能力。这将通过推理计算资源的战略扩展来实现。这些努力旨在突破当前技术限制的界限，为更广泛的人工智能领域做出贡献。

7位作者

核心贡献者：杨安、杨宝松、张北辰、许斌源、郑波、余博文、李成远、刘大一恒、黄飞、魏浩然、林焕、杨建、涂建红、张建伟、杨建新、杨家喜、周静仁、林俊阳、党凯、卢克明、包克勤、杨克新、余乐、李美、薛明峰、张培、朱勤、门睿、润吉林、李天浩、唐天一、夏婷玉、

Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu

Contributors: Biao Sun, Bin Luo, Bin Zhang, Binghai Wang, Chaojie Yang, Chang Si, Cheng Chen, Chengpeng Li, Chujie Zheng, Fan Hong, Guanting Dong, Guobin Zhao, Hangrui Hu, Hanyu Zhao, Hao Lin, Hao Xiang, Haoyan Huang, Humen Zhong, Jialin Wang, Jialong Tang, Jiandong Jiang, Jianqiang Wan, Jianxin Ma, Jianyuan Zeng, Jie Zhang, Jin Xu, Jinkai Wang, Jinzheng He, Jun Tang, Ke Yi, Keqin Chen, Langshi Chen, Le Jiang, Lei Zhang, Liang Chen, Man Yuan, Mingkun Yang, Minmin Sun, Na Ni, Nuo Chen, Peng Wang, Peng Zhu, Pengcheng Zhang, Pengfei Wang, Qiaoyu Tang, Qing Fu, Rong Zhang, Ru Peng, Ruize Gao, Shanghaoran Quan, Shen Huang, Shuai Bai, Shuang Luo, Sibo Song, Song Chen, Tao He, Ting He, Wei Ding, Wei Liao, Weijia Xu, Wenbin Ge, Wenbiao Yin, Wenyuan Yu, Xianyan Jia, Xianzhong Shi, Xiaodong Deng, Xiaoming Huang, Ximing Zhou, Xinyu Wang, Xipin Wei, Xuejing Liu, Yang Liu, Yang Yao, Yang Zhang, Yibo Miao, Yidan Zhang, Yikai Zhu, Yinger Zhang, Yong Jiang, Yong Li, Yongan Yue, Yuanzhi Zhu, Yunfei Chu, Zekun Wang, Zhaohai Li, Zheren Fu, Zhi Li, Zhibo Yang, Zhifang Guo, Zhipeng Zhang, Zhiying Xu, Zile Qiao, Ziye Meng

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219, 2024.

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, Aleksander Ficek, Denys Fridman, Shaona Ghosh, Boris Ginsburg, Igor Gitman, Tomasz Grzegorzek, Robert Hero, Jining Huang, Vibhu Jawa, Joseph Jennings, Aastha Jhunjhunwala, John Kamalu, Sadaf Khan, Oleksii Kuchaiev, Patrick LeGresley, Hui Li, Jiwei Liu, Zihan Liu, Eileen Peters Long, Ameya Mahabaleshwarkar, Somshubra Majumdar, James Maki, Miguel Martinez, Maer Rodrigues de Melo, Ivan Moshkov, Deepak Narayanan, Sean Narenthiran, Jesus Navarro, Phong Nguyen, Osvald Nitski, Vahid Noroozi, Guruprasad Nutheti, Christopher Parisien, Jupinder Parmar, Mostofa Patwary, Krzysztof Pawelec, Wei Ping, Shrimai Prabhumoye, Rajarshi Roy, Trisha Saar, Vasanth Rao Naik Sabavat, Sanjeev Satheesh, Jane Polak Scowcroft, Jason D. Sewall, Pavel Shamis, Gerald Shen, Mohammad Shoeybi, Dave Sizer, Misha Smelyanskiy, Felipe Soares, Makesh Narsimhan Sreedhar, Dan Su, Sandeep Subramanian, Shengyang Sun, Shubham Toshniwal, Hao Wang, Zhilin Wang, Jiaxuan You, Jiaqi Zeng, Jimmy Zhang, Jing Zhang, Vivienne Zhang, Yian Zhang, and Chen Zhu. Nemotron-4 340B technical report. *CoRR*, abs/2406.11704, 2024.

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.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon series of open language models. *CoRR*, abs/2311.16867, 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. Introducing Claude, 2023a. URL https://www.anthropic.com/index/introducing-claude.

Anthropic. Claude 2. Technical report, Anthropic, 2023b. URL https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

任兴章、任宣成、范杨、苏杨、张一昌、万宇、刘玉琼、崔泽宇、张振如、邱子涵

贡献者：孙彪、罗斌、张斌、王冰海、杨超杰、常思、陈程、李成鹏、郑楚杰、洪范、董官廷、赵国斌、胡航瑞、赵涵宇、林浩、翔浩、黄浩彦、钟虎门、王家林、唐家龙、江建东、万建强、马建新、曾建元、张杰、徐金、王金凯、何金正、唐军、可毅、陈克勤、陈朗士、蒋乐、张雷、陈亮、袁曼、杨明坤、孙敏敏、倪娜、陈诺、王鹏、朱鹏、张鹏程、王鹏飞、唐巧宇、付庆、张蓉、彭茹、高瑞泽、权尚浩然、沉黄、白帅、罗双、宋思波、陈松、何涛、何婷、丁伟、魏廖伟佳、葛文斌、尹文标、于文渊、贾贤彦、史献忠、邓晓东、黄晓明、周锡明、王新宇、魏西品、刘学静、刘洋、姚洋、张阳、苗一波、张一丹、朱一凯、张颖儿、蒋勇、李勇、岳永安、朱远志、褚云飞、王泽坤、李兆海、付哲人、李志、志波杨、郭志芳、张志鹏、徐志英、乔子乐、孟子业

参考

Marah I Abdin、Sam Ade Jacobs、Ammar Ahmad Awan、Jyoti Aneja、Ahmed Awadallah、Hany Awadalla、Nguyen Bach、Amit Bahree、Arash Bakhtiari、Harkirat S. Behl、Alon Benhaim、Misha Bilenko、Johan Bjorck、Sébastien Bubeck、Martin Cai、Caio César Teodoro Mendes、 Weizhu Chen、Vishrav Chaudhary、Parul Chopra、Allie Del Giorno、Gustavo de Rosa、Matthew Dixon、Ronen Eldan、Dan Iter、Amit Garg、Abhishek Goswami、Suriya Gunasekar、Emman Haider、Junhenghao、Russell J. Hewett、Jamie Huynh、Mojan Javaheripi、Xin Jin、Piero Kauffmann、Nikos Karampatziakis、Dongwoo Kim、 Mahoud Khademi、Lev Kurilenko、James R. Lee、Yin Tat Lee、李远志、陈亮、刘伟雄、Eric Lin、Zeqi Lin、Piyush Madan、Arindam Mitra、Hardik Modi、Anh Nguyen、Brandon Norick、Barun Patra、Daniel Perez-Becker、Thomas Portet、Reid Pryzant、Heyangqin、Marko Radmilac、Corby Rosset、Sambudha Roy、 Olatunji Ruwase、Olli Saarikivi、Amin Saied、Adil Salim、Michael Santacroce、Shital Shah、Ning Shang、Hiteshi Sharma、Xia Song、Masahiro Tanaka、Xin Wang、Rachel Ward、Guanhua Wang、Philipp Witte、Michael Wyatt、Can Xu、Jiahang Xu、Sonali Yadav、Fan Yang、Ziyi Yang、Donghan Yu、Chengruidong 张、Cyril 张、Jianwen 张、Li张琳娜、张毅、张悦、张雨楠和周习仁。 Phi-3 技术报告：手机本地的高性能语言模型。 *CoRR*，abs/2404.14219，2024。

博·阿德勒、尼凯特·阿加瓦尔、阿什瓦斯·艾瑟尔、Dong H.Anh、帕拉布·巴塔查亚、安妮卡·布伦丁、贾里德·卡斯帕、布莱恩·卡坦扎罗、莎朗·克莱、乔纳森·科恩、沙克·达斯、阿尤什·达塔古普塔、奥利维尔·德拉洛、莱昂·德尔钦斯基、易东、丹尼尔·埃格特、埃莉·埃文斯、亚历山大·菲塞克、丹尼斯Fridman、Shaona Ghosh、Boris Ginsburg、Igor Gitman、Tomasz Grzegorzek、Robert Hero、Jining Huang、Vibhu Jawa、Joseph Jennings、Aastha Jhunjhunwala、John Kamalu、Sadaf Khan、Oleksii Kuchaiev、Patrick LeGresley、李慧、刘继伟、刘子涵、Eileen Peters Long、Ameya Mahabaleshwarkar、Somshubra Majumdar、James Maki、Miguel Martinez、Maer Rodrigues de Melo、Ivan Moshkov、Deepak Narayanan、Sean Narenthiran、Jesus Navarro、Phong Nguyen、Osvald Nitski、Vahid Noroozi、Guruprasad Nutheti、Christopher Parisien、Jupinder Parmar、Mostofa Patwary、Krzysztof Pawelec、Wei Ping、Shrimai Prabhumoye、 Rajarshi Roy、Trisha Saar、Vasanth Rao Naik Sabavat、Sanjeev Satheesh、Jane Polak Scowcroft、Jason D. Sewall、Pavel Shamis、Gerald Shen、Mohammad Shoeybi、Dave Sizer、Misha Smelyanskiy、Felipe Soares、Makesh Narsimhan Sreedhar、Dan Su、Sandeep Subramanian、Shengyang Sun、Shubham Toshniwal、王浩、Zhilin王、嘉轩尤、曾嘉琪、张吉米、张静、张薇薇、张宜安和朱陈。 Nemotron-4 340B 技术报告。 *CoRR*，abs/2406.11704，2024。

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

Ebtesam Almazrouei、Hamza Alobeidli、Abdulaziz Alshamsi、Alessandro Cappelli、Ruxandra Cojocaru、Mérouane Debah、Étienne Goffinet、Daniel Hesslow、Julien Launay、Quentin Malartic、Daniele Mazzotta、Badreddine Noune、Baptiste Pannier 和 Guilherme佩内多。 Falcon系列开放语言模型。 *CoRR*，abs/2311.16867，2023。

安陈鑫、黄飞、张军、龚山三、邱西鹏、周常、孔令鹏。大型语言模型的免训练长上下文扩展。 *CoRR*，abs/2402.17463，2024。

人类。介绍Claude，2023a。 URL https://www.anththropic.com/index/introducing-claude。

人类。 Claude 2。技术报告，人类，2023b。 URL https://www-files.anththropic.com/pro duction/images/model-card-claude-2.pdf。

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic, AI, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

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.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. LongAlign: A recipe for long context alignment of large language models. In *EMNLP (Findings)*, pp. 1376–1395. Association for Computational Linguistics, 2024.

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.

Boxi Cao, Keming Lu, Xinyu Lu, Jiawei Chen, Mengjie Ren, Hao Xiang, Peilin Liu, Yaojie Lu, Ben He, Xianpei Han, Le Sun, Hongyu Lin, and Bowen Yu. Towards scalable automated alignment of LLMs: A survey. *CoRR*, abs/2406.01252, 2024.

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.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A theorem-driven question answering dataset. In *EMNLP*, pp. 7889–7901. Association for Computational Linguistics, 2023a.

Zhihong Chen, Shuo Yan, Juhao Liang, Feng Jiang, Xiangbo Wu, Fei Yu, Guiming Hardy Chen, Junying Chen, Hongbo Zhang, Li Jianquan, Wan Xiang, and Benyou Wang. MultilingualSIFT: Multilingual supervised instruction fine-tuning, 2023b. URL https://github.com/FreedomIntelligence/MultilingualSIFT.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018.

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.

人为的。 Claude 3 型号系列：Opus、Sonnet、Haiku。技术报告，Anthropic，AI，2024 年。URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model Card Claud e 3.pdf。

Jacob Austin、Augustus Odena、Maxwell I. Nye、Maarten Bosma、Henryk Michalewski、David Dohan、Ellen Jiang、Carrie J. Cai、Michael Terry、Quoc V. Le 和 Charles Sutton。使用大型语言模型进行程序综合。*CoRR*，abs/2108.07732，2021。

j白英泽、白帅、褚云飞、崔泽宇、党凯、邓晓东、范杨、葛文斌、韩宇、黄飞、惠斌源、罗吉、李梅、林俊阳、林润吉、刘大一恒、高刘、路成强、陆克明、马建新、门瑞、任兴章、任宣成、谭传奇、谭思南、涂建宏、王鹏、王士杰、王伟、吴胜光、徐本峰、徐进、安阳、杨浩、杨建、杨树生、姚洋、于博文、袁宏毅、袁征、张建伟、张兴轩、张宜昌、张振如、周常、周静仁、周小欢和朱天行。 Qwen技术报告。 *CoRR*，abs/2309.16609，2023。

白雨诗、吕鑫、张家杰、何雨泽、季琪、侯磊、唐杰、董雨晓和李娟子。 LongAlign：大型语言模型的长上下文对齐的秘诀。在*EMNLP (Findings)*中，第1376-1395页。计算语言学协会，2024。

卢卡斯·班达卡、戴维斯·梁、本杰明·穆勒、米克尔·阿泰克斯、萨蒂亚·纳拉扬·舒克拉、唐纳德·胡萨、纳曼·戈亚尔、阿比南丹·克里希南、卢克·泽特莫耶和马迪安·卡布萨。 Belebele 基准：122 种语言变体的并行阅读理解数据集。 *CoRR*，abs/2308.16884，2023。

汤姆·B·布朗、本杰明·曼、尼克·莱德、梅兰妮·苏比亚、贾里德·卡普兰、普拉富拉·达里瓦尔、阿尔文德·尼拉坎坦、普拉纳夫·希亚姆、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·赫尼根、Rewon Child、阿迪亚·拉梅什、丹尼尔·M·齐格勒、杰弗里·吴、克莱门斯 Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。语言模型是小样本学习者。在 *NeurIPS*，2020 年。

曹博熙、卢克明、卢新宇、陈嘉伟、任梦洁、浩翔、刘培林、卢耀杰、何本、韩先培、孙乐、林宏宇和余博文。迈向法学硕士的可扩展自动调整：一项调查。 *CoRR*，abs/2406.01252，2024。

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 和 Abhinav Jangda。 MultiPL-E：一种可扩展的多语言方法，用于对神经代码生成进行基准测试。*IEEE Trans. Software Eng.*，49(7)：3675–3691，2023。

m ark Chen、Jerry Tworek、Heewoo Jun、袁启明、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、米哈伊尔·巴甫洛夫、Alethea Power、卢卡斯·凯撒、穆罕默德·巴伐利亚、克莱门斯·温特、菲利普·蒂莱、费利佩·佩特罗斯基·萨奇、戴夫·卡明斯、马蒂亚斯·普拉珀特、福蒂奥斯·尚齐斯、伊丽莎白·巴恩斯、阿里尔·赫伯特-沃斯、威廉·赫伯根·格斯、亚历克斯·尼科尔、亚历克斯·佩诺、尼古拉斯·特扎克、唐杰、伊戈尔·巴布什金、苏其尔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。

陈文虎、尹铭、Max Ku、潘路、万一鑫、马雪光、徐建宇、王欣怡和夏托尼。 TheoremQA：定理驱动的问答数据集。在*EMNLP*中，第7889–7901页。计算语言学协会，2023a。

陈志宏、严硕、梁巨豪、姜峰、吴翔波、于飞、陈桂明、陈俊英、张宏波、李建泉、万翔和王本友。 MultilingualSIFT：多语言监督指令微调，2023b。网址 https://github.com/FreedomIntelligence/MultilingualSIFT。

Peter Clark、Isaac Cowhey、Oren Etzioni、Tushar Khot、Ashish Sabharwal、Carissa Schoenick 和 Oyvind Tafjord。您认为您已经解决了问答问题吗？尝试 ARC，AI2 推理挑战赛。 *CoRR*，abs/1803.05457，2018。

Karl Cobbe、Vineet Kosaraju、Mohammad Bavarian、Mark Chen、Heewoo Jun、Lukasz Kaiser、Matthias Plappert、Jerry Tworek、Jacob Hilton、Reiichiro Nakano、Christopher Hesse 和 John Schulman。培训验证员解决数学应用题。 *CoRR*，abs/2110.14168，2021。

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

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

Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. Self-play with execution feedback: Improving instruction-following capabilities of large language models. *CoRR*, abs/2406.13542, 2024.

Shihan Dou, Jiazheng Zhang, Jianxiang Zang, Yunbo Tao, Haoxiang Jia, Shichun Liu, Yuming Yang, Shenxi Wu, Shaoqing Zhang, Muling Wu, et al. Multi-programming language sandbox for llms. *CoRR*, abs/2410.23074, 2024.

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.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39, 2022.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton A. Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. MERA: A comprehensive LLM evaluation in russian. *CoRR*, abs/2401.04531, 2024.

Evan Frick, Peter Jin, Tianle Li, Karthik Ganesan, Jian Zhang, Jiantao Jiao, and Banghua Zhu. Athene-70b: Redefining the boundaries of post-training for open models, July 2024a. URL https://nexusflow.ai/blogs/athene.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF. *CoRR*, abs/2410.14872, 2024b.

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.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google, 2024. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118, 2024.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans. Assoc. Comput. Linguistics*, 10: 522–538, 2022.

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

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

Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. Self-play with execution feedback: Improving instruction-following capabilities of large language models. *CoRR*, abs/2406.13542, 2024.

Shihan Dou, Jiazheng Zhang, Jianxiang Zang, Yunbo Tao, Haoxiang Jia, Shichun Liu, Yuming Yang, Shenxi Wu, Shaoqing Zhang, Muling Wu, et al. Multi-programming language sandbox for llms. *CoRR*, abs/2410.23074, 2024.

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.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39, 2022.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton A. Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. MERA: A comprehensive LLM evaluation in russian. *CoRR*, abs/2401.04531, 2024.

Evan Frick, Peter Jin, Tianle Li, Karthik Ganesan, Jian Zhang, Jiantao Jiao, and Banghua Zhu. Athene-70b: Redefining the boundaries of post-training for open models, July 2024a. URL https://nexusflow.ai/blogs/athene.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF. *CoRR*, abs/2410.14872, 2024b.

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.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google, 2024. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118, 2024.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans. Assoc. Comput. Linguistics*, 10: 522–538, 2022.

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.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022.

Keith Hoskin. The "awful idea of accountability": Inscribing people into the measurement of objects. *Accountability: Power, ethos and the technologies of managing*, 1996.

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.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kai Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *CoRR*, abs/2404.06395, 2024.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *CoRR*, abs/2310.06825, 2023a.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *CoRR*, abs/2401.04088, 2024a.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *arXiv preprint arXiv:2407.02490*, 2024b.

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, 2023b.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *EMNLP*, pp. 12359–12374. Association for Computational Linguistics, 2023.

Nathan Lambert, Valentina Pyatkin, Jacob Daniel Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Raghavi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hanna Hajishirzi. RewardBench: Evaluating reward models for language modeling. *CoRR*, abs/2403.13787, 2024.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. *CoRR*, abs/2006.16668, 2020.

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.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022.

Keith Hoskin. The "awful idea of accountability": Inscribing people into the measurement of objects. *Accountability: Power, ethos and the technologies of managing*, 1996.

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.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kai Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *CoRR*, abs/2404.06395, 2024.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *CoRR*, abs/2310.06825, 2023a.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *CoRR*, abs/2401.04088, 2024a.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *arXiv preprint arXiv:2407.02490*, 2024b.

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, 2023b.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *EMNLP*, pp. 12359–12374. Association for Computational Linguistics, 2023.

Nathan Lambert, Valentina Pyatkin, Jacob Daniel Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Raghavi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hanna Hajishirzi. RewardBench: Evaluating reward models for language modeling. *CoRR*, abs/2403.13787, 2024.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. *CoRR*, abs/2006.16668, 2020.

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.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *ACL (1)*, pp. 3214–3252. Association for Computational Linguistics, 2022a.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In *EMNLP*, pp. 9019–9052. Association for Computational Linguistics, 2022b.

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*, 2023.

Keming Lu, Bowen Yu, Fei Huang, Yang Fan, Runji Lin, and Chang Zhou. Online merging optimizers for boosting rewards and mitigating tax in alignment. *CoRR*, abs/2405.17931, 2024a.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *CoRR*, abs/2401.12474, 2024b.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In *ACL (1)*, pp. 15991–16111. Association for Computational Linguistics, 2023.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Pérez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Ki-Woong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, José Camacho-Collados, and Alice Oh. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *CoRR*, abs/2406.09948, 2024.

OpenAI. GPT4 technical report. *CoRR*, abs/2303.08774, 2023.

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

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

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

Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In *EMNLP (1)*, pp. 2362–2376. Association for Computational Linguistics, 2020.

Shanghaoran Quan, Tianyi Tang, Bowen Yu, An Yang, Dayiheng Liu, Bofei Gao, Jianhong Tu, Yichang Zhang, Jingren Zhou, and Junyang Lin. Language models can self-lengthen to generate long texts. *CoRR*, abs/2410.23933, 2024.

Qwen Team. Code with CodeQwen1.5, 2024a. URL https://qwenlm.github.io/blog/codeqwen1.5/.

Qwen Team. Introducing Qwen1.5, 2024b. URL https://qwenlm.github.io/blog/qwen1.5/.

Qwen Team. Introducing Qwen2-Math, 2024c. URL https://qwenlm.github.io/blog/qwen2-math/.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

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.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *ACL (1)*, pp. 3214–3252. Association for Computational Linguistics, 2022a.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In *EMNLP*, pp. 9019–9052. Association for Computational Linguistics, 2022b.

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*, 2023.

Keming Lu, Bowen Yu, Fei Huang, Yang Fan, Runji Lin, and Chang Zhou. Online merging optimizers for boosting rewards and mitigating tax in alignment. *CoRR*, abs/2405.17931, 2024a.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *CoRR*, abs/2401.12474, 2024b.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In *ACL (1)*, pp. 15991–16111. Association for Computational Linguistics, 2023.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Pérez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Ki-Woong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, José Camacho-Collados, and Alice Oh. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *CoRR*, abs/2406.09948, 2024.

OpenAI. GPT4 technical report. *CoRR*, abs/2303.08774, 2023.

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

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

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

Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In *EMNLP (1)*, pp. 2362–2376. Association for Computational Linguistics, 2020.

Shanghaoran Quan, Tianyi Tang, Bowen Yu, An Yang, Dayiheng Liu, Bofei Gao, Jianhong Tu, Yichang Zhang, Jingren Zhou, and Junyang Lin. Language models can self-lengthen to generate long texts. *CoRR*, abs/2410.23933, 2024.

Qwen Team. Code with CodeQwen1.5, 2024a. URL https://qwenlm.github.io/blog/codeqwen1.5/.

Qwen Team. Introducing Qwen1.5, 2024b. URL https://qwenlm.github.io/blog/qwen1.5/.

Qwen Team. Introducing Qwen2-Math, 2024c. URL https://qwenlm.github.io/blog/qwen2-math/.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. DeepSpeed-MoE: Advancing mixture-of-experts inference and training to power next-generation AI scale. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18332–18346. PMLR, 2022.

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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, 2021.

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.

Jianlin Su. The magical effect of the Bias term: RoPE + Bias = better length extrapolation, 2023. URL https://spaces.ac.cn/archives/9577.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. Secrets of RLHF in large language models part II: Reward modeling. *CoRR*, abs/2401.06080, 2024a.

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

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, 2024b.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. HelpSteer2-Preference: Complementing ratings with preferences. *CoRR*, abs/2410.01257, 2024c.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. DeepSpeed-MoE: Advancing mixture-of-experts inference and training to power next-generation AI scale. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18332–18346. PMLR, 2022.

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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, 2021.

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.

Jianlin Su. The magical effect of the Bias term: RoPE + Bias = better length extrapolation, 2023. URL https://spaces.ac.cn/archives/9577.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. Secrets of RLHF in large language models part II: Reward modeling. *CoRR*, abs/2401.06080, 2024a.

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

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, 2024b.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. HelpSteer2-Preference: Complementing ratings with preferences. *CoRR*, abs/2410.01257, 2024c.

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.

Hao Xiang, Bowen Yu, Hongyu Lin, Keming Lu, Yaojie Lu, Xianpei Han, Le Sun, Jingren Zhou, and Junyang Lin. Aligning large language models via self-steering optimization. *CoRR*, abs/2410.17131, 2024.

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.

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, 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, 2024b.

Jian Yang, Jiaxi Yang, Ke Jin, Yibo Miao, Lei Zhang, Liqun Yang, Zeyu Cui, Yichang Zhang, Binyuan Hui, and Junyang Lin. Evaluating and aligning codellms on human preference. *CoRR*, abs/2412.05210, 2024c.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *EMNLP/IJCNLP (1)*, pp. 3685–3690. Association for Computational Linguistics, 2019.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.AI. *CoRR*, abs/2403.04652, 2024.

Tao Yuan, Xuefei Ning, Dong Zhou, Zhijie Yang, Shiyao Li, Minghui Zhuang, Zheyue Tan, Zhuyu Yao, Dahua Lin, Boxun Li, Guohao Dai, Shengen Yan, and Yu Wang. LV-Eval: A balanced long-context benchmark with 5 length levels up to 256K. *CoRR*, abs/2402.05136, 2024.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. Scaling relationship on learning mathematical reasoning with large language models. *CoRR*, abs/2308.01825, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *ACL (1)*, pp. 4791–4800. Association for Computational Linguistics, 2019.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS*, 2023.

Enyu Zhou, Guodong Zheng, Bing Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. RMB: Comprehensively benchmarking reward models in LLM alignment. *CoRR*, abs/2410.09893, 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.

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.

Hao Xiang, Bowen Yu, Hongyu Lin, Keming Lu, Yaojie Lu, Xianpei Han, Le Sun, Jingren Zhou, and Junyang Lin. Aligning large language models via self-steering optimization. *CoRR*, abs/2410.17131, 2024.

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.

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, 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, 2024b.

Jian Yang, Jiaxi Yang, Ke Jin, Yibo Miao, Lei Zhang, Liqun Yang, Zeyu Cui, Yichang Zhang, Binyuan Hui, and Junyang Lin. Evaluating and aligning codellms on human preference. *CoRR*, abs/2412.05210, 2024c.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *EMNLP/IJCNLP (1)*, pp. 3685–3690. Association for Computational Linguistics, 2019.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.AI. *CoRR*, abs/2403.04652, 2024.

Tao Yuan, Xuefei Ning, Dong Zhou, Zhijie Yang, Shiyao Li, Minghui Zhuang, Zheyue Tan, Zhuyu Yao, Dahua Lin, Boxun Li, Guohao Dai, Shengen Yan, and Yu Wang. LV-Eval: A balanced long-context benchmark with 5 length levels up to 256K. *CoRR*, abs/2402.05136, 2024.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. Scaling relationship on learning mathematical reasoning with large language models. *CoRR*, abs/2308.01825, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *ACL (1)*, pp. 4791–4800. Association for Computational Linguistics, 2019.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS*, 2023.

Enyu Zhou, Guodong Zheng, Bing Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. RMB: Comprehensively benchmarking reward models in LLM alignment. *CoRR*, abs/2410.09893, 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.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models. *CoRR*, abs/2202.08906, 2022.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models. *CoRR*, abs/2202.08906, 2022.