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.

在本报告中，我们介绍了Qwen2.5，这是一个全面的大型语言模型（LLMs）系列，旨在满足多样化的需求。与之前的版本相比，Qwen2.5在预训练和后训练阶段都得到了显著改进。在预训练方面，我们将高质量预训练数据集从之前的7万亿个标记扩展到18万亿个标记，这为常识、专家知识和推理能力提供了坚实的基础。在后训练方面，我们实施了超过100万个样本的精细监督微调，以及多阶段强化学习，包括离线学习DPO和在线学习GRPO。后训练技术显著增强了人类偏好，并显著改善了长文本生成、结构化数据分析和指令遵循能力

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.

为了有效处理多样化的用例，我们提供了丰富的Qwen2.5 LLM系列配置。开放权重的产品包括0.5B、1.5B、3B、7B、14B、32B和72B参数的基础模型和指令调优模型。还提供了指令调优模型的量化版本。超过100个模型可以从Hugging Face Hub、ModelScope和Kaggle访问。此外，对于托管解决方案，专有模型目前包括两个专家混合（MoE）变体：Qwen2.5Turbo和Qwen2.5-Plus，两者均可从阿里云Model Studio获取

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.

Qwen2.5在评估语言理解、推理、数学、编码、人类偏好对齐等方面的广泛基准测试中展示了顶级性能。具体而言，开放权重的旗舰产品Qwen2.5-72B-Instruct在多个开放和专有模型中表现出色，并与当前最先进的开放权重模型Llama-3-405B-Instruct（大约是其5倍大小）竞争。Qwen2.5-Turbo和Qwen2.5-Plus在成本效益方面表现出色，分别与GPT-4o-mini和GPT-4o竞争。此外，作为基础，Qwen2.5模型在训练专门模型（如Qwen2.5-Math（Yang等，2024b）、Qwen2.5-Coder（Hui等，2024）、QwQ（Qwen团队，2024d）和多模态模型）中发挥了重要作用

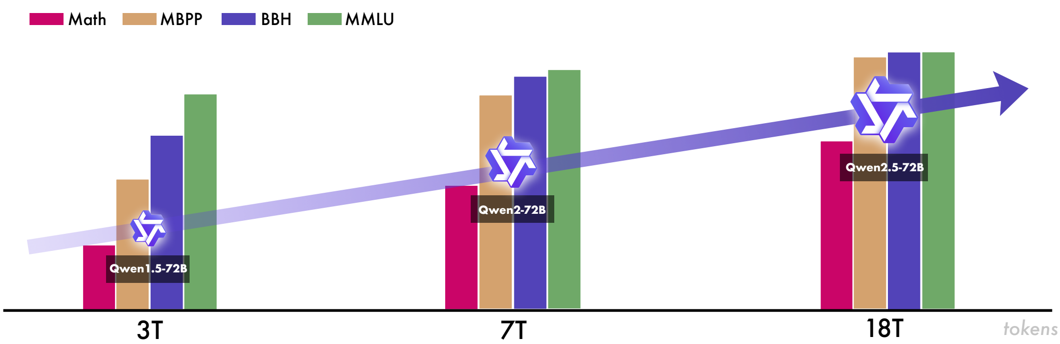

Figure 1: In the iterative development of the Qwen series, data scaling has played a crucial role. Qwen 2.5, which leverages 18 trillion tokens for pre-training, has demonstrated the most advanced capabilities within the Qwen series, especially in terms of domain expertise, underscoring the importance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

图1：在Qwen系列的迭代开发过程中，数据扩展发挥了至关重要的作用。Qwen 2.5利用18万亿个标记进行预训练，展现了Qwen系列中最先进的能力，特别是在领域专业知识方面，凸显了规模与混合在提升模型能力中的重要性

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.

人工通用智能（AGI）的火花通过大型基础模型，尤其是大型语言模型（LLMs）的快速发展日益显现（Brown等，2020；OpenAI，2023；2024a；Gemini团队，2024；Anthropic，2023a；b；2024；Bai等，2023；Yang等，2024a；Touvron等，2023a；b；Dubey等，2024）。模型和数据规模的持续进步，结合大规模预训练后高质量监督微调（SFT）和基于人类反馈的强化学习（RLHF）的范式（Ouyang等，2022），使得大型语言模型（LLMs）在语言理解、生成和推理方面展现出涌现能力。 在此基础上，推理时间规模化的最新突破，特别是o1（OpenAI，2024b）所展示的，通过逐步推理和反思增强了LLMs的深度思考能力。这些进展提升了语言模型的潜力，表明随着它们继续展现出更具通用性的人工智能的涌现能力，它们可能在科学探索中取得重大突破

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; 2024), 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.

除了模型能力的快速发展外，近两年在大型语言模型（LLM）领域涌现出一批开放权重（open-weight）的大型语言模型，例如Llama系列（Touvron等，2023a；b；Dubey等，2024）、Mistral系列（Jiang等，2023a；2024）以及我们的Qwen系列（Bai等，2023；Yang等，2024a；Qwen团队，2024a；Hui等，2024；Qwen团队，2024c；Yang等，2024b）。这些开放权重模型使得普通用户和开发者能够更便捷地访问大型语言模型，促进了更广泛的研究参与，通过社区协作推动了创新，并加速了人工智能应用在多个领域的开发进程

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.

近日，我们发布了Qwen系列最新版本Qwen2.5的详细信息。在开源权重部分，我们发布了包括0.5B、1.5B、3B、7B、14B、32B和72B在内的7种规模的预训练和指令调优模型，不仅提供了bfloat16精度的原始模型，还提供了不同精度的量化模型。具体而言，旗舰模型Qwen2.5-72B-Instruct在与当前最先进的开源权重模型Llama-3-405B-Instruct的对比中展现出竞争力，而后者规模约为前者的5倍。此外，我们还发布了专家混合模型（Mixture-of-Experts, MoE, Lepikhin等，2020; Fedus等，2022; Zoph等，2022）的专有模型，即Qwen2.5-Turbo和Qwen2.5-Plus1，它们分别在与GPT-4o-mini和GPT-4o的对比中表现出竞争力

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:

在本技术报告中，我们介绍了Qwen2.5，这是我们持续努力创造更优大型语言模型（LLMs）的成果。下文展示了最新版Qwen的关键特性

- **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.

在规模上更优：相较于Qwen2，除了0.5B、1.5B、7B和72B模型外，Qwen2.5重新引入了3B、14B和32B模型，这些模型在资源受限的场景下更具成本效益，并且在当前开放基础模型领域中代表性不足。Qwen2.5Turbo和Qwen2.5-Plus在准确性、延迟和成本之间实现了良好的平衡

- **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).

在数据上更优：预训练和后训练数据得到了显著改进。预训练数据从7万亿个标记增加到18万亿个标记，重点关注知识、编码和数学领域。预训练分阶段进行，以便在不同混合之间进行过渡。后训练数据达到100万个示例，涵盖监督微调（SFT，Ouyang等，2022）、直接偏好优化（DPO，Rafailov等，2023）和群体相对策略优化（GRPO，Shao等，2024）阶段

- **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.

在使用上更优：Qwen2在使用中的几个关键限制已被消除，包括更大的生成长度（从2K标记增加到8K标记）、更好地支持结构化输入和输出（例如表格和JSON），以及更便捷的工具使用。此外，Qwen2.5-Turbo支持高达100万个标记的上下文长度

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.

基本上，Qwen2.5系列包括开源密集模型，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。 以下，我们将提供有关模型架构的详细信息

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.

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

对于密集模型，我们保持了基于Transformer的解码器架构（Vaswani等，2017；Radford等，2018），如Qwen2（Yang等，2024a）所示。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力（GQA，Ainslie等，2023）、用于非线性激活的SwiGLU激活函数（Dauphin等，2017）、用于编码位置信息的旋转位置嵌入（RoPE，Su等，2024）、注意力机制中的QKV偏置（Su，2023）以及预归一化的RMSNorm（Jiang等，2023b）以确保训练稳定性

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.

基于密集模型架构，我们将其扩展至混合专家（MoE）模型架构。这一扩展通过将标准的前馈网络（FFN）层替换为专门的MoE层实现，其中每一层包含多个FFN专家及一个路由机制，该机制负责将令牌分配给前K个专家。 遵循Qwen1.5-MoE（Yang等人，2024a）所展示的方法，我们实施了细粒度的专家分割（Dai等人，2024）以及共享专家路由（Rajbhandari等人，2022；Dai等人，2024）。 这些架构上的创新在下游任务中显著提升了模型性能

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.

在分词方面，我们采用了Qwen的分词器（Bai等人，2023），该分词器实现了字节级字节对编码（BBPE，Brown等人，2020；Wang等人，2020；Sennrich等人，2016），并拥有151,643个常规词汇单元。相较于之前的Qwen版本，我们将控制词汇单元从3个扩展至22个，新增了两个用于工具功能的词汇单元，并将剩余的词汇单元分配给其他模型功能。这一扩展为所有Qwen2.5模型建立了统一的词汇表，增强了模型间的一致性，并减少了潜在的兼容性问题

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.

我们的语言模型预训练过程包含若干关键组成部分。首先，我们通过精密的筛选与评分机制，结合策略性的数据混合，精心策划高质量的训练数据。其次，我们进行了广泛的超参数优化研究，以有效地训练不同规模的模型。最后，我们引入了专门的长上下文预训练，以增强模型处理和理解长序列的能力。下文将详细阐述我们在数据准备、超参数选择及长上下文训练方面的具体方法

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:

Qwen2.5相较于其前身Qwen2，在预训练数据质量方面展现出显著提升。这些改进源自以下几个关键方面：

(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.

更优的数据过滤机制。高质量的预训练数据对模型性能至关重要，因此数据质量评估与过滤成为我们流程中的关键环节。我们采用Qwen2-Instruct模型作为数据质量过滤器，执行全面、多维度的分析，以评估并评分训练样本。相较于先前用于Qwen2的方法，此过滤技术实现了显著进步，得益于Qwen2在更大规模多语言语料库上的扩展预训练。增强的能力使得质量评估更为细致，不仅提高了高质量训练数据的保留率，也加强了对多种语言中低质量样本的有效过滤

(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.

更优的数学与代码数据。在Qwen2.5的预训练阶段，我们整合了来自Qwen2.5-Math（Yang等人，2024b）和Qwen2.5-Coder（Hui等人，2024）的训练数据。这一数据整合策略被证明极为有效，因为这些专门的数据集对于在数学和编码任务上实现顶尖性能起到了关键作用。通过在预训练期间利用这些高质量的领域特定数据集，Qwen2.5继承了在数学推理和代码生成方面的强大能力

(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.

更优质的合成数据。为生成高质量的合成数据，特别是在数学、代码及知识领域，我们采用了Qwen2-72B-Instruct（杨等人，2024a）与Qwen2Math-72B-Instruct（Qwen团队，2024c）两种模型。通过我们专有的通用奖励模型及专门的Qwen2-Math-RM-72B（Qwen团队，2024c）模型进行严格筛选，进一步提升了这些合成数据的质量

(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.

优化数据混合比例。为了优化预训练数据的分布，我们采用Qwen2-Instruct模型对不同领域的内容进行分类与平衡。分析表明，电子商务、社交媒体及娱乐等领域在网络规模数据中占比显著过高，这些领域常包含重复、模板化或机器生成的内容。相反，技术、科学及学术研究等领域虽蕴含更高质量的信息，却历来在数据集中代表性不足。通过策略性地对过度代表领域进行下采样，并对高价值领域进行上采样，我们确保了训练数据集更加均衡且信息丰富，从而更好地服务于模型的学习目标

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.

基于这些技术，我们开发了一个规模更大、质量更高的预训练数据集，从Qwen2（Yang等，2024a）中使用的7万亿个标记扩展到了18万亿个标记

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.

我们基于Qwen2.5的预训练数据（Hoffmann等人，2022；Kaplan等人，2020）开发了超参数的缩放定律。尽管先前的研究（Dubey等人，2024；Almazrouei等人，2023；Hoffmann等人，2022）主要利用缩放定律在给定计算预算的情况下确定最佳模型规模，但我们则利用这些定律来识别跨模型架构的最佳超参数。具体而言，我们的缩放定律有助于确定关键训练参数，如批量大小B和学习率μ，适用于不同规模的密集模型和混合专家（MoE）模型

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.

通过广泛的实验，我们系统性地研究了模型架构与最优训练超参数之间的关系。具体而言，我们分析了最优学习率μopt和批量大小Bopt如何随模型规模N和预训练数据规模D的变化而变化。我们的实验涵盖了广泛的架构范围，包括参数数量从44M到14B的密集模型以及激活参数数量从44M到1B的混合专家（MoE）模型，这些模型在0.8B到600B个token的数据集上进行了训练。 利用这些最优超参数预测，我们进一步将最终损失建模为模型架构和训练数据规模的函数

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.

此外，我们利用缩放定律来预测并比较不同参数规模的混合专家模型（MoE）与其密集模型对应物的性能。这一分析指导了我们为MoE模型配置超参数，通过精细调整激活参数与总参数，使我们能够实现与特定密集模型变体（如Qwen2.5-72B和Qwen2.5-14B）的性能相当

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).

为了达到最佳的训练效率，Qwen2.5采用了双阶段预训练策略：首先进行4096个标记长度的上下文训练，随后进入针对更长序列的扩展阶段。借鉴Qwen2的策略，在最终预训练阶段，除Qwen2.5-Turbo外的所有模型变体，其上下文长度均从4096扩展至32768个标记。同时，我们运用ABF技术（Xiong等人，2023年），将RoPE的基础频率从10,000提升至1,000,000

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.

对于Qwen2.5-Turbo，我们在训练过程中实施了一种渐进式上下文长度扩展策略，依次推进四个阶段：32,768个标记、65,536个标记、131,072个标记，最终达到262,144个标记，RoPE基础频率为10,000,000。在每个阶段，我们精心策划训练数据，确保包含40%当前最大长度的序列和60%较短的序列。这种渐进式训练方法使得模型能够平稳适应不断增加的上下文长度，同时保持其有效处理和泛化不同长度序列的能力

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.

为增强模型在推理过程中处理更长序列的能力，我们实施了两项关键策略：YARN（Peng等人，2023年）和双块注意力机制（Dual Chunk Attention，DCA，An等人，2024年）。通过这些创新，我们实现了序列长度处理能力的四倍提升，使得Qwen2.5-Turbo能够处理高达100万个标记，而其他模型则能处理多达131,072个标记。值得注意的是，这些方法不仅通过降低困惑度改善了长序列的建模效果，还保持了模型在短序列上的强劲性能，确保了不同输入长度下的一致质量

4 Post-training

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

相较于Qwen 2，Qwen 2.5在其后训练设计中引入了两项重大改进

(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

扩展监督微调数据覆盖范围：监督微调过程利用了一个包含数百万高质量示例的大规模数据集。此次扩展特别针对先前模型表现出局限性的关键领域，如长序列生成、数学问题解决、编码、指令遵循、结构化数据理解、逻辑推理、跨语言迁移以及鲁棒系统指令

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

两阶段强化学习：Qwen 2.5中的强化学习（RL）过程被划分为两个独立阶段：离线强化学习与在线强化学习

(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.

离线强化学习：此阶段着重于培养奖励模型难以评估的能力，如推理、事实准确性及指令遵循。通过精心构建与验证训练数据，我们确保离线强化学习信号既易于学习又可靠（Xiang等，2024），使模型能有效掌握这些复杂技能

- *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.

在线强化学习：在线强化学习阶段利用奖励模型识别输出质量细微差别的能力，包括真实性、帮助性、简洁性、相关性、无害性及去偏见性。它使模型能够生成精确、连贯且结构良好的响应，同时保持安全性与可读性。因此，模型的输出始终符合人类质量标准与期望

- *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

在本节中，我们详细阐述了Qwen2.5在SFT阶段所做出的关键改进，重点关注以下几个关键领域

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

长序列生成：Qwen2.5能够生成高质量内容，其输出上下文长度可达8,192个标记，相较于通常训练后响应长度（通常不超过2,000个标记）有了显著提升。为弥补这一差距，我们开发了长响应数据集（Quan等，2024）。我们采用回译技术从预训练语料库中生成针对长文本数据的查询，施加输出长度限制，并利用Qwen2筛选出低质量的配对数据

(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.

数学：我们引入了Qwen2.5-Math（Yang等，2024b）的思维链数据，该数据涵盖了多样化的查询来源，包括公共数据集、K-12问题集以及合成问题。为确保高质量的推理过程，我们采用了拒绝采样（Yuan等，2023）结合奖励建模和带注释的答案进行指导，生成了逐步的推理过程

(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.

编码：为提升编码能力，我们整合了Qwen2.5Coder（Hui等人，2024年）的指令调优数据。通过将多种语言特定代理融入协作框架，我们在近40种编程语言中生成多样且高质量的指令对。我们通过从代码相关问答网站合成新示例及从GitHub收集算法代码片段，进一步扩充了指令数据集。采用一个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，确保代码质量与正确性（Dou等人，2024年；Yang等人，2024c）

(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).

指令遵循：为确保高质量的指令遵循数据，我们实施了一套严格的基于代码的验证框架。在此方法中，大型语言模型（LLMs）不仅生成指令，还生成相应的验证代码，并配备全面的单元测试以进行交叉验证。通过基于执行反馈的拒绝采样，我们精心筛选用于监督微调的训练数据，从而确保模型忠实遵循预定指令（Dong等，2024）

(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).

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，该数据集不仅涵盖了传统任务，如表格问答、事实核查、错误修正和结构理解，还包括涉及结构化和半结构化数据的复杂任务。通过将推理链融入模型的响应中，我们显著增强了其从结构化数据中推断信息的能力，从而提升了其在各种任务中的表现。这一方法不仅拓宽了数据集的范围，还加深了模型从复杂数据结构中推理和提取有意义见解的能力

(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.

逻辑推理：为增强模型的逻辑推理能力，我们引入了涵盖多个领域的70,000个新查询。这些查询包括选择题、判断题及开放式问题。模型被训练以系统化方式处理问题，采用多种推理方法，如演绎推理、归纳概括、类比推理、因果推理及统计推理。通过迭代优化，我们系统地筛选出包含错误答案或推理过程有缺陷的数据。这一过程逐步强化了模型进行逻辑准确推理的能力，确保其在各类推理任务中表现稳健

(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.

跨语言迁移：为了促进模型在多种语言间通用能力的迁移，我们采用翻译模型将高资源语言的指令转换为多种低资源语言，从而生成相应的响应候选。为确保这些响应的准确性和一致性，我们评估了每种多语言响应与其原始对应项之间的语义对齐。这一过程保留了原始响应的逻辑结构和风格细微差别，从而在不同语言间保持了其完整性和连贯性

(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.

稳健系统指令构建：我们构建了数百个通用系统提示，以增强训练后系统提示的多样性，确保系统提示与对话之间的一致性。采用不同系统提示的评估结果显示，模型保持了良好的性能（Lu等，2024b），且方差降低，表明其稳健性得到了提升

(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.

响应过滤：为评估回答质量，我们采用多种自动标注方法，包括专用的批评模型与多代理协作评分系统。所有回答均需经过严格评估，唯有在所有评分系统中均被视为无瑕疵者方得以保留。此全面方法确保我们的输出维持最高质量标准

(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.

最终，我们构建了一个包含超过100万条SFT样本的数据集。模型以32,768个标记的序列长度进行了两个周期的微调。为了优化学习过程，学习率从7×10^{-}逐步降低至7×10^{-7}。为了防止过拟合，我们采用了0.1的权重衰减，并将梯度范数限制在最大值1.0以内

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

相较于在线强化学习（RL），离线RL允许预先准备训练信号，这对于存在标准答案但难以通过奖励模型进行评估的任务尤为有利。在本研究中，我们聚焦于数学、编程、指令遵循及逻辑推理等客观查询领域，这些领域中获取准确评估可能较为复杂。 在前一阶段，我们广泛采用了执行反馈和答案匹配等策略以确保回答质量。当前阶段，我们复用该流程，利用SFT模型对新一组查询进行回答重采样。通过质量检查的回答被用作正面示例，而未通过者则作为直接偏好优化（DPO）训练的负面示例（Rafailov等人，2023）。为进一步提升训练信号的可靠性与准确性，我们结合了人工与自动化审查流程（Cao等人，2024）。这种双重方法确保了训练数据不仅可学习，而且与人类期望保持一致。最终，我们构建了一个包含约150,000个训练对的数据集。随后，使用在线合并优化器（Lu等人，2024a）以7×10^{-7}的学习率对模型进行一个周期的训练

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

为了开发一个稳健的在线强化学习奖励模型，我们遵循了一套精心定义的标注标准。这些标准确保模型生成的响应不仅高质量，而且符合伦理和以用户为中心的标准（Wang等，2024a）。数据标注的具体指导原则如下：

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:

真实性：回应必须基于事实准确性，忠实反映所提供的背景和指示。模型应避免生成虚假或未经给定数据支持的信息

- **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.

帮助性：模型的输出应真正有用，有效解决用户的查询，同时提供积极、引人入胜、教育性强且相关的内容。它应严格遵循给定的指示，并为用户提供价值

- **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.

简洁性：回应应简明扼要，避免不必要的冗长。目标是清晰高效地传达信息，而不会用过多的细节压倒用户

- **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.

相关性：回应的所有部分都应直接与用户的查询、对话历史和助手的背景相关。模型应调整其输出，确保其完全符合用户的需求和期望

- **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.

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

- **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.

- **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.

去偏化：模型应生成无偏见的回应，包括但不限于性别、种族、国籍和政治等方面。它应平等且公正地对待所有话题，遵循广泛接受的道德与伦理标准

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.

用于训练奖励模型的查询来源于两个不同的数据集：公开可用的开源数据和一个以更高复杂性为特征的专有查询集。响应是从Qwen模型的检查点生成的，这些模型在训练的不同阶段使用了不同的方法——监督微调（SFT）、直接偏好优化（DPO）和强化学习（RL）——进行了微调。为了引入多样性，这些响应是在不同的温度设置下采样的。偏好对通过人工和自动化标注过程创建，DPO的训练数据也被整合到这个数据集中

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.

在我们的在线强化学习（RL）框架中，我们采用了群体相对策略优化（Group Relative Policy Optimization, GRPO, Shao et al, 2024）。用于训练奖励模型的查询集与RL训练阶段所使用的相同。训练过程中查询的处理顺序由其响应得分的方差决定，该方差由奖励模型评估。具体而言，响应得分方差较高的查询会被优先处理，以确保更有效的学习。我们为每个查询采样8个响应。所有模型均以2048的全局批量大小进行训练，每轮训练包含2048个样本，其中每个样本由一对查询和响应组成

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.

为进一步扩展Qwen2.5-Turbo的上下文长度，我们在后训练阶段引入了更长的监督微调（SFT）样本，使其在长查询任务中能更好地与人类偏好保持一致

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.

在SFT（监督式微调）阶段，我们采用了一种两阶段的方法。第一阶段，模型仅通过包含最多32,768个标记的简短指令进行微调。此阶段使用的数据和训练步骤与其他Qwen2.5模型相同，确保了在简短任务上的强劲表现。 第二阶段，微调过程结合了简短指令（最多32,768个标记）和长指令（最多262,144个标记）。这种混合方法有效提升了模型在长上下文任务中的指令遵循能力，同时保持了其在简短任务上的性能

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.

在强化学习（RL）阶段，我们采用了与其他Qwen2.5模型相似的训练策略，仅专注于短指令。这一设计选择主要基于两点考虑：首先，针对长上下文任务进行RL训练在计算上成本高昂；其次，目前缺乏能够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，仅对短指令实施RL训练，仍能显著提升模型在长上下文任务中与人类偏好的对齐程度

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.

通过预训练生成的基础模型与经过后训练调整的指令调优模型，均采用一套全面的评估体系进行相应评价。该评估体系囊括了广泛使用的公开基准测试及针对特定技能的内部数据集，其设计旨在实现主要自动化评估，最大限度减少人工干预

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 $|\text{LCS}(\mathbf{s}_t, \mathbf{s}_e)| \geq 0.6 \times \min(|\mathbf{s}_t|, |\mathbf{s}_e|)$.

为防止测试数据泄露，在构建预训练和后训练数据集时，我们采用n-gram匹配方法排除潜在污染数据。依据Qwen2所采用的标准，若存在测试序列se，使得标记化后的训练序列st与se之间的最长公共子序列（LCS）长度满足|LCS(st, se)| ⩾ 13且|LCS(st, se)| ⩾ 0.6 × min(|st|, |se|)，则将该训练序列st从训练数据中移除

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.

我们对Qwen2.5系列的基础语言模型进行了全面评估。基础模型的评估主要侧重于其在自然语言理解、通用问答、编程、数学、科学知识、推理以及多语言能力方面的表现

The evaluation datasets include: 评估数据集包括：

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).

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

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**

数学与科学任务 GPQA（Rein 等人，2023）（5-shot）、定理问答（Chen 等人，2023a）（5-shot）、GSM8K（Cobbe 等人，2021）（4-shot）、MATH（Hendrycks 等人，2021b）（4-shot）

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).

编程任务 HumanEval（Chen 等，2021）（零样本），HumanEval+（Liu 等，2023）（零样本），MBPP（Austin 等，2021）（零样本），MBPP+（Liu 等，2023）（零样本），MultiPL-E（Cassano 等，2023）（零样本）（Python、C++、JAVA、PHP、TypeScript、C#、Bash、JavaScript）

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).

多语言任务 我们将其分为四类：(a) 考试类：M3Exam（5-shot，我们仅选择不需要图像的示例）、IndoMMLU（Koto等，2023）（3-shot）、ruMMLU（Fenogenova等，2024）（5-shot）以及翻译版MMLU（Chen等，2023b）（针对阿拉伯语、西班牙语、法语、葡萄牙语、德语、意大利语、日语和韩语的5-shot）；(b) 理解类：BELEBELE（Bandarkar等，2023）（5-shot）、XCOPA（Ponti等，2020）（5-shot）、XWinograd（Muennighoff等，2023）（5-shot）、XStoryCloze（Lin等，2022b）（0-shot）和PAWS-X（Yang等，2019）（5-shot）；(c) 数学类：MGSM（Goyal等，2022）（8-shot CoT）；以及(d) 翻译类：Flores-101（Goyal等，2022）（5-shot）

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).

对于基础模型，我们从参数规模的角度将Qwen2.5模型与Qwen2模型及其他领先的开源权重模型进行了比较

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., 2024), 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.

Qwen2.5-72B与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源权重基础模型进行了比较：Llama3-70B（Dubey等，2024）、Llama3-405B（Dubey等，2024）、Mixtrail-8x22B（Jiang等，2024）以及我们之前的72B版本Qwen2-72B（Yang等，2024a）。Qwen2.5-72B基础模型在广泛的任务中显著优于同类别其他模型。它在仅使用五分之一参数的情况下，取得了与Llama3-405B相当的结果。此外，与上一代Qwen2-72B相比，Qwen2.5-72B在几乎所有基准评估中均表现出显著提升，尤其在通用任务、数学和编程挑战方面表现尤为突出。Qwen2.5-Plus在显著降低训练和推理成本的同时，与Qwen2.5-72B和Llama3-405B相比，取得了极具竞争力的性能表现，在Hellaswag、TheoremQA、MATH、GSM8K、MultiPL-E、Multi-Mathematics和Multi-Translation等任务上优于其他基线模型。此外，Qwen2.5-Plus在MMLU-Pro上取得了64.0的分数，比Qwen2.5-72B高出5.9分

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.,

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

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**

Qwen2.5-14B/32B 和 Qwen2.5-Turbo 的评估与相似规模的基线模型进行了对比。这些基线模型包括 Yi-1.5-34B（Young 等，2024）、Gemma2-27B（Gemma 团队等，2024）和 Qwen1.5-32B（Qwen 团队，2024b）。结果如表 3 所示。Qwen2.5-14B 模型在各种任务中表现出色，尤其在 MMLU 和 BBH 等通用任务中表现突出，分别取得了 79.7 和 78.2 的分数，超越了更大规模的竞争对手。与此同时，Qwen2.5-32B 展现了卓越的能力，在许多情况下超越了相似规模的更大模型。值得注意的是，它在数学和编程等具有挑战性的领域中显著优于其前身 Qwen1.5-32B，特别是在 MATH 和 MBPP 任务中分别取得了 57.7 和 84.5 的分数。对于 Qwen2.5-Turbo，尽管其训练成本和推理成本显著低于 Qwen2.5-14B，但其表现与之相当，其 MMLU-Pro 分数甚至优于 Qwen2.5-32B

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.

在7B级别的模型中，我们重点比较了Qwen2.5-7B与其他领先的7B+模型，包括Mistral-7B（Jiang等，2023a）、Llama3-8B（Dubey等，2024）、Gemma2-9B（Gemma团队等，2024）以及我们的前代模型Qwen2-7B（Yang等，2024a）。结果如表4所示。需要注意的是，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Qwen2.5-7B的非嵌入参数较少，但在众多基准测试中超越了其前代和同类模型。它在各种任务上表现出显著的改进，在通用基准测试如MMLU（Hendrycks等，2021a）上达到74.2分，在数学挑战如MATH（Hendrycks等，2021b）上达到49.8分，在编程任务如HumanEval（Chen等，2021）上达到57.9分

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).

对于边缘端模型，我们将Qwen2.5-0.5B、1.5B和3B与现有基线进行比较：Qwen2-0.5B/1.5B（Yang等，2024a）和Gemma2-2.6B（Gemma团队等，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

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.

为了批判性地评估指令调优模型，我们采用了多方面的研究方法。基础技能和人类偏好通过开放数据集和基准进行评估。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方面的表现。特别关注的是对长上下文能力的评估。接下来的章节将概述评估方法并展示结果

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).

为了全面评估指令调优模型的质量，我们结合了自动评估与人工评估，以衡量其能力及人类偏好。在基础能力评估方面，我们采用了与预训练模型评估相似的数据集，这些数据集聚焦于自然语言理解、编程、数学和推理能力。具体而言，我们使用MMLU-Pro、MMLU-redux和LiveBench 0831（White等人，2024年）进行通用评估，GPQA、GSM8K和MATH用于科学与数学领域，HumanEval、MBPP、MultiPL-E及LiveCodeBench 2305-2409（Jain等人，2024年）用于编程能力测试，IFEval（Zhou等人，2023年）则用于指令遵循能力的评估。此外，我们还通过MT-Bench（Zheng等人，2023年）和Arena-Hard（Li等人，2024年）等基准测试，评估了模型在人类偏好对齐及指令遵循方面的表现

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*.

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

如表6所示，我们将Qwen2.5-72B-Instruct和Qwen2.5-Plus与其他领先的开源权重指令调优模型进行了比较：Llama3.1-70B-Instruct（Dubey等人，2024）、Llama3.1-405B-Instruct（Dubey等人，2024）以及我们之前的72B版本Qwen2-72BInstruct（Yang等人，2024a）。Qwen2.5-72B-Instruct模型表现出色，甚至在多个关键基准测试中超越了更大的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个上表现优于Qwen2.5-72B-Instruct

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.

Qwen2.5-14B/32B-Instruct与Qwen2.5-Turbo的性能评估及与相似规模基线的对比 本研究对Qwen2.5-Turbo、Qwen2.5-14BInstruct及Qwen2.5-32B-Instruct模型的性能进行了评估，并与相似规模的基线模型进行了对比。基线模型包括GPT4o-mini、Gemma2-27B-IT（Gemma团队等，2024年）及Qwen2-57BA14BInstruct（Yang等，2024a）。评估结果汇总于表7。Qwen2.5-32B-Instruct模型在多数任务中相较于其他相似规模模型展现出更优的性能。值得注意的是，我们的开源权重模型Qwen2.5-14B-Instruct在所有基准测试中均取得了具有竞争力的结果，与GPT-4o-mini相媲美。尽管Qwen2.5-Turbo模型的训练与推理成本显著降低，其在十项基准测试中的八项上均超越了Qwen2.5-14B-Instruct。这表明Qwen2.5-Turbo在效率与效果上均取得了显著成就，使其成为资源受限环境下的理想选择

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.

其他指令调优模型 如表8所示，Qwen2.5-7B-Instruct模型在除IFEval外的所有任务中均显著优于其竞争对手Gemma2-9B-IT和Llama3.1-8B-Instruct。值得注意的是，Qwen2.5-7B-Instruct在数学（MATH: 75.5）和编程（HumanEval: 84.8）方面表现出明显优势。对于边缘端指令模型，Qwen2.5-3B-Instruct模型尽管参数数量少于Phi3.5-mini-instruct（Abdin等，2024）和MiniCPM3-4B-Instruct（Hu等，2024）模型，但在数学和编程任务中超越了它们，如表9所示。此外，它在语言理解方面也取得了具有竞争力的结果。Qwen2.5-1.5B-Instruct和Qwen2.5-0.5B-Instruct模型相较于其先前版本也实现了显著的性能提升，具体如表10所示。这些改进使它们特别适合在资源高度受限的环境中进行边缘端应用

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

尽管已有多个开放的基准数据集可供评估，但我们认为这些数据集尚不足以全面捕捉大型语言模型（LLMs）的能力。为此，我们开发了一系列内部数据集，旨在评估模型在知识理解、文本生成、编码等多个方面的表现。这些评估分别在中文和英文环境下进行。此外，我们还特别评估了指令调优模型的多语言性能。评估结果分别总结在表11（英文）、表12（中文）、表13（70B+指令模型的多语言性能）和表14（7B-14B模型）中

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.

中英文评估 我们比较了Qwen2.5-Instruct模型与包括GPT-4、Claude3.5-sonnet、Qwen2和Llama-3.1在内的多个领先语言模型在英文和中文上的表现。我们的分析聚焦于模型规模及其对性能的影响，以及我们最新的Qwen2.5系列与之前版本及竞争模型的对比情况。对于较小模型，我们观察到Qwen2.5-0.5B模型的表现与Qwen2-1.5B模型相当甚至超越。这表明Qwen2.5系列优化了参数使用，使得中等规模模型能够达到与上一代更大模型相似的性能水平。Qwen2.5-3B模型的表现与Qwen2-7B模型相当。值得注意的是，Qwen2.5-32B模型相较于Qwen2-72B模型展现出显著提升。 我们的旗舰模型Qwen2.5-72B进一步缩小了Qwen与GPT-4、Claude3.5-sonnet等顶尖模型之间的差距。特别是，Qwen2.5-72B在除指令跟随外的所有指标上均与Llama-3.1-405B相当或超越。这一成就凸显了Qwen2.5-72B在广泛语言处理任务中的竞争力，同时也指出了未来改进的方向。Qwen2.5-Plus针对之前中文指令跟随的不足进行了改进，并进一步增强了其他领域的优势

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

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

多语言评估 为了全面评估指令调优模型的多语言能力，我们遵循P-MMEval（Zhang等，2024）的方法，并扩展了以下几个基准测试：(1) IFEval（多语言）：我们将原本为英语的IFEval基准扩展到包含多语言示例。为确保语言中立性，我们移除了包含语言特定内容（例如，“以字母A开头”）的实例。(2) 知识利用：为了评估Qwen2.5系列模型在多种语言中的知识利用能力，我们采用了五个类似MMLU的基准测试（多项选择格式）。这些基准测试包括：AMMLU（阿拉伯语）、JMMLU（日语）、KMMLU（韩语）、IndoMMLU（印尼语）和TurkishMMLU（土耳其语）。此外，我们还评估了模型在翻译版MMLU基准（okapi MMLU）上的表现，该基准已从其原始英语形式适应为多种语言。(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：为了评估模型捕捉文化细微差别的能力，我们使用了BLEnD基准（Myung等，2024）。该基准专门设计用于测试LLM对文化细微差别的理解

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在指令遵循、多语言知识及数学推理方面展现出与同类规模模型相媲美的竞争力。尽管相较于前代Qwen2，其在捕捉文化细微差异方面取得了显著进步，但在这一领域仍有进一步优化的空间

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

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**

奖励模型作为指导强化学习（RL）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评估。我们的评估基准包括Reward Bench（Lambert等，2024）、RMB（Zhou等，2024）、PPE（Frick等，2024b）以及内部收集的跨领域中文人类偏好基准（Human-Preference-Chinese），以提供全面的分析。为了进行比较，我们纳入了基线模型，如Nemotron-4-340BReward（Adler等，2024）、Llama-3.1-Nemotron-70B-Reward（Wang等，2024c）和Athene-RM70B（Frick等，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Reward Bench上表现优异，而Athene-RM-70B在RMB基准上表现最佳。Qwen2.5-RM-72B在PPE和Human-Preference-Chinese评估中领先，在RMB上仅次于Athene-RM-70B，并在Reward Bench上达到了与Nemotron-4-340B-Reward相当的性能水平，尽管略逊于Llama-3.1-Nemotron-70B-Reward

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.

由于缺乏对奖励模型的评估方法，当前的奖励模型通常使用Reward Bench进行评估。然而，我们从多个奖励模型基准测试中的评估结果表明，过度优化特定基准可能会触发Goodhart定律（Hoskin, 1996），导致在其他基准上的性能下降，并可能影响下游的对齐性能。这凸显了需要在多样化的基准上对奖励模型进行全面评估，而不是仅仅依赖单一基准

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.

更重要的是，通过迭代实验，我们也认识到一个关键限制：当前的奖励模型评估基准并不能准确预测在其指导下训练的强化学习模型的表现。换言之，奖励模型基准测试中的较高分数并不必然意味着由此产生的强化学习模型具有更优的性能。这一认识强调了进一步研究更具预测性的奖励模型评估方法的必要性

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.

我们采用三个基准来评估Qwen2.5模型的长上下文能力：RULER（Hsieh等人，2024年）、LV-Eval（Yuan等人，2024年）以及Longbench-Chat（Bai等人，2024年）。在LV-Eval中，我们采用关键词召回率作为报告分数，以缓解原始指标中高假阴性率的问题

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.

结果如表16和表17所示。我们可以观察到，配备了长度外推技术（即DCA + YARN）的Qwen2.5模型在三个数据集上展示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表现最为突出，显著优于现有的开源长上下文模型以及GPT-4o-mini和GPT-4等专有模型

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 introduce a sparse attention mechanism 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.

此外，如图2所示，Qwen2.5-Turbo在1M-token的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捉细节信息的卓越能力。我们引入了一种稀疏注意力机制，显著提升了推理速度，这对于处理长上下文时的用户体验至关重要。对于1M token的序列，该方法将注意力机制的计算负载减少了12.5倍。图3展示了Qwen2.5-Turbo在不同硬件配置下的首次令牌时间（TTFT），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

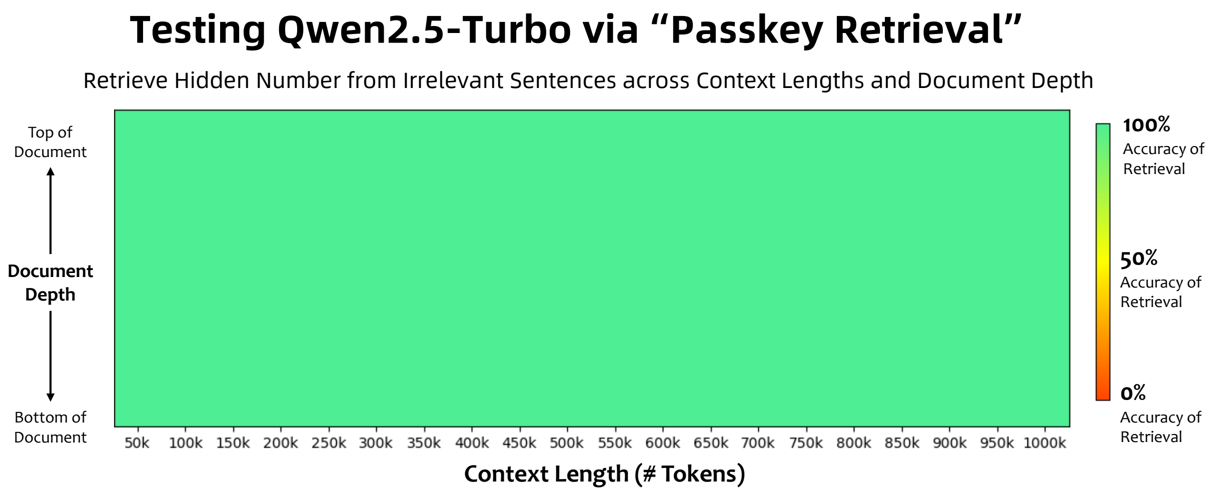

Figure 2: **Performance of Qwen2.5-Turbo on Passkey Retrieval Task with 1M Token Lengths.**

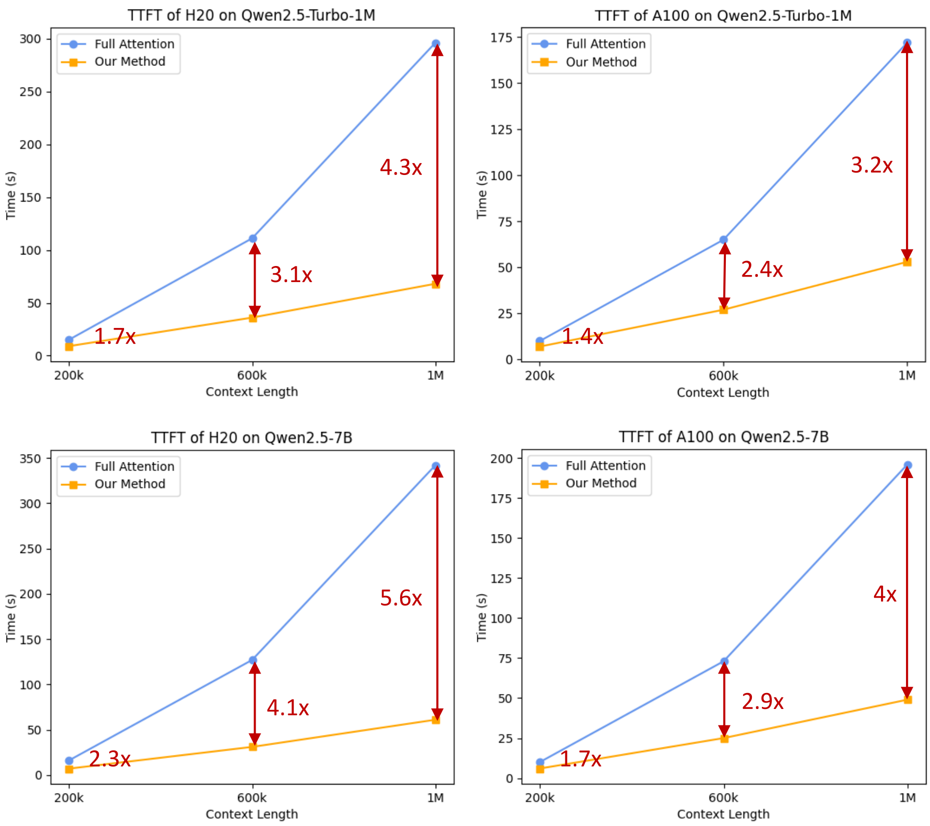

Figure 3: **TTFT (Time To First Token) of Qwen2.5-Turbo and Qwen2.5-7B with Full Attention and Our Method.**

Qwen2.5 在大规模语言模型（LLMs）领域取得了显著进展，通过增强的预训练技术，在18万亿个标记上进行训练，并采用了包括监督微调和多阶段强化学习在内的精细化后训练方法。这些改进提升了模型在人类偏好对齐、长文本生成以及结构化数据分析方面的能力，使得Qwen2.5在执行指令跟随任务时表现出色。Qwen2.5 提供多种配置，既有从0.5B到72B参数的开源权重模型，也有专有模型，其中包括经济高效的混合专家（MoE）变体，如Qwen2.5-Turbo和Qwen2.5-Plus。 实证评估表明，尽管Qwen2.5-72B-Instruct的规模仅为当前最先进的Llama-3-405B-Instruct的六分之一，但其性能与之相当。此外，Qwen2.5 还作为专用模型的基础，展示了其在特定领域应用中的广泛适应性。我们相信，Qwen2.5 的强大性能、灵活架构以及广泛的可用性，使其成为学术研究和工业应用的宝贵资源，并有望成为未来创新的关键推动者

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.

未来，我们将致力于推进稳健的基础模型发展。首先，通过融入更广泛、更多样化、更高质量的数据，我们将迭代优化基础及指令调优的大型语言模型（LLMs）。其次，我们也将持续开发多模态模型，旨在将多种模态整合至一个统一框架内，以此促进跨文本、视觉和听觉领域的无缝端到端信息处理。第三，我们致力于提升模型的推理能力，这将通过战略性扩展推理计算资源来实现。这些努力旨在突破当前技术限制的边界，为人工智能的广阔领域做出贡献

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, Tingyu Xia, 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 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.

Nvidia 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`.

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.

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, 2024.

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

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.

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.