# Qwen1.5-110B：Qwen1.5系列的首个千亿参数开源模型

2024年4月25日 · 1 分钟 · 113 字 · Qwen Team | 语言:

GITHUB HUGGING FACE MODELSCOPE DEMO DISCORD

## 简介#

近期开源社区陆续出现了千亿参数规模以上的大模型，这些模型都在各项评测中取得杰出的成绩。今天，我们开源1100亿参数的Qwen1.5系列首个千亿参数模型Qwen1.5-110B，该模型在基础能力评估中与Meta-Llama3-70B相媲美，在Chat评估中表现出色，包括MT-Bench和AlpacaEval 2.0。

## 模型特性#

Qwen1.5-110B与其他Qwen1.5模型相似，采用了相同的Transformer解码器架构。它包含了分组查询注意力（GQA），在模型推理时更加高效。该模型支持32K tokens的上下文长度，同时它仍然是多语言的，支持英、中、法、西、德、俄、日、韩、越、阿等多种语言。

## 模型效果#

我们对基础语言模型进行了一系列评估，并与最近的SOTA语言模型Meta-Llama3-70B以及Mixtral-8x22B进行了比较。

| QWEN1.5-110B | QWEN1.5-72B | LLAMA-3-70B | MIXTRAL-8X22B |
|---|---|---|---|
| MMLU | 80.4 | 77.5 | 79.5 |
| TheoremQA | 34.9 | 29.3 | 32.0 |
| GPQA | 35.9 | 36.3 | 36.4 |
| Hellaswag | 87.5 | 86.0 | 88.0 |
| BBH | 74.8 | 65.5 | 76.6 |
| ARC-C | 69.6 | 65.9 | 68.8 |
| GSM8K | 85.4 | 79.5 | 79.2 |
| MATH | 49.6 | 34.1 | 41.0 |
| HumanEval | 52.4 | 41.5 | 45.7 |
| MBPP | 58.1 | 53.4 | 55.1 |

上述结果显示，新的110B模型在基础能力方面至少与Llama-3-70B模型相媲美。在这个模型中，我们没有对预训练的方法进行大幅改变，因此我们认为与72B相比的性能提升主要来自于增加模型规模。

我们还在MT-Bench和AlpacaEval 2.0上进行了Chat评估，结果如下：

| Models | MT-Bench | AlpacaEval 2.0 |
|---|---|---|
| | Avg. Score | LC Win Rate |
| Llama-3-70B-Instruct | 8.85 | 34.40 |
| Qwen1.5-72B-Chat | 8.61 | 36.60 |
| Qwen1.5-110B-Chat | 8.88 | 43.90 |

与之前发布的72B模型相比，在两个Chat模型的基准评估中，110B表现显著更好。评估结果的持续改善表明，即使在没有大幅改变后训练方法的情况下，更强大、更大规模的基础语言模型也可以带来更好的Chat模型。

## 使用Qwen1.5-110B#

我们建议您阅读Qwen1.5的博客了解更多关于在transformers、llama.cpp、vLLM、Ollama、LMStudio、SkyPilot、Axolotl、LLaMA-Factory等框架上使用的方法。

## 结语#

Qwen1.5-110B是Qwen1.5系列中规模最大的模型，也是该系列中首个拥有超过1000亿参数的模型。它在与最近发布的SOTA模型Llama-3-70B的性能上表现出色，并且明显优于72B模型。这告诉我们，在模型大小扩展方面仍有很大的提升空间。虽然Llama-3的发布表明预训练数据规模具有重要意义，但我们相信通过在未来的发布中同时扩展数据和模型大小，我们可以同时获得两者的优势。敬请期待Qwen2！

## 引用#

```
@misc{qwen1.5,
title = {Introducing Qwen1.5},
url = {https://qwenlm.github.io/blog/qwen1.5/},
author = {Qwen Team},
month = {February},
year = {2024}
}
```