QWEN TECHNICAL REPORT

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, Tianhang Zhu.

Qwen Team, Alibaba Group*

ABSTRACT

Large language models (LLMs) have revolutionized the field of artificial intelligence, enabling natural language processing tasks that were previously thought to be exclusive to humans. In this work, we introduce QWEN[1], the first installment of our large language model series. QWEN is a comprehensive language model series that encompasses distinct models with varying parameter counts. It includes QWEN, the base pretrained language models, and QWEN-CHAT, the chat models finetuned with human alignment techniques. The base language models consistently demonstrate superior performance across a multitude of downstream tasks, and the chat models, particularly those trained using Reinforcement Learning from Human Feedback (RLHF), are highly competitive. The chat models possess advanced tool-use and planning capabilities for creating agent applications, showcasing impressive performance even when compared to bigger models on complex tasks like utilizing a code interpreter. Furthermore, we have developed coding-specialized models, CODE-QWEN and CODE-QWEN-CHAT, as well as mathematics-focused models, MATH-QWEN-CHAT, which are built upon base language models. These models demonstrate significantly improved performance in comparison with open-source models, and slightly fall behind the proprietary models.

*Authors are ordered alphabetically by the last name. Correspondence to: ericzhou.zc@alibaba-inc.com.

[1]QWEN is a moniker of Qianwen, which means "thousands of prompts" in Chinese. The pronunciation of "QWEN" can vary depending on the context and the individual speaking it. Here is one possible way to pronounce it: /kwɛn/.

Contents

1 INTRODUCTION

Large language models (LLMs) (Radford et al., 2018; Devlin et al., 2018; Raffel et al., 2020; Brown et al., 2020; OpenAI, 2023; Chowdhery et al., 2022; Anil et al., 2023; Thoppilan et al., 2022; Touvron et al., 2023a;b) have revolutionized the field of artificial intelligence (AI) by providing a powerful foundation for complex reasoning and problem-solving tasks. These models have the ability to compress vast knowledge into neural networks, making them incredibly versatile agents. With a chat interface, LLMs can perform tasks that were previously thought to be the exclusive domain of humans, especially those involving creativity and expertise (OpenAI, 2022; Ouyang et al., 2022; Anil et al., 2023; Google, 2023; Anthropic, 2023a;b). They can engage in natural language conversations with humans, answering questions, providing information, and even generating creative content such as stories, poems, and music. This has led to the development of a wide range of applications, from chatbots and virtual assistants to language translation and summarization tools.

LLMs are not just limited to language tasks. They can also function as a generalist agent (Reed et al., 2022; Bai et al., 2022a; Wang et al., 2023a; AutoGPT, 2023; Hong et al., 2023), collaborating with external systems, tools, and models to achieve the objectives set by humans. For example, LLMs can understand multimodal instructions (OpenAI, 2023; Bai et al., 2023; Liu et al., 2023a; Ye et al., 2023; Dai et al., 2023; Peng et al., 2023b), execute code (Chen et al., 2021a; Zheng et al., 2023; Li et al., 2023d), use tools (Schick et al., 2023; LangChain, Inc., 2023; AutoGPT, 2023), and more. This opens up a whole new world of possibilities for AI applications, from autonomous vehicles and robotics to healthcare and finance. As these models continue to evolve and improve, we can expect to see even more innovative and exciting applications in the years to come. Whether it's helping us solve complex problems, creating new forms of entertainment, or transforming the way we live and work, LLMs are poised to play a central role in shaping the future of AI.

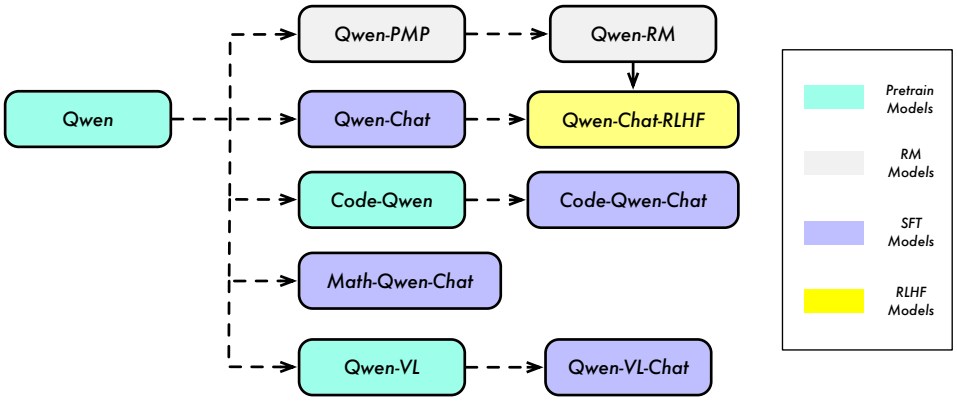

Figure 1: **Model Lineage of the Qwen Series.** We have pretrained the language models, namely QWEN, on massive datasets containing trillions of tokens. We then use SFT and RLHF to align QWEN to human preference and thus we have QWEN-CHAT and specifically its improved version QWEN-CHAT-RLHF. Additionally, we also develop specialized models for coding and mathematics, such as CODE-QWEN, CODE-QWEN-CHAT, and MATH-QWEN-CHAT based on QWEN with similar techniques. Note that we previously released the multimodal LLM, QWEN-VL and QWEN-VL-CHAT (Bai et al., 2023), which are also based on our QWEN base models.

Despite their impressive capabilities, LLMs are often criticized for their lack of reproducibility, steerability, and accessibility to service providers. In this work, we are pleased to present and release the initial version of our LLM series, QWEN. QWEN is a moniker that derives from the Chinese phrase Qianwen, which translates to "thousands of prompts" and conveys the notion of embracing a wide range of inquiries. QWEN is a comprehensive language model series that encompasses distinct models with varying parameter counts. The model series include the base pretrained language models, chat models finetuned with human alignment techniques, i.e., supervised finetuning (SFT), reinforcement learning with human feedback (RLHF), etc., as well as specialized models in coding and math. The details are outlined below:

1. The base language models, namely QWEN, have undergone extensive training using up to 3 trillion tokens of diverse texts and codes, encompassing a wide range of areas. These models have consistently demonstrated superior performance across a multitude of downstream tasks, even when compared to their more significantly larger counterparts.
2. The QWEN-CHAT models have been carefully finetuned on a curated dataset relevant to task performing, chat, tool use, agent, safety, etc. The benchmark evaluation demonstrates that the SFT models can achieve superior performance. Furthermore, we have trained reward models to mimic human preference and applied them in RLHF for chat models that can produce responses preferred by humans. Through the human evaluation of a challenging test, we find that QWEN-CHAT models trained with RLHF are highly competitive, still falling behind GPT-4 on our benchmark.
3. In addition, we present specialized models called CODE-QWEN, which includes CODE-QWEN-7B and CODE-QWEN-14B, as well as their chat models, CODE-QWEN-14B-CHAT and CODE-QWEN-7B-CHAT. Specifically, CODE-QWEN has been pre-trained on extensive datasets of code and further fine-tuned to handle conversations related to code generation, debugging, and interpretation. The results of experiments conducted on benchmark datasets, such as HumanEval (Chen et al., 2021b), MBPP (Austin et al., 2021), and HumanEvalPack (Muennighoff et al., 2023), demonstrate the high level of proficiency of CODE-QWEN in code understanding and generation.
4. This research additionally introduces MATH-QWEN-CHAT specifically designed to tackle mathematical problems. Our results show that both MATH-QWEN-7B-CHAT and MATH-QWEN-14B-CHAT outperform open-sourced models in the same sizes with large margins and are approaching GPT-3.5 on math-related benchmark datasets such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021).
5. Besides, we have open-sourced QWEN-VL and QWEN-VL-CHAT, which have the versatile ability to comprehend visual and language instructions. These models outperform the current open-source vision-language models across various evaluation benchmarks and support text recognition and visual grounding in both Chinese and English languages. Moreover, these models enable multi-image conversations and storytelling. Further details can be found in Bai et al. (2023).

Now, we officially open-source the 14B-parameter and 7B-parameter base pretrained models QWEN and aligned chat models QWEN-CHAT[2]. This release aims at providing more comprehensive and powerful LLMs at developer- or application-friendly scales.

The structure of this report is as follows: Section 2 describes our approach to pretraining and results of QWEN. Section 3 covers our methodology for alignment and reports the results of both automatic evaluation and human evaluation. Additionally, this section describes details about our efforts in building chat models capable of tool use, code interpreter, and agent. In Sections 4 and 5, we delve into specialized models of coding and math and their performance. Section 6 provides an overview of relevant related work, and Section 7 concludes this paper and points out our future work.

2 PRETRAINING

The pretraining stage involves learning vast amount of data to acquire a comprehensive understanding of the world and its various complexities. This includes not only basic language capabilities but also advanced skills such as arithmetic, coding, and logical reasoning. In this section, we introduce the data, the model design and scaling, as well as the comprehensive evaluation results on benchmark datasets.

2.1 DATA

The size of data has proven to be a crucial factor in developing a robust large language model, as highlighted in previous research (Hoffmann et al., 2022; Touvron et al., 2023b). To create an effective pretraining dataset, it is essential to ensure that the data are diverse and cover a wide range

[2]GitHub: `https://github.com/QwenLM/Qwen`.

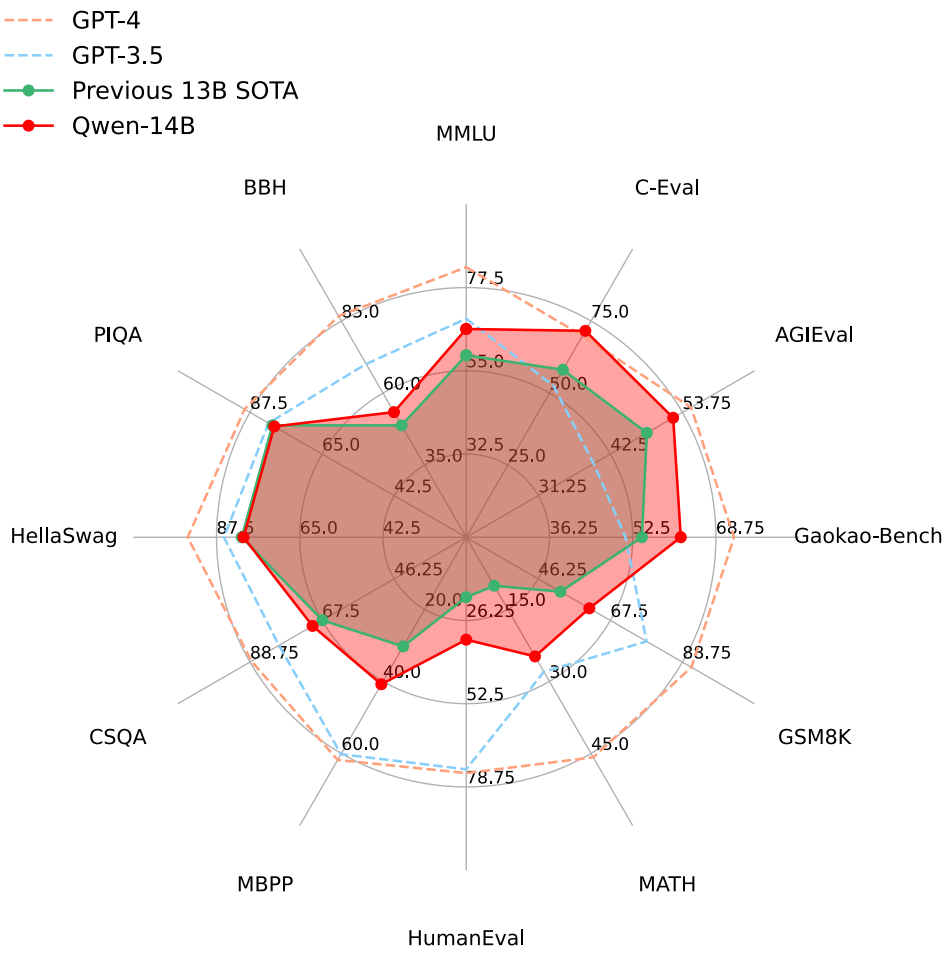

Figure 2: **Performance of GPT-4, GPT-3.5, the previous 13B SOTA, as well as QWEN-14B.** We demonstrate the results on 12 datasets covering multiple domains, including language understanding, knowledge, reasoning, etc. QWEN significantly outperforms the previous SOTA of similar model sizes, but still lag behind both GPT-3.5 and GPT-4.

of types, domains, and tasks. Our dataset is designed to meet these requirements and includes public web documents, encyclopedia, books, codes, etc. Additionally, our dataset is multilingual, with a significant portion of the data being in English and Chinese.

To ensure the quality of our pretraining data, we have developed a comprehensive data preprocessing procedure. For public web data, we extract text from HTML and use language identification tools to determine the language. To increase the diversity of our data, we employ deduplication techniques, including exact-match deduplication after normalization and fuzzy deduplication using MinHash and LSH algorithms. To filter out low-quality data, we employ a combination of rule-based and machine-learning-based methods. Specifically, we use multiple models to score the content, including language models, text-quality scoring models, and models for identifying potentially offensive or inappropriate content. We also manually sample texts from various sources and review them to ensure their quality. To further enhance the quality of our data, we selectively up-sample data from certain sources, to ensure that our models are trained on a diverse range of high-quality content. Finally, we have built a dataset of up to 3 trillion tokens.