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`.