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.

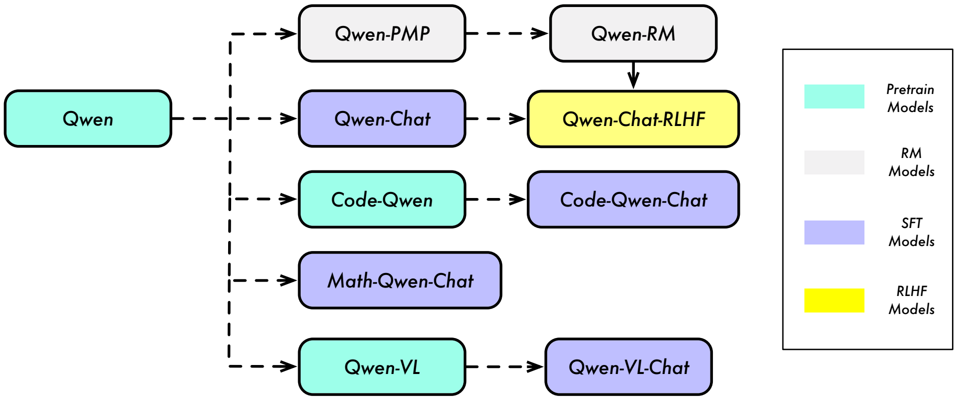

Figure 1: **Model Lineage of the Qwen Series.** We have pretrained the language models, namely Qwen, on massive datasets containing trillions of tokens. We then use SFT and RLHF to align Qwen to human preference and thus we have Qwen-Chat and specifically its improved version Qwen-Chat-RLHF. Additionally, 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: