in the math SFT dataset are examination questions, and it is easy for the model to predict the input format and it is meaningless for the model to predict the input condition and numbers which could be random. Thus, we mask the inputs of the system and user to avoid loss computation on them and find masking them accelerates the convergence during our preliminary experiments. For optimization, we use the AdamW optimizer with the same hyperparameters of SFT except that we use a peak learning rate of 2×10^{-5} and a training step of $50\,000$.

5.2 Evaluation

We evaluate models on the test sets of GSM8K (Grade school math) (Cobbe et al., 2021), MATH (Challenging competition math problems) (Hendrycks et al., 2021), Math401 (Arithmetic ability) (Yuan et al., 2023b), and Math23K (Chinese grade school math) (Wang et al., 2017). We compare MATH-QWEN with proprietary models ChatGPT and Minerva (Lewkowycz et al., 2022) and open-sourced math-specialized model RFT (Yuan et al., 2023a), WizardMath (Luo et al., 2023a), and GAIRMath-Abel (Chern et al., 2023a) in Table 12. MATH-QWEN-CHAT models show better math reasoning and arithmetic abilities compared to open-sourced models and QWEN-CHAT models of similar sizes. Compared to proprietary models, MATH-QWEN-CHAT-7B outperforms Minerva-8B in MATH. MATH-QWEN-14B-CHAT is chasing Minerva-62B and GPT-3.5 in GSM8K and MATH and delivers better performance on arithmetic ability and Chinese math problems.

6 Related Work

6.1 Large Language Models

The excitement of LLM began with the introduction of the Transformer architecture (Vaswani et al., 2017), which was then applied to pretraining large-scale data by researchers such as Radford et al. (2018); Devlin et al. (2018); Liu et al. (2019). These efforts led to significant success in transfer learning, with model sizes growing from 100 million to over 10 billion parameters (Raffel et al., 2020; Shoeybi et al., 2019).

In 2020, the release of GPT-3, a massive language model that is 10 times larger than T5, demonstrated the incredible potential of few-shot and zero-shot learning through prompt engineering and in-context learning, and later chain-of-thought prompting (Wei et al., 2022c). This success has led to a number of studies exploring the possibilities of further scaling these models (Scao et al., 2022; Zhang et al., 2022; Du et al., 2021; Zeng et al., 2022; Lepikhin et al., 2020; Fedus et al., 2022; Du et al., 2022; Black et al., 2022; Rae et al., 2021; Hoffmann et al., 2022; Chowdhery et al., 2022; Thoppilan et al., 2022). As a result, the community has come to view these large language models as essential foundations for downstream models (Bommasani et al., 2021).

The birth of ChatGPT (OpenAI, 2022) and the subsequent launch of GPT-4 (OpenAI, 2023) marked two historic moments in the field of artificial intelligence, demonstrating that large language models (LLMs) can serve as effective AI assistants capable of communicating with humans. These events have sparked interests among researchers and developers in building language models that are aligned with human values and potentially even capable of achieving artificial general intelligence (AGI) (Anil et al., 2023; Anthropic, 2023a;b).

One notable development in this area is the emergence of open-source LLMs, specifically LLaMA (Touvron et al., 2023a) and LLAMA 2 (Touvron et al., 2023b), which have been recognized as the most powerful open-source language models ever created. This has led to a surge of activity in the open-source community (Wolf et al., 2019), with a series of large language models being developed collaboratively to build upon this progress (Mosaic ML, 2023; Almazrouei et al., 2023; ChatGLM2 Team, 2023; Yang et al., 2023; InternLM Team, 2023).

6.2 Alignment

The community was impressed by the surprising effectiveness of alignment on LLMs. Previously, LLMs without alignment often struggle with issues such as repetitive generation, hallucination, and deviation from human preferences. Since 2021, researchers have been diligently working on developing methods to enhance the performance of LLMs in downstream tasks (Wei et al., 2022a;