A Appendix

A.1 More Training Details

A.1.1 Data Format for Qwen-Chat

Different from conventional pretraining based on autoregressive next-token prediction, despite using a similar training task, there should be a specially design data format for SFT and RLHF to build a conversational AI assistant model. Common formats include "human-assistant" and ChatML formats. As to our knowledge, the earliest example of the human-assistant format comes from Anthropic (Bai et al., 2022b), which adds a special phrase "`\n\nhuman: `" in front of the user input and "`\n\nassistant: `" in front of the assistant response. It is easy for the base language model to transfer to the pattern of conversational AI. However, as the specific phrases are common words, it might be hard for the model to disambiguate from these words in other contexts.

Instead, we turned to the ChatML format proposed by OpenAI.[5] This format allows the use of special tokens, i.e., "`<im_start>`" and "`<im_end>`", that do not appear in pretraining, and thus resolve the aforementioned problem. We demonstrate an example of the format below.

ChatML Format

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
Hello!<|im_end|>
<|im_start|>assistant
Hello! How can I assist you today?<|im_end|>
```

A.2 Evaluation

A.2.1 Automatic Evaluation

To provide a whole picture of the performance of our model series Qwen, here in this section we illustrate the detailed performance of our models as well as the baselines in the comprehensive benchmark evaluation proposed by OpenCompass Team (2023). We report the results in multiple tables based on the officially provided categories, including examination, language, knowledge, understanding, and reasoning. In terms of the performance of the baseline models, we report the higher results between the reported ones and those on the leaderboard.

Examination Here we evaluate the models on a series of datasets relevant to the examination. The datasets include

- **MMLU** (Hendrycks et al., 2020) Massive Multi-task Language Understanding is designed for measuring language understanding capabilities. We report 5-shot results.
- **C-Eval** (Huang et al., 2023) C-Eval is a Chinese evaluation dataset spanning 52 diverse disciplines. We report 5-shot results.
- **CMMLU** (Li et al., 2023c) CMMLU is designed for assessing language understanding capabilities in Chinese. We report 5-shot results.
- **AGIEval** (Zhong et al., 2023a) This is a benchmark consisting of human-centric examinations, including college entrance exams, law school admission tests, math competitions, and lawyer qualification tests. We report zero-shot results.
- **Gaokao-Bench** (Zhang et al., 2023b) This is a benchmark with Gaokao (Chinese college-entrance examination) questions. We report zero-shot results.
- **ARC** (Clark et al., 2018) ARC is a dataset consisting of grade-school level, multiple-choice science questions. It includes an easy set and a challenge set, which are referred by ARC-e and ARC-c. We report zero-shot results.