Despite this, we have reservations about the ability of traditional benchmark evaluation to accurately measure the performance and potential of chat models trained with alignment techniques in today's landscape. The results mentioned earlier provide some evidence of our competitive standing, but we believe that it is crucial to develop new evaluation methods specifically tailored to aligned models.

We believe that human evaluation is crucial, which is why we have created a carefully curated dataset for this purpose. Our process involved collecting 300 instructions in Chinese that covered a wide range of topics, including knowledge, language understanding, creative writing, coding, and mathematics. To evaluate the performance of different models, we chose the SFT version of QWEN-CHAT-7B and the SFT and RLHF version of QWEN-CHAT-14B, and added two strong baselines, GPT-3.5 and GPT-4[4], for comparison. For each instruction, we asked three annotators to rank the model responses by the overall score of helpfulness, informativeness, validity, and other relevant factors. Our dataset and evaluation methodology provides a comprehensive and rigorous assessment of the capabilities of different language models in various domains.

Figure 4 illustrates the win rates of the various models. For each model, we report the percentage of wins, ties, and losses against GPT-3.5, with the segments of each bar from bottom to top representing these statistics. The experimental results clearly demonstrate that the RLHF model outperforms the SFT models by significant margins, indicating that RLHF can encourage the model to generate responses that are more preferred by humans. In terms of overall performance, we find that the RLHF model significantly outperforms the SFT models, slightly falling behind GPT-4. This indicates the effectiveness of RLHF for aligning to human preference. To provide a more comprehensive understanding of the models' performance, we include a case study with examples from different models in Appendix A.2.2.

3.4 TOOL USE, CODE INTERPRETER, AND AGENT

Table 6: Performance of QWEN on the in-house Chinese benchmark that evaluates its ability to use unseen tools via ReAct prompting.

Model	Params	Tool Selection (Acc.↑)	Tool Input (Rouge-L↑)	False Positive Error (%)↓
GPT-4	-	95	90	15.0
GPT-3.5	-	85	88	75.0
QWEN-CHAT	1.8B	92	89	19.3
	7B	**98**	91	7.3
	14B	**98**	**93**	**2.4**

The QWEN models, which are designed to be versatile, have the remarkable ability to assist with (semi-)automating daily tasks by leveraging their skills in tool-use and planning. As such, they can serve as agents or copilots to help streamline various tasks. We explore QWEN's proficiency in the following areas:

- Utilizing unseen tools through ReAct prompting (Yao et al., 2022) (see Table 6).
- Using a Python code interpreter to enhance math reasoning, data analysis, and more (see Table 7 and Table 8).
- Functioning as an agent that accesses Hugging Face's extensive collection of multimodal models while engaging with humans (see Table 9).

To enhance QWEN's capabilities as an agent or copilot, we employ the self-instruct (Wang et al., 2023c) strategy for supervised fine-tuning (SFT). Specifically, we utilize the in-context learning capability of QWEN for self-instruction. By providing a few examples, we can prompt QWEN to generate more relevant queries and generate outputs that follow a specific format, such as ReAct. We then apply rules and involve human annotators to filter out any noisy samples. Afterwards, the samples are incorporated into QWEN's training data, resulting in an updated version of QWEN that is more dependable for self-instruction. We iterate through this process multiple times until we gather

[4]To obtain the results from the models, we use the OpenAI APIs of GPT-3.5-turbo-0613 and GPT-4-0613.