Table 9: Results of QWEN-Chat on the Hugging Face Agent benchmark.

Task	Model	Params	Metric		
			Tool Selection ↑	Tool Used ↑	Code Correctness ↑
Run Mode	GPT-4	-	100	100	97.4
	GPT-3.5	-	95.4	96.3	87.0
	Starcoder-Base	15B	86.1	87.0	68.9
	Starcoder	15B	87.0	88.0	68.9
	QWEN-CHAT	1.8B	85.2	84.3	61.1
		7B	87.0	87.0	71.5
		14B	93.5	94.4	87.0
Chat Mode	GPT-4	-	97.9	97.9	98.5
	GPT-3.5	-	97.3	96.8	89.6
	Starcoder-Base	15B	97.9	97.9	91.1
	Starcoder	15B	97.9	97.9	89.6
	QWEN-CHAT	1.8B	93.6	93.6	73.2
		7B	94.7	94.7	85.1
		14B	97.9	97.9	95.5

an ample number of samples that possess both exceptional quality and a wide range of diversity. As a result, our final collection consists of around 2000 high-quality samples.

During the fine-tuning process, we mix these high-quality samples with all the other general-purpose SFT samples, rather than introducing an additional training stage. By doing so, we are able to retain essential general-purpose capabilities that are also pertinent for constructing agent applications.

Using Tools via ReAct Prompting We have created and made publicly available a benchmark for evaluating QWEN's ability to call plugins, tools, functions, or APIs using ReAct Prompting (see Qwen Team, Alibaba Cloud, 2023b). To ensure fair evaluation, we have excluded any plugins that were included in QWEN's training set from the evaluation set. The benchmark assesses the model's accuracy in selecting the correct plugin from a pool of up to five candidates, as well as the plausibility of the parameters passed into the plugin and the frequency of false positives. In this evaluation, a false positive occurs when the model incorrectly invokes a plugin in response to a query, despite not being required to do so.

The results presented in Table 6 demonstrate that QWEN consistently achieves higher accuracy in identifying the relevance of a query to the available tools as the model size increases. However, the table also highlights that beyond a certain point, there is little improvement in performance when it comes to selecting the appropriate tool and providing relevant arguments. This suggests that the current preliminary benchmark may be relatively easy and may require further enhancement in future iterations. It is worth noting that GPT-3.5 stands out as an exception, displaying suboptimal performance on this particular benchmark. This could potentially be attributed to the fact that the benchmark primarily focuses on the Chinese language, which may not align well with GPT-3.5's capabilities. Additionally, we observe that GPT-3.5 tends to attempt to use at least one tool, even if the query cannot be effectively addressed by the provided tools.

Using Code Interpreter for Math Reasoning and Data Analysis The Python code interpreter is widely regarded as a powerful tool for augmenting the capabilities of an LLM agent. It is worth investigating whether QWEN can harness the full potential of this interpreter to enhance its performance in diverse domains, such as mathematical reasoning and data analysis. To facilitate this exploration, we have developed and made publicly available a benchmark that is specifically tailored for this purpose (see Qwen Team, Alibaba Cloud, 2023a).

The benchmark encompasses three primary categories of tasks: math problem-solving, data visualization, and other general-purpose tasks like file post-processing and web crawling. Within the