Table 4: List of all Apps and their corresponding functions in the TRAILBench.

Domain	APPs	Functions
Shopping	Taobao, JD, Pinduoduo	*search_goods*
Transport	Hello, Didi, FlowerPig, T3, GaodeMap, BaiduMap TencentMap, Ctrip, Fliggy, Qunar, CEAir, 12306	*get_recommend_video*, *get_taxi*, *find_flights*, *find_stations*, *find_trains*
Entertainment	Bilibili, Douyin, Kuaishou, IQIYI TencentVideo, MangoTV, Maoyan, Damai	*get_recommend_video*, *search_video*, *search_movie*, *search_show*
Food	Meituan, Eleme, TaobaoWaimai, JDwaimai, Dianping, KFC, McDonalds	*search_restaurant*, *recommend_restaurant*
Knowledge	Xiaohongshu, Zhihu, Google	*get_knowledge*
Reminder	SystemClock	*add_event*, *add_alarm*, *view_event_by_time*
News	Toutiao, Zhihu, Weibo, Hupu	*get_daily_news*
sport$_{a}nd_{h}ealth$	AppleHealth, HuaweiHealth, Keep	*search_exercise_plan*, *log_exercise*

Table 5: Statistics of the datasets (by Scenario).

Scenario	#APPs(Tools)	#Functions	avg #param	avg #required
Food	7	11	9	2
Shopping	3	3	6	1
Transport	12	13	6.38	3.23
Reminder	1	3	3	2.67
Entertainment	8	15	5.13	0.67
Sport & Health	2	4	2.75	1.5
Knowledge	3	3	4	1
News	3	3	1	0

A The Use of LLM

In this paper, Large language models (LLMs) were used as a general-purpose assist tool to improve the clarity and grammar of the manuscript. The models were not used for research ideation, data analysis, or the generation of any core content. Their role was limited to minor editing and polishing of the text to enhance readability.

B Dataset Details

This section shows all Apps and APIs defined in our benchmark. The statistics are shown in 4. Besides, we provided a detailed display of the number of apps, functions, average parameters, and required parameters in each of scenarios. The statistics are illustrated in 5.

C Experiment Setup Details

C.1 Evaluation Metrics

The calculation of various metrics in PTBench are formulated as follows:

- **Format Accuracy** refers to the proportion of the LLM's generated output in conforming to our required output template, which indicates the instruction-following ability.

$$\text{format\_acc} = \frac{\#\text{parsable samples}}{\#\text{total samples}} \tag{1}$$

- **APP Accuracy** refers to the proportion of function calls generated by the LLM where the selected app is the same as the ground-truth tool, which indicates the tool comprehension ability.

$$\text{APP\_acc} = \frac{\#\text{correct APP samples}}{\#\text{total samples}} \tag{2}$$

- **Function Name Accuracy** is the proportion of samples where the function name in an LLM-generated tool call is an exact match to the ground truth, which indicates the function selection ability.

$$\text{function\_name\_acc} = \frac{\#\text{correct function name samples}}{\#\text{total samples}} \tag{3}$$

- **Function Parameter Accuracy** is the proportion of samples where both the number and the names of the parameters in an LLM-generated tool call are an exact match to the ground truth.

$$\text{function\_param\_acc} = \frac{\#\text{correct function param samples}}{\#\text{total samples}} \tag{4}$$

- **Function Parameter-Value Accuracy** is the proportion of samples where the parameter names and their corresponding values in an LLM-generated tool call are both an exact match to the ground truth, which indicate the parameter personalization and temporal context personalization ability.

$$\text{function\_value\_acc} = \frac{\#\text{correct parameter value samples}}{\#\text{total samples}} \tag{5}$$

- **Function Temporal Context Parameter-Value Accuracy** is the proportion of samples where the **temporal context parameter** names and their corresponding values in an LLM-generated tool call are both an exact match to the ground truth, which indicate the temporal context personalization ability **only**. In the following formula, we use **TCP** to represent temporal context parameter.

$$\text{function\_TCP\_value\_acc} = \frac{\#\text{correct TCP value samples}}{\#\text{total samples containing TCP}} \tag{6}$$

- **Overall Accuracy** indicate the overall personalized function calling ability.

$$\text{overall\_acc} = \frac{\#\text{full correct samples}}{\#\text{total samples}} \tag{7}$$