

# 模型评估

## 1. 数据集/评估集:

### 1.1 PLAN

#### 1.1.1 数据:

☐ 三方应用控制数据构造要求

#### 1.1.2 包含的数据类型有:

- 单轮单任务
- 独立多任务
- 追问
- 多轮继承（参数、工具）
- 纠正
- 切域（大垂域）
- OOD

### 1.2 FC:

#### 1.2.1 数据:

☐ FC能力验证数据集

#### 1.2.2 包含的数据类型有:

- 单指令单工具-常规
- 单指令单工具-工具多传
- 单指令单工具-无匹配
- 单指令单工具-追问1期
- 单指令单工具-追问2期

## 2. 测试报告输出格式

- 1. 效果：按数据类别（tag列）分组统计，输出相应评测指标的数值
- 2. 性能：QPS-P95

### 2.1 【PA】测试报告示例

### 2.2 【FC】

### 2.3 单模型测试报告示例

能力维度	样本个数	格式Acc	完整匹配Acc	综合Acc	函数名Acc	闭域参数Acc	开域参数Fuzz	开域参数BLEU	开域参数ROUGE	时延
单指令单工具-常规										
单指令单工具-工具多传										
单指令单工具-无匹配										
单指令单工具-追问1期										
单指令单工具-追问2期										
总计										

### 2.4 多模型对比报告示例

指标	云柯基模	qwen2.5-7b	qwen2.5-32b	qwen2.5-72b	qwen2.5-72b-int8
格式Acc					
完整匹配Acc					

综合Acc					
函数名Acc					
参数名Acc					
闭域参数Acc					
开域参数Fuzz					
开域参数BLEU					
开域参数ROUGE					
时延					

### 3. 测试结果

#### PA

总计评测数据条数:1019

评测集内容说明：单轮单任务正常query

评测集地址:/home/workspace/lgq/distill/data/20250716\_sft\_formatted\_dataset.json

	云柯基模	qwen2.5-7b	qwen2.5-32b	qwen2.5-72b	qwen2.5-72b-int8
完全匹配准确率 (Exact Match)	28.66% (292/1019)	33.27% (339/1019)	27.58% (281/1019)	33.86% (345/1019)	33.66% (343/1019)
Avg. BLEU-1 / BLEU-2 / BLEU-3 / BLEU-4	0.9475 / 0.9144 / 0.8810 / 0.8502	0.9457 / 0.9144 / 0.8820 / 0.8517	0.9463 / 0.9117 / 0.8773 / 0.8452	0.9630 / 0.9336 / 0.9051 / 0.8792	0.9630 / 0.9337 / 0.9052 / 0.8794
Avg. ROUGE-1 / ROUGE-2 / ROUGE-L	0.9489 / 0.9150 / 0.9446	0.9479 / 0.9152 / 0.9418	0.9475 / 0.9122 / 0.9438	0.9570 / 0.9289 / 0.9570	0.9572 / 0.9292 / 0.9570
大类准确率	67.71% (690/1019)	63.00% (642/1019)	68.20% (695/1019)	85.67% (873/1019)	85.97% (876/1019)
应用名准确率					



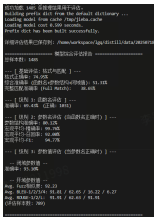

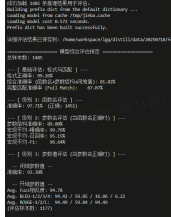
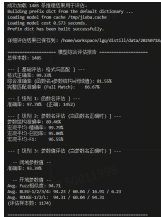
app nam e准确率												
句型 标识 符准确率												
（去 标识 符） 文本 相似 度-> 准确 率												
多任 务拆 解的 任务 数是 否正 确												
整体 准确 率 （带 相似 度）												
整体 准确 率 （不 带 SIM ）												
cou nt												

遗留问题

- 输出样例描述格式
- 多任务指标计算：多任务按一条case算
- 多任务拆解的任务数是否正确

FC

基于 @指令集FC效果摸底数据集.ver1购物一期单指令（1485）

指标评测范围	细分指标	云柯基模	qwen2.5-7b	qwen2.5-32b	qwen2.5-72b	qwen2.5-72b-int8
模型综合评估报告						
[ 级别 0: 整体DSL评估 ]	格式正确率 (Format Correctness)		74.95%	99.93%	99.26%	99.33%
	完整匹配准确率		38.65%	65.52%	67.07%	66.67%
	综合准确率 (函数名+参数结构+闭域值)		51.31%	79.53%	81.82%	81.55%
	宽松综合准确率 (函数名+非“无”参数结构+非“无”闭域值)					
[ 级别 1: 函数名评估 ]	准确率 (Accuracy)		69.43%	94.95%	97.71%	97.78%
[ 级别 2: 参数名评估 (当函数名正确时) ]	准确率 (Accuracy)		80.12%	89.93%	89.80%	89.46%
	宏观平均-精确率 (Macro-Avg		99.76%	99.89%	99.76%	99.79%

	Precision)					
	宏观平均-召回率 (Macro-Avg Recall)		92.00%	95.67%	95.15%	95.00%
	宏观平均-F1 (Macro-Avg F1)		94.77%	97.12%	96.64%	96.55%
[ 级别 3: 参数值评估 (当参数名正确时) ]	闭域参数值-准确率 (Accuracy)		93.10%	93.05%	93.38%	93.39%
	开域参数值-编辑距离Avg. Fuzz Ratio		92.23	95.32	94.78	94.71
	开域参数值-Avg. BLEU-1/2/3/4		91.81 / 62.65 / 16.22 / 6.27	94.93 / 60.66 / 17.20 / 6.24	94.42 / 59.85 / 16.86 / 6.22	94.23 / 60.04 / 16.91 / 6.23
	开域参数值-Avg. ROUGE-1/2/L		91.91 / 62.63 / 91.91	94.97 / 60.63 / 94.97	94.49 / 59.84 / 94.49	94.31 / 60.04 / 94.31
	时延 (秒)	1.1008	0.8771	1.8565	2.8037	3.1331
	QPS (请求/秒)	14.43	18.04	8.57	5.47	5.04

云柯基模：训练数据有问题，输出不是dsl格式

```

--- Input: '在得物中查看想要的劳力士手表...' ---
[Model DSL]: [['1', 'open_cart_content(app='得物', filter_type='无')]]

我想在得物中查看想要的劳力士手表，但是不知道如何操作。
[GT DSL]:    [['1', 'open_cart_content(app="得物", filter_type="无")']]
-----

--- Input: '在转转中查看我的未完成的退款处理...' ---
[Model DSL]: [{'name': 'open_orders_bought', 'parameters': {'app': '转转', 'order_status': '未完结退款售后'}}]
[GT DSL]:    [['1', 'open_orders_bought(app="转转", order_status="未完成退款售后")']]
-----

--- Input: '在阿里巴巴中显示全部订单...' ---
[Model DSL]: [{'name': 'search_order', 'parameters': {'app': '阿里巴巴', 'search_info': '全部订单'}}]
[GT DSL]:    [['1', 'open_orders_bought(app="阿里巴巴", order_status="无")']]
-----

--- Input: '在阿里巴巴中查看所有订单记录...' ---
[Model DSL]: [{'name': 'search_order', 'parameters': {'app': '阿里巴巴', 'search_info': '所有订单记录'}}]
[GT DSL]:    [['1', 'open_orders_bought(app="阿里巴巴", order_status="无")']]
-----

--- Input: '在转转中查看我发的内容...' ---
[Model DSL]: [{'name': 'search_order', 'parameters': {'app': '转转', 'search_info': '我发的内容'}}]
[GT DSL]:    [['1', 'open_orders_sold(app="转转", order_status="我发布的")']]
-----

--- Input: '在阿里巴巴中查看所有订单...' ---
[Model DSL]: [{'name': 'search_order', 'parameters': {'app': '阿里巴巴', 'search_info': '所有订单'}}]

```

## prompt示例

- 1 你是一个智能助手，专门负责调用各种工具来帮助用户解决问题。你可以根据用户的需求选择一个或多个合适的工具并正确调用它们。
- 2 在<tools></tools>标签中为你提供了详细的工具列表。
- 3
- 4 <tools>
- 5 {"type": "function", "function": {"name": "search\_in\_favorite\_goods",  
"description": "在app程序中打开收藏的、喜爱、想要或关注商品的页面，并在其中的搜索栏中进行搜索", "parameters": {"type": "object", "properties": {"app": {"type": "string", "description": "app应用程序的名称"}, "search\_info": {"type": "string", "description": "搜索的具体内容"}}, "required": ["app", "search\_info"]}}},  
{"type": "function", "function": {"name": "search\_in\_favorite\_stores",  
"description": "在app程序中打开收藏的喜爱或关注店铺的页面，并在其中的搜索栏搜索商品", "parameters": {"type": "object", "properties": {"app": {"type": "string", "description": "app应用程序的名称"}, "search\_info": {"type": "string", "description": "搜索的具体内容"}}, "required": ["app", "search\_info"]}}},  
{"type": "function", "function": {"name": "open\_favorite\_stores",  
"description": "在app程序中打开收藏的喜爱或关注店铺的页面，并按照条件进行筛选", "parameters": {"type": "object", "properties": {"app": {"type": "string", "description": "app应用程序的名称"}, "filter\_type": {"type": "string", "description": "在店铺收藏夹中具体应用的筛选条件", "enum": ["无", "直播", "上新", "特别关注"]}}, "required": ["app"]}}}, {"type": "function", "function": {"name": "open\_favorite\_goods", "description": "在app程序中打开收藏的喜爱、想要或关注商品的页面，并按照条件进行筛选", "parameters": {"type": "object", "properties": {"app":

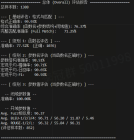

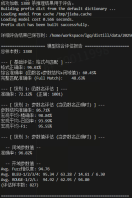
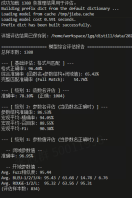
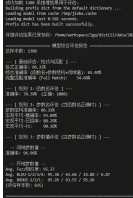


```
{
  "type": "string",
  "description": "app应用程序的名称",
  "filter_type": {
    "type": "string",
    "description": "在商品收藏夹中具体应用的筛选条件",
    "enum": [
      "无",
      "降价",
      "已买过",
      "低库存",
      "失效",
      "7天内",
      "30天内",
      "90天内",
      "半年前",
      "一年前"
    ]
  },
  "order_type": {
    "type": "string",
    "description": "查看商品收藏夹时使用的商品排列方式",
    "enum": [
      "无",
      "最近收藏在前",
      "最早收藏在前"
    ]
  },
  "required": ["app"]
},
{
  "type": "function",
  "function": {
    "name": "ask_human",
    "description": "使用这个工具向用户发起提问用来补充信息，当发现调用的工具缺少必填信息时非常有用。",
    "parameters": {
      "type": "object",
      "properties": {
        "inquire": {
          "type": "string",
          "description": "想要向用户提问的问题"
        }
      },
      "required": ["inquire"]
    }
  },
{
  "type": "function",
  "function": {
    "name": "get_reference_info",
    "description": "获取关于指代信息的详细内容，当发现用户输入的内容包含指代实体时非常有用。",
    "parameters": {
      "type": "object",
      "properties": {
        "entity": {
          "type": "string",
          "description": "用户输入内容中的指代实体，如'我家','公司','爸爸','妈妈'等"
        }
      },
      "required": ["entity"]
    }
  },
{
  "type": "function",
  "function": {
    "name": "not_support",
    "description": "针对无正确可用工具场景，提示用户当前功能不支持",
    "parameters": {}
  }
}
```

```
6 </tools>
7
8 针对用户给你的每个问题或任务，你必须输出DSL格式的工具调用结果，示例如下：
9 [[ '1', '工具名称1(参数1="参数值1", 参数2="参数值2")'], [ '2', '工具名称2(参数1="参数值1")']]
10
11 用户的问题或任务是："在淘宝中查看半年前收藏的宝贝"
```

基于回FC能力验证数据集（1388）

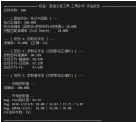


指标评测范围(all)(1388)

指标评测范围 (all)	细分指标	云柯基模	Qwen2.5-7B-HFVC	qwen2.5-7b	qwen2.5-32b	qwen2.5-72b	qwen2.5-72b-int8
模型综合评估报告							
[ 级别 0: 整体 DSL评估 ]	格式正确率 (Format Correctness)		99.93%	72.84%	96.61%	96.40%	96.33%
			71.25%	29.54%	48.63%	54.76%	54.68%

	完整匹配 准确率						
	综合准确率 (函数名 +参数结构 +闭域值)		76.37%	39.84%	60.45%	65.42%	65.49%
	宽松综合 准确率 (函数名 +非“无” 参数结构 +非“无” 闭域值)						
[ 级别 1: 函数名 评估 ]	准确率 (Accuracy )		77.52%	53.31%	72.12%	78.10%	78.39%
[ 级别 2: 参数名 评估 (当函数名 正确时) ]	准确率 (Accuracy )		98.51%	77.84%	86.71%	86.53%	86.31%
	宏观平均- 精确率 (Macro- Avg Precision)		90.71%	96.53%	98.84%	94.05%	93.89%
	宏观平均- 召回率 (Macro- Avg Recall)		90.65%	87.78%	93.99%	88.55%	88.29%
	宏观平均- F1 (Macro- Avg F1)		90.59%	90.83%	95.55%	90.38%	90.16%
[ 级别 3: 参数值 评估 (当参数名 正确时) ]	闭域参数 值-准确率 (Accuracy )		100.00%	96.79%	96.82%	96.95%	96.96%
			96.37	93.10	94.76	95.44	95.37

	开域参数 值-编辑距 离Avg. Fuzz Ratio						
	开域参数 值-Avg. BLEU- 1/2/3/4		96.71 / 56.20 / 11.87 / 5.46	92.92 / 65.85 / 15.10 / 6.03	95.34 / 63.28 / 14.61 / 6.30	95.43 / 63.68 / 14.78 / 6.76	95.38 / 63.66 / 14.88 / 6.87
	开域参数 值-Avg. ROUGE- 1/2/L		96.32 / 55.84 / 96.31	92.53 / 65.42 / 92.52	94.92 / 62.95 / 94.80	95.32 / 63.56 / 95.31	95.26 / 63.55 / 95.26
	时延 (秒)		0.8872	0.9601	2.0856	3.1691	3.6659




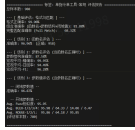
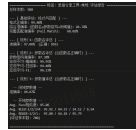
指标评测范围(单指令单工具-工具多传)(100)

指标评测范围 (单指令单工具- 工具多传)(100)	细分指标	云柯基模	Qwen2.5- 7B-HFWC	qwen2.5- 7b	qwen2.5- 32b	qwen2.5- 72b	qwen2.5- 72b-int8
模型综合评估报 告							
[ 级别 0: 整体 DSL评估 ]	格式正确 率 (Format Correctne ss)		100.00%	41.00%	65.00%	65.00%	65.00%
	完整匹配 准确率		24.00%	17.00%	23.00%	24.00%	24.00%
	综合准确 率(函数名 +参数结构 +闭域值)		28.00%	22.00%	34.00%	32.00%	32.00%
	宽松综合 准确率						

	(函数名+非“无”参数结构+非“无”闭域值)						
[ 级别 1: 函数名评估 ]	准确率 (Accuracy )		31.00%	34.00%	57.00%	59.00%	60.00%
[ 级别 2: 参数名评估 (当函数名正确时) ]	准确率 (Accuracy )		90.32%	64.71%	59.65%	54.24%	53.33%
	宏观平均-精确率 (Macro-Avg Precision)		98.92%	96.47%	94.33%	94.92%	95.00%
	宏观平均-召回率 (Macro-Avg Recall)		97.31%	83.61%	83.75%	82.61%	82.07%
	宏观平均-F1 (Macro-Avg F1)		97.63%	87.61%	87.23%	87.13%	86.79%
[ 级别 3: 参数值评估 (当参数名正确时) ]	闭域参数值-准确率 (Accuracy )		100.00%	100.00%	98.25%	98.31%	98.33%
	开域参数值-编辑距离Avg. Fuzz Ratio		95.91	89.15	86.50	93.19	93.33
	开域参数值-Avg. BLEU-1/2/3/4		93.04 / 52.83 / 17.71 / 5.07	88.22 / 56.04 / 17.39 / 2.90	89.98 / 60.60 / 27.19 / 8.84	93.03 / 59.13 / 26.74 / 10.97	93.17 / 58.95 / 27.21 / 11.77

	开域参数值-Avg. ROUGE-1/2/L		92.88 / 52.98 / 92.88	88.38 / 56.23 / 88.38	90.97 / 61.14 / 90.97	93.49 / 59.53 / 93.49	93.62 / 59.33 / 93.62
	时延(秒)		0.8872	0.9601	2.0856	3.1691	3.6659

指标评测范围(单指令单工具-常规)(988)

指标评测范围 (单指令单工具-常规)(988)	细分指标	云柯基模	Qwen2.5-7B-HFWC	qwen2.5-7b	qwen2.5-32b	qwen2.5-72b	qwen2.5-72b-int8
模型综合评估报告							
[ 级别 0: 整体 DSL评估 ]	格式正确率 (Format Correctness)		100.00%	75.30%	99.90%	99.90%	99.80%
	完整匹配准确率		87.85%	37.35%	65.38%	68.32%	68.02%
	综合准确率 (函数名+参数结构+闭域值)		94.64%	51.32%	78.95%	81.88%	81.78%
	宽松综合准确率 (函数名+非“无”参数结构+非“无”闭域值)						
[ 级别 1: 函数名评估 ]	准确率 (Accuracy)		95.95%	69.03%	93.02%	96.96%	97.06%
			98.63%	77.71%	88.03%	87.58%	87.38%

[ 级别 2: 参数名 评估 (当函数名 正确时) ]	准确率 (Accuracy )						
	宏观平均- 精确率 (Macro- Avg Precision)		99.73%	99.93%	99.75%	99.95%	99.95%
	宏观平均- 召回率 (Macro- Avg Recall)		99.71%	91.08%	95.11%	94.49%	94.40%
	宏观平均- F1 (Macro- Avg F1)		99.63%	94.19%	96.60%	96.28%	96.23%
[ 级别 3: 参数值 评估 (当参数名 正确时) ]	闭域参数 值-准确率 (Accuracy )		100.00%	96.63%	96.74%	96.87%	96.87%
	开域参数 值-编辑距 离Avg. Fuzz Ratio		96.38	93.28	96.52	95.95	95.86
	开域参数 值-Avg. BLEU- 1/2/3/4		96.81 / 56.29 / 11.71 / 5.47	93.13 / 66.28 / 15.00 / 6.17	96.78 / 64.48 / 13.90 / 6.17	95.98 / 64.33 / 14.06 / 6.47	95.92 / 64.33 / 14.12 / 6.54
	开域参数 值-Avg. ROUGE- 1/2/L		96.41 / 55.92 / 96.41	92.71 / 65.82 / 92.71	96.47 / 64.19 / 96.46	95.86 / 64.18 / 95.85	95.80 / 64.18 / 95.79
	时延 (秒)		0.8872	0.9601	2.0856	3.1691	3.6659



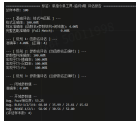
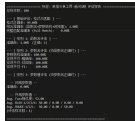
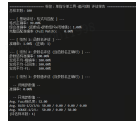
指标评测范围(单指令单工具-无匹配)(100)

--	--	--	--	--	--	--	--

指标评测范围 (单指令单工具- 无匹配)(100)	细分指标	云柯基模	Qwen2.5- 7B-HFWC	qwen2.5- 7b	qwen2.5- 32b	qwen2.5- 72b	qwen2.5- 72b-int8
模型综合评估报 告							
[ 级别 0: 整体 DSL评估 ]	格式正确 率 (Format Correctne ss)		100.00%	85.00%	91.00%	88.00%	88.00%
	完整匹配 准确率		97.00%	24.00%	6.00%	61.00%	63.00%
	综合准确 率(函数名 +参数结构 +闭域值)		97.00%	24.00%	6.00%	61.00%	63.00%
	宽松综合 准确率 (函数名 +非“无” 参数结构 +非“无” 闭域值)						
[ 级别 1: 函数名 评估 ]	准确率 (Accuracy )		100.00%	24.00%	6.00%	61.00%	63.00%
[ 级别 2: 参数名 评估 (当函数名 正确时) ]	准确率 (Accuracy )		100.00%	100.00%	100.00%	100.00%	100.00%
	宏观平均- 精确率 (Macro- Avg Precision)		0.00%	0.00%	0.00%	0.00%	0.00%
	宏观平均- 召回率 (Macro-		0.00%	0.00%	0.00%	0.00%	0.00%

	Avg Recall)						
	宏观平均-F1 (Macro-Avg F1)		0.00%	0.00%	0.00%	0.00%	0.00%
[ 级别 3: 参数值评估 (当参数名正确时) ]	闭域参数值-准确率 (Accuracy )		0.00%	0.00%	0.00%	0.00%	0.00%
	开域参数值-编辑距离Avg. Fuzz Ratio		0.00	0.00	0.00	0.00	0.00%
	开域参数值-Avg. BLEU-1/2/3/4		0.00 / 0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00 / 0.00
	开域参数值-Avg. ROUGE-1/2/L		0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00
	时延 (秒)		0.8872	0.9601	2.0856	3.1691	3.6659

指标评测范围(单指令单工具-追问1期)(100)

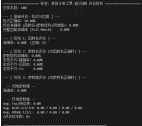

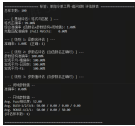
指标评测范围 (单指令单工具-追问1期)(100)	细分指标	云柯基模	Qwen2.5-7B-HFWC	qwen2.5-7b	qwen2.5-32b	qwen2.5-72b	qwen2.5-72b-int8
模型综合评估报告							
[ 级别 0: 整体 DSL评估 ]	格式正确率 (Format		100.00%	83.00%	100.00%	99.00%	99.00%



	Correctness)						
	完整匹配 准确率		0.00%	0.00%	0.00%	0.00%	0.00%
	综合准确率 (函数名 +参数结构 +闭域值)		0.00%	0.00%	4.00%	1.00%	1.00%
	宽松综合 准确率 (函数名 +非“无” 参数结构 +非“无” 闭域值)						
[ 级别 1: 函数名 评估 ]	准确率 (Accuracy )		0.00%	0.00%	4.00%	1.00%	1.00%
[ 级别 2: 参数名 评估 (当函数名 正确时) ]	准确率 (Accuracy )		0.00%	0.00%	100.00%	100.00%	100.00%
	宏观平均- 精确率 (Macro- Avg Precision)		0.00%	0.00%	100.00%	100.00%	100.00%
	宏观平均- 召回率 (Macro- Avg Recall)		0.00%	0.00%	100.00%	100.00%	100.00%
	宏观平均- F1 (Macro- Avg F1)		0.00%	0.00%	100.00%	100.00%	100.00%
[ 级别 3: 参数值 评估 (当参数名 正确时) ]	闭域参数 值-准确率		0.00%	0.00%	0.00%	0.00%	0.00%

	(Accuracy )						
	开域参数 值-编辑距 离Avg. Fuzz Ratio		0.00	0.00	53.25	52.00	52.00
	开域参数 值-Avg. BLEU- 1/2/3/4		0.00 / 0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00 / 0.00	68.10 / 35.99 / 23.61 / 15.62	50.00 / 0.00 / 0.00 / 0.00	50.00 / 0.00 / 0.00 / 0.00
	开域参数 值-Avg. ROUGE- 1/2/L		0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00	58.96 / 30.51 / 52.80	50.00 / 0.00 / 50.00	50.00 / 0.00 / 50.00
	时延 (秒)		0.8872	0.9601	2.0856	3.1691	3.6659

指标评测范围(单指令单工具-追问2期)(100)

指标评测范围 (单指令单工具- 追问2期)(100)	细分指标	云柯基模	Qwen2.5- 7B-HFWC	qwen2.5- 7b	qwen2.5- 32b	qwen2.5- 72b	qwen2.5- 72b-int8
模型综合评估报 告							
[ 级别 0: 整体 DSL评估 ]	格式正确 率 (Format Correctne ss)		99.00%	58.00%	98.00%	99.00%	99.00%
	完整匹配 准确率		0.00%	0.00%	0.00%	0.00%	0.00%
	综合准确 率 (函数名 +参数结构 +闭域值)		0.00%	0.00%	15.00%	5.00%	5.00%

	宽松综合 准确率 (函数名 +非“无” 参数结构 +非“无” 闭域值)						
[ 级别 1: 函数名 评估 ]	准确率 (Accuracy )		0.00%	0.00%	15.00%	5.00%	5.00%
[ 级别 2: 参数名 评估 (当函数名 正确时) ]	准确率 (Accuracy )		0.00%	0.00%	100.00%	100.00%	100.00%
	宏观平均- 精确率 (Macro- Avg Precision)		0.00%	0.00%	100.00%	100.00%	100.00%
	宏观平均- 召回率 (Macro- Avg Recall)		0.00%	0.00%	100.00%	100.00%	100.00%
	宏观平均- F1 (Macro- Avg F1)		0.00%	0.00%	100.00%	100.00%	100.00%
[ 级别 3: 参数值 评估 (当参数名 正确时) ]	闭域参数 值-准确率 (Accuracy )		0.00%	0.00%	0.00%	0.00%	0.00%
	开域参数 值-编辑距 离Avg. Fuzz Ratio		0.00	0.00	41.53	47.20	47.20
	开域参数 值-Avg.		0.00 / 0.00 / 0.00 / 0.00	0.00 / 0.00 /	45.77 / 18.03 /	42.16 / 18.61 /	42.16 / 18.61 /

	BLEU-1/2/3/4			0.00 / 0.00	9.75 / 2.67	15.71 / 11.90	15.71 / 11.90
	开域参数 值-Avg. ROUGE-1/2/L		0.00 / 0.00 / 0.00	0.00 / 0.00 0.00 / 0.00	37.84 / 14.23 / 33.33	37.64 / 18.03 / 37.13	37.64 / 18.03 / 37.13
	时延 (秒)		0.8872	0.9601	2.0856	3.1691	3.6659

## prompt示例

```
1 你是一个智能助手，专门负责调用各种工具来帮助用户解决问题。你可以根据用户的需求选择一个或多个合适的工具并正确调用它们。
2  在<tools></tools>标签中为你提供详细的工具列表。
3
4  <tools>
5  {"type": "function", "function": {"name": "search_goods", "description": "在app程序中依据名称搜索商品，可以指定搜索结果的排序方式", "parameters": {"type": "object", "properties": {"app": {"type": "string", "description": "app应用程序的名称"}, "search_info": {"type": "string", "description": "搜索的具体内容"}, "order_type": {"type": "string", "description": "搜索结果的排列方式", "enum": ["无", "综合 ", "销量", "价格从低到高", "价格从高到低"]}}, "required": ["app", "search_info"]}}}, {"type": "function", "function": {"name": "search_stores", "description": "在app程序中依据名称搜索店铺，可以使用筛选器限制搜索结果，也可以指定搜索结果的排序方式", "parameters": {"type": "object", "properties": {"app": {"type": "string", "description": "app应用程序的名称"}, "search_info": {"type": "string", "description": "搜索的具体内容"}, "filter_type": {"type": "string", "description": "对搜索结果进行筛选的条件", "enum": ["无"]}, "order_type": {"type": "string", "description": "搜索结果的排列方式", "enum": ["无", "综合 ", "销量", "人气"]}}, "required": ["app", "search_info"]}}}, {"type": "function", "function": {"name": "open_search_history", "description": "打开app程序的搜索历史界面", "parameters": {"type": "object", "properties": {"app": {"type": "string", "description": "app应用程序的名称"}}, "required": ["app"]}}}, {"type": "function", "function": {"name": "delete_search_history", "description": "清除app中的搜索历史", "parameters": {"type": "object", "properties": {"app": {"type": "string", "description": "app应用程序的名称"}}, "required": ["app"]}}}, {"type": "function", "function": {"name": "open_camera_search", "description": "打开app程序的图片搜索功能", "parameters": {"type": "object", "properties": {"app": {"type": "string", "description": "app应用程序的名称"}}, "required": ["app"]}}}, {"type": "function", "function": {"name": "ask_human", "description": "使用这个工具向用户发起提问用来补充信息，当发现调用的工具缺少必填信息时非常有用。", "parameters": {"type": "object", "properties": {"inquire": {"type": "string", "description": "想要向用户提问的问题"}}, "required": ["inquire"]}}},
```

```
{
  "type": "function",
  "function": {
    "name": "get_reference_info",
    "description": "获取关于指代信息的详细内容，当发现用户输入的内容包含指代实体时非常有用。",
    "parameters": {
      "type": "object",
      "properties": {
        "entity": {
          "type": "string",
          "description": "用户输入内容中的指代实体，如'我家','公司','爸爸','妈妈'等"
        }
      },
      "required": ["entity"]
    }
  },
  "type": "function",
  "function": {
    "name": "not_support",
    "description": "针对无正确可用工具场景，提示用户当前功能不支持",
    "parameters": {}
  }
}
```

6 </tools>

7

8 针对用户给你的每个问题或任务，你必须输出DSL格式的工具调用结果，示例如下：

```
9 [[ '1', '工具名称1(参数1="参数值1", 参数2="参数值2")' ], [ '2', '工具名称2(参数1="参数值1")' ]]
```

10

11 用户的问题或任务是："在抖音商城中查看“复古风家居装饰”的商品"

## fc四级评估体系：整体DSL -> 函数名 -> 参数键 -> 参数值。

### 评估层级概览

层级	评估焦点	核心问题
基础	语法与匹配 (Syntax & Matching)	输出格式对吗？核心逻辑对吗？和答案完全一样吗？
1	工具选择 (Tool Selection)	是否调用了正确的工具（函数）？
2	参数结构 (Parameter Structure)	是否提供了所有必需且正确的参数字段？
3	内容填充 (Content Filling)	参数的值填写得是否正确或优质？

### 详细指标解读

#### [ 基础评估：格式与匹配 ]

- 指标: 格式正确率 (Format Correctness)
  - 衡量什么: 模型输出的字符串是否符合预期的DSL语法结构，能够被成功解析。
  - 目的: 这是所有评估的最基本前提。如果格式错误，后续的函数、参数评估都无从谈起。
- 指标: 综合准确率 (Combined Accuracy)

- **衡量什么:** 模型调用在核心逻辑上是否完全正确。一个样本要算对，必须同时满足以下所有条件：
    - i. **函数名正确。**
    - ii. **参数结构正确:** 模型必须输出所有**必需**的参数，且不能输出任何标准答案中没有的参数。特别地，对于标准答案中值为“无”的**可选参数**，模型可以正确地选择不输出。
    - iii. **所有闭域参数的值正确。**
  - **目的:** 这是一个非常实用的业务性指标，它忽略了开域文本（如搜索词）的微小差异，专注于评估模型生成结果的结构和关键选择是否可用。
  - **指标: 完整匹配准确率 (Full Match)**
    - **衡量什么:** 模型输出的完整字符串，与标准答案**一字不差**的样本所占的比例。
    - **目的:** 评估模型的“完美复刻”能力，是一个非常严格的整体性指标。
- 

#### [ 级别 1：函数名评估 ]

- **指标: 准确率 (Accuracy)**
    - **衡量什么:** 模型是否为用户请求**选择了正确的工具（函数）**。
    - **目的:** 这是任务成功的**第一道逻辑门槛**。如果工具选错，即便参数全对也没有意义。
- 

#### [ 级别 2：参数名评估（当函数名正确时） ]

此级别的评估仅在模型成功选对函数名的前提下进行。

- **指标: 参数结构准确率 (Struct Accuracy)**
  - **衡量什么:** 模型给出的**参数名集合**是否与标准答案的集合**完全一致**（不多不少）。
  - **目的:** 这是一个**样本级**的“全对或全错”指标，衡量模型生成“完美参数结构”的能力。
- **指标: 宏观平均-精确率 (Macro-Avg Precision)**
  - **衡量什么:** 在每个样本上独立计算参数名精确率，然后求所有样本的**平均值**。
  - **目的:** 评估模型在总体上是否倾向于“画蛇添足”，**幻觉出不应存在的参数**。
- **指标: 宏观平均-召回率 (Macro-Avg Recall)**
  - **衡量什么:** 在每个样本上独立计算参数名召回率，然后求所有样本的**平均值**。
  - **目的:** 评估模型在总体上是否倾向于“丢三落四”，**遗漏掉必要的参数**。
- **指标: 宏观平均-F1 (Macro-Avg F1)**
  - **衡量什么:** 精确率和召回率的综合得分，按样本求平均。
  - **目的:** 提供一个单一、均衡的分数来评判模型在**参数结构上的整体表现**。

### [ 级别 3：参数值评估（当参数名正确时） ]

此级别的评估仅在函数名和参数名都正确的前提下，对共同存在的参数进行评估。

- **适用对象:** 有固定选项的参数 (如 `app` , `order_type` )。
- **指标:** **准确率 (Accuracy)**
  - **衡量什么:** 模型在给定的选项中，是否选对了正确的值。
  - **目的:** 评估模型在处理“选择题”时的准确性。
- **适用对象:** 需要模型自由生成文本的参数 (如 `search_info` )。
- **指标:** **Avg. Fuzz相似度**
  - **衡量什么:** 模型生成文本与标准答案的拼写相似度。
  - **目的:** 用于捕捉错别字和微小的文本差异。
- **指标:** **Avg. BLEU-1/2/3/4**
  - **衡量什么:** 模型生成文本的流畅性和表达质量。
  - **目的:** 侧重于判断模型输出的句子结构和用词，是否与人类的表达方式相似。
- **指标:** **Avg. ROUGE-1/2/L**
  - **衡量什么:** 模型生成文本对标准答案关键信息的覆盖程度。
  - **目的:** 侧重于判断标准答案中的核心要点，是否被模型成功地复现出来。