模型评估

1. 数据集/评估集:

1.1 PLAN

1.1.1 数据:

回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 ROUG E	时延
	7	IE 50011998		東西群 5001	1998		李冠群 50011998			50011998
单指令单工 具-常规			李冠曆 50011998			SEE 2001788				
单指令单工 具-工具多传	李	E# 50011998		李冠舞 5001	(Aag		李冠群 50011998		李冠	50011998
兵-工兵夕19	111998		李冠群 50011998		型"	西 牌 50011998		李冠群 5001	1998	
单指令单工 具-无匹配										50011998
李冠群与	21998		李冠群 50011993		李	E 18 50031098		李冠群 5003	098	
单指令单工 具-追问1期										50011 ⁹⁹⁸
单指令单工	117888		李阳群 50011998		9	E 50011998		李冠群 5001	998	
具-追问2期										¥50011998
总计	11998		- con11998			w c0011998		c001	998	

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					EN 20011998
完整匹配Acc					#RB 50011998

综合Acc			李冠群 50011998				
函数名Acc		李冠群 50011998		李冠牌 50011998		李祖時 50011998	
参数名Acc	李冠群 50011998		李冠群 50011998		李祖斯 50011998		李冠群 50011998
闭域参数Acc		李冠群 50011998		李冠群 50011998		李节群 50011998	
开域参数Fuzz	李冠碑 50011998		李冠群 50077938		李冠群 50011998		李冠群 50011995
开域参数BLEU		泰冠群 50011998		率招牌 50011998		李科群 50011998	
开域参数ROUGE	李冠科 50011908		李冠郡 50011308		#18 FF 5001.1998		李冠群 50011508
<mark>时延</mark>		李冠群 50011950		李冠牌 50011998		李节群 50011358	

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		# FETTH 5001		1888 S0011958	李冠群 50011998

**************************************	92.54% (943/1019)	91.46% (932/1019)	96.47% (983/1019)	97.64% (995/1019)	97.55% (994/1019)
时延(s)	1.01	0.89	1.8271	2.6388	3.0074
详细指标					18

评估细则

输出样例描述

单任务-正常:

<Plan>#大类#在\$app\$动作<mark>%plan成功%</mark></Plan><Finish>

单任务-追问:

<Plan>#大类#在\$app\$动作<mark>%reply%</mark></Plan><Finish>

单任务-拒识:

<Plan><mark>#未知#</mark>在\$app\$动作<mark>%暂不支持该操作%</mark></Plan><Finish>

多任务-正常:

<Plan>#大类#在\$app\$动作<mark>%plan成功%</mark></Plan>
<Plan>#大类#在\$app\$动作<mark>%plan成功%</mark></Plan><Finish>

多任务-追问:

<Plan>#大类#在\$app\$动作<mark>%plan成功%</mark></Plan>
<Plan>#大类#在\$app\$动作<mark>%reply%</mark></Plan><Finish>

多任务-拒识:

<Plan>#大类#在\$app\$动作<mark>%plan成功%</mark></Plan>
<Plan><mark>#未知#</mark>在\$app\$动作<mark>%暂不支持该操作%</mark></Plan><Finish>

指标计算描述

单轮-	单轮-	单轮-	单轮-	单轮-	单轮-	单轮	多轮-	多轮-	多轮-	多轮-	多轮-	多轮-	43
单任	单任	单任	多任	多任	多任	整体	单任	单任	单任	多任	多任	多任	杢
务-正	务-追	务-拒	务-正	务-追	务-拒		务-正	务-追	务-拒	务-正	务-追	务-拒	
常	问	识	常	问	识		常	问	识	常	问	识	
李冠昂	1998 1998		李节	1 5001 1998		1	FOR 50011 098			李冠群 50011998			李节
		李冠群 5001199											
- 1	50017338			1# 50011998			·福斯 50011998			李冠群 50011998			45
		李冠群 5001199											
	单任 务-正	单任 单任 务-正 务-追	单任 单任 单任 务-正 务-追 务-拒	单任 单任 单任 多任 务-正 务-追 务-拒 务-正	单任 单任 单任 多任 多任 务-正 务-追 务-拒 务-正 务-追	单任 单任 单任 多任 多任 多任 务-正 务-追 务-拒 务-正 务-追 务-拒	单任 单任 单任 多任 多任 多任 整体 务-正 务-追 务-拒 务-正 务-追 务-拒	单任 单任 单任 多任 多任 多任 整体 单任 务-正 务-追 务-拒 务-正 务-追 务-拒 务-正	单任 单任 单任 多任 多任 多任 整体 单任 单任 务-正 务-追 务-追 务-追 务-拒 务-正 务-追	单任 单任 单任 多任 多任 多任 整体 单任 单任 单任 务-正 务-追 务-正 务-追 务-拒 务-正 务-正 务-拒	单任 单任 单任 多任 多任 多任 整体 单任 多任 务-正 务-追 务-正 务-追 务-正 务-正 务-追 务-正 务-正	单任 单任 单任 多任 多任 整体 单任 单任 单任 单任 多任 多任 务-正 务-追 务-追 务-追 务-拒 务-正 务-追 务-拒 务-追 务-追	单任 单任 单任 多任 多任 多任 整体 单任 单任 单任 多任 多任 多任 务-正 务-追 务-拒 务-拒 务-正 务-正

арр			李冠郡 50011998		李冠群 50013198		李冠群 50011998		李冠昭 20011998	
nam e准				李冠 8 50011998		李冠期 50011998		李冠群 50011998		李
确率			李冠群 50011998		李冠群 50011893		實質期 50011998		李冠斯 5)011998	
句型 标识				李冠耳 50011998		李昭朝 50011998		李冠群 50011998		李
符准 确率			李元章 2001199	4 50011998	李岩縣 50011月95	See11002	#18# 50 (1990	80011998	李昭和 50011999	
(去 标识			李冠斯 50011998	910	李冠科 50011月98		李冠郡 500,11998	39.00-00	李冠郡 50011998	39
符) 文本 相似			- muy99	李昭 身 50011998	- may 198	李冠群 50011998	- mg1393	李冠群 50011998	m11988	李
度-> 准确			李冠林 35000	李冠 [450011998	李冠郡 2000年	湖冠部 50011998	李冠群立山	李冠群 50011998	#1ED	77.7
率			李冠隊 50011998		第四届 50011898		李冠牌 50011398		李冠群 30011998	
多任 务拆				李阳 # 50011998		李冠群 50011998		李冠群 50011998		李
解的 任务 数是			李冠群 50011998	李冠章 50011998	李冠朝 50011798	#EE 50011998	李光章 500 11998	李冠群 50011998	#FEM 50011998	
否正确			李冠群 50011998		李冠器 50011/98		griffs 50ct 1998		李冠斯 59011998	
整体准确	李冠郎	50011998	20013998	李昭 年 50011998	-001V98	a filth 50011998	1993	李冠群 50011998	- 01/01/1988	7
率(带			東記師 300	李冠耳50011998	李冠目 300	新冠期 50011998	李冠籍	李冠暦 50011998	#1210 P	李
相似 度)			李冠群 50011998		李冠群 50011 198		泰岩區 500 11998		李冠群 53011998	
整体准确	李冠群	60011998	李冠師 50011998	#18 N 50021999	李冠郡 50011 798	第四 50011998	李昭朝 500,1398	#18# 50011998	李阳原 5011998	李
率 (不 带				李阳 每 50011998		李昭朝 50011998		李冠群 50011998		41
SIM			泰冠群 50011998	Rans	978EN 50011/98	Rong	李冠斯 5001.1998	, nas	#HH 50011988	
cou	東田野	60011998	李冠群 50011998	東昭 年 50011999	\$188 5001 Jp8	河南市50011999	李冠斯 50(11998	李冠群 50011990	#78M 501198	草
cou nt		-0111995	李冠群 50011398	- CON 1990	#1810 S0011/98		李昭朝 50011998		978EBI 57011998	

- 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

基于回指令集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			Action and deposits also an actionary country and action and action and action and action act	STORE AND PROPERTY OF A PARTY OF THE PARTY O	STATE OF SERVICE PARTY AND ADMINISTRATION OF THE PARTY ADMINISTRATION OF THE PARTY AND ADMINISTRATION OF THE PARTY AND ADMINISTRATION OF THE PARTY ADM	and the depotent and an artist and artist artist and artist
			### 175 500	THE PROPERTY IN THE PROPERTY I	###SEPTRE - 1885 - 188	**************************************
[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%
# 50011998	综合准确率 (函数名+参数结构+闭域值)	₉ 50011898	51.31%	79.53%	81.82%	81.55%
	宽松综合准确率 (函数名 +非"无"参数 结构		#182.00 - 0.00 L1 1918	\$1811 200774. 0 20077488		
	+非"无"闭域值)		李冠	, 50011988 		
[级别 1: 函数名评 估]	准确率 (Accuracy)	g 50011998	69.43%	94.95%	97.71%	97.78%
			被阻	50011998		
[级别 2:参数名评估(当函数名正确	准确率 (Accuracy)		80.12%	89.93%	89.80%	89.46%
时)]	宏观平均-精确 率 (Macro-Avg		99.76%	99.89%	99.76%	99.79%

	Precision)		1772 M 50011998	李冠群 50011		
	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%
	闭域参数值-准 确率 (Accuracy)		93.10%	93.05%	93.38%	93.39%
[级别 3:参数值评估(当参数名正确	开域参数值-编 辑距离Avg. Fuzz Ratio	g 50011998	92.23	95.32	94.78	94.71
时)]	开域参数值- Avg. BLEU- 1/2/3/4	g 50011998	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	N 20011998	91.91 / 62.63 / 91.91	94.97 / 60.63 / 94.97	94.49 / 59.84 / 94.49	94.31 / 60.04 / 94.31
AMM 20077985	时延(秒)	1.1008	0.8771	1.8565	2.8037	3.1331
\$18.00 Takes	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。
           [['1', 'open_cart_content(app="得物", filter_type="无")']]
[GT DSL]:
--- 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'}}]
           [['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'}}]
           [['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'}}]
           [['1', 'open_orders_bought(app="阿里巴巴", order_status="无")']]
[GT DSL]:
-- 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'}}]
           [['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

3

- 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。

```
4
   <tools>
   {"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"

基于画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-HFWC	qwen2.5- 7b	qwen2.5- 32b	qwen2.5- 72b	qwen2.5- 72b-int8
模型综合评估报 告	938 90011999 978 970011999	中記載 50011948 中記載 50011948	STATE SEE THE STATE OF THE STAT	Code in Code Code Code Code Code Code Code Code	A SIGN OF PROPERTY AND	College of the Colleg	College on College of
[级别 0: 整体 DSL评估]	格式正确 率 (Format Correctne ss)		99.93%	72.84%	96.61%	96.40%	96.33%
			71.25%	29.54%	48.63%	54.76%	54.68%

	完整匹配		李冠群 50011998				李冠群 50011998	
	准确率	- 1 M 50011998						
	综合准确		-011998	95 70 100	1998		201398	- 370
	率 (函数名		70 2706	20.0406	CO 450/	CE 420%	CE 400%	
	+参数结构	- mt 50011998	76.37%	39.84%	60.45%	65.42%	65.49%	
	+闭域值)	Wilder.						ale to
	李冠第5001199		李冠群 50011098		李祖縣 50011998		李冠群 50011998	-
	<mark>宽松</mark> 综合							
	准确率	李冠群 50011998						李节
	(函数名		=0011998				cn011998	
	+非"无"		李冠牌				李冠麟为	
	参数结构							40.50
	+非"无"	3510.						Skin
	闭域值)		李冠群 50011998				李冠群 50011998	
.0017	准确率					-2011998		-
级别 1: 函数名	/ 正明年 (Accuracy	李冠群 500-	77.52%	53.31%	72.12%	78.10%	78.39%	李节
评估]	(Accuracy		11.3270	J3.31 /0	1 Z. 1 Z / U	10.10 /0	10.35 /0	
)							_
	准确率	李冠群 50011998						李节
	(Accuracy		98.51%	77.84%	86.71%	86.53%	86.31%	
)		第70 m				李冠矿	
	→ 14	李冠群 50011998		李冠群 50011998		李冠群 50011998		李阳
	宏观平均-		01198				98	
	精确率		9/8/H 50011	/	李冠科 50011-		#RE 500115	
	(Macro-	-50011998	90.71%	96.53%	98.84%	94.05%	93.89%	
[级别 2: 参数名	Avg	藥冠群 500-						李节
评估 (当函数名	Precision)		李冠群 50011998				李冠群 50011998	
正确时)]	宏观平均-							
李冠群 50011	召回率	李祖牒 50011996						李节
	(Macro-		90.65%	87.78%	93.99%	88.55%	88.29%	
	Avg		- Spire		914		Shine -	
	Recall)	观剂群 50011998						- 東南
	-01199		011998	-	-011998			-
	宏观平均-		李冠群 5002-				李冠群 5001	
	F1	-2011998	90.59%	90.83%	95.55%	90.38%	90.16%	
	(Macro-	李冠群为	30.3370	30.0370	33.3370	30.30 /0	30.1070	李节
	Avg F1)		李冠群 50011998				李冠群 50011998	
[级别 3: 参数值	闭域参数							1
评估 (当参数名	值-准确率	李担牒 50011996						李军
正确时)]			100.00%	96.79%	96.82%	96.95%	96.96%	
TC1/用 L7 7]	(Accuracy							
	p8 /	#54W 50011998		50011998		#50011998		李元
	-0.199		96.37	93.10	94.76	95.44	95.37	

京高橋 50011998		开域参数 值-编辑距 离Avg. Fuzz Ratio	李冠斯 20017288	#REB 50011998			##EN 50011998	ARM 20017249
安祝朝 50011998		开域参数 值-Avg. BLEU- 1/2/3/4	#18 III 50011948	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
#SEE 50011948		开域参数 值-Avg. ROUGE- 1/2/L	\$1810 con17588	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
FIETS 5002 (STATE)	李冠群 500115	时延 (秒)	李岩斯 50011998	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			Company of the Compan	Communication of the communica	Land and the second a	AND THE STREET, AND THE STREET	The second secon
[级别 0: 整体 DSL评估]	格式正确 率 (Format Correctne ss)	#### 20011248	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)	₩RM 50011998	28.00%	22.00%	34.00%	32.00%	32.00%
	<mark>宽松</mark> 综合 准确率	李冠閣 50011998	#3E# 50011993	李冠朝 50011998	# 1816 50011998	#18 th 50011993	#1818 50011998

	(函数名 +非"无" 参数结构	李冠暦 50011998	学冠科 50011998				#1921 500119 ¹⁰
	+非 "无" 闭域值)		李冠斯 50011998				李冠科 50011998
[级别 1: 函数名	准确率	李冠群 5001155					
评估]	(Accuracy	- 4011993	31.00%	34.00%	57.00%	59.00%	60.00%
20 Jan 19	准确率	geros.	-08	SE TOP		9.100	
	(Accuracy	銀冠隊 50011993	90.32%	64.71%	59.65%	54.24%	53.33%
	宏观平均-		\$18EE 50011988		李祖斯 50011988		#15EE 50017068
	(Macro-	李冠群 50011998	98.92%	96.47%	94.33%	94.92%	95.00%
[级别 2: 参数名 评估 (当函数名	Avg Precision)		李冠碑 50011998		李冠群 50011998		李冠群 50011998
正确时)]	宏观平均-	李冠群 50011998					
	召回率		- # 50011998				- mr 50011998
	(Macro-		97.31%	83.61%	83.75%	82.61%	82.07%
	Avg Recall)	李冠群 50011998					
	宏观平均-		9-14-45 SUV		39 Table 5019 2-2		9/19/1
	F1	李冠群 50011998	07 620%	07 610%	07 220%	87.13%	86.79%
	(Macro- Avg F1)		97.63%	87.61%	87.23%	81.13%	86.79%
[级别 3: 参数值	闭域参数	李冠群 50011998		李冠群 50011998		李阳群 50011998	
评估 (当参数名	值-准确率		100.00%	100.00%	98.25%	98.31%	98.33%
正确时)]	(Accuracy		100.00 /0	100.00 /0	30.23 /0	30.3± /0	50.5570
)	李冠群 50011998					
	开域参数		李冠群 50011998				李冠群 50011998
	值-编辑距	003	05.01	0Ω 1 5	86.50	02.10	93.33
	离Avg.	李冠群 500112-	95.91	89.15	80.50	93.19	33.33
	Fuzz Ratio		等冠群 50011998				李冠群 50011998
	开域参数	-184 50011998	93.04 /	88.22 /	89.98 /	93.03 /	93.17 /
	值-Avg.	李祖	52.83 /	56.04 /	60.60 /	59.13 /	58.95 /
	BLEU-		17.71 /	17.39 /	27.19 /	26.74 /	27.21/
	1/2/3/4		5.07	2.90	8.84	10.97	11.77
		36 30 M	011998	91100	211998	99 NAMES SULVER	241998
	李冠器 5001		李冠群 50011-				李冠群 50011

#HEM 50011998	开域参数 值-Avg. ROUGE- 1/2/L	9782B 50011955	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
\$FEIR 50011	时延 (秒)	李程曆 50011998	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	SEE SOOTION		THE REPORT OF THE PROPERTY OF	Description of the second of t	A STATE OF THE STA	The second secon	A CONTROL OF THE PROPERTY OF T
# 50011988 # 50011988	格式正确 率 (Format Correctne ss)	東記 爾 90011998 東記蘭 90011998	100.00%	75.30%	99.90%	99.90%	99.80%
	完整匹配准确率	李冠曆 50011998	87.85%	37.35%	65.38%	68.32%	68.02%
[级别 0: 整体 DSL评估]	综合准确 率 (函数名 +参数结构 +闭域值)	\$15E 20011808	94.64%	51.32%	78.95%	81.88%	81.78%
	宽松综合 准确率 (函数名 +非"无"	#18EB 20011988	9/88 S00119/8	\$78.00 SOULISES	学院前 50011998	#12EE 50011958	STREET STOLLESSE
	参数结构 +非"无" 闭域值)				\$150 S0011998		李冠都 50011998
[级别 1: 函数名 评估]	准确率 (Accuracy)	WEEK SHILLING	95.95%	69.03%	93.02%	96.96%	97.06%
EN 50011998	東西部 50011998		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)	李冠師 50011998					# TSE 50011998
	宏观平均-	-n011 ⁹⁹⁸	97EB 50011998	sn011988	#18# 500119 ⁹⁸		学形態 50011998
	(Macro- Avg	李冠郎 30	99.73%	99.93%	99.75%	99.95%	99.95%
	Precision)	THE 50011998		5091 50011998		50011998	
	宏观平均-	Spi Inc.					東西郡 50011998
	召回率 (Macro-	李冠群 50011998	99.71%	91.08%	95.11%	94.49%	94.40%
	Avg Recall)						李冠群 50011998
	宏观平均-	李阳原 50011998		李冠科 50011998		李冠群 50011998	
	F1 (Macro- Avg F1)		99.63%	94.19%	96.60%	96.28%	96.23%
李冠群 50011	AVG (1)	李冠群 50011998		率冠群 50011998		李冠群 50011998	
	闭域参数 值-准确率 (Accuracy		100.00%	96.63%	96.74%	96.87%	96.87%
)	李冠群 300					
	开域参数 值-编辑距	李冠群 50011998	00.20	02.20	06.53	05.05	05.00
[级别 3: 参数值 评估 (当参数名	离Avg. Fuzz Ratio		96.38	93.28	96.52	95.95	95.86
正确时)]	开域参数	李阳曆 50011998	96.81/	93.13 /	96.78 /	95.98 /	95.92 /
	值-Avg.		56.29 /	66.28 /	64.48 /	64.33 /	64.33 /
	BLEU- 1/2/3/4		11.71 / 5.47	15.00 / 6.17	13.90 / 6.17	14.06 / 6.47	14.12 / 6.54
	1/2/3/4	3/10/11	3.71	0.11	0.17	0.11	0.54
	开域参数		96.41/	92.71 /	96.47 /	95.86 /	95.80 /
	值-Avg.		55.92 /	65.82 /	64.19 /	64.18 /	64.18 /
	ROUGE- 1/2/L	20120	96.41	92.71	96.46	95.85	95.79
#RB 5001X	时延 (秒)	李冠群 50011998	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	978 FF 902109	97ER (0011998	particular to the second of th	The second secon	Comment of the section of the sectio	The second secon	TOTAL OF THE STATE
#### 20017 #### 20017	格式正确 率 (Format Correctne ss)	\$78H 50011998	100.00%	85.00%	91.00%	88.00%	88.00%
	完整匹配准确率	98	97.00%	24.00%	6.00%	61.00%	63.00%
[级别 0: 整体 DSL评估]	综合准确 率 (函数名 +参数结构 +闭域值)	\$120 000100 \$120 0001008	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)	#FEER 10011998	\$1818 S0011998	\$12 M 50011988	STREET STOLLARS	# 15 E 500 L1 9 6 6	\$100 STATE S
[级别 1: 函数名 评估]	准确率 (Accuracy)	97EE 5001,998	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)	#1819 50011998	100.00%	100.00%	100.00%	100.00%	100.00%
	宏观平均- 精确率 (Macro- Avg Precision)	#FER 5001(998)	0.00%	0.00%	0.00%	0.00%	0.00%
	宏观平均- 召回率 (Macro-	\$18TH :0017889	0.00%	0.00%	0.00%	0.00%	0.00%

	Avg Recall)					-rs 50011938	李冠群 50011998
	宏观平均- F1 (Macro- Avg F1)		0.00%	0.00%	0.00%	0.00%	0.00%
\$78 B 50011988	闭域参数 值-准确率 (Accuracy)	4018 50011998	0.00%	0.00%	0.00%	0.00%	0.00%
[级别 3: 参数值 评估 (当参数名	开域参数 值-编辑距 离Avg. Fuzz Ratio	\$150 S0011918	0.00	0.00	0.00	0.00	0.00%
正确时)]	开域参数 值-Avg. BLEU- 1/2/3/4	\$182 8 20011988	0.00 / 0.00 / 0.00 / 0.00				
	开域参数 值-Avg. ROUGE- 1/2/L	#18.05 50011.993	0.00 / 0.00 / 0.00				
李祖郎 50011998	时延 (秒)	\$100 PM	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	\$125 50010 ¹⁰	\$1818 00011998 \$1818 10011998	TO TAKE THE REPORT OF TAKE THE PROPERTY OF TAKE THE	The second secon	10 10 10 10 10 10 10 10 10 10 10 10 10 1	The second secon	AND
[级别 0: 整体 DSL评估]	格式正确率		100.00%	83.00%	100.00%	99.00%	99.00%
	(Format						# 元素 50011993

	Correctne ss)	-1998			家居器 50011993		\$18 SOULES
	完整匹配 准确率	pp 11 500 L	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)	#FEM 50011998	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)	#1210 50011995	#20011998 #20011998	#### 50011998	WHEN 00011998	\$120 SOO 11938	978 H 50011998
[级别 1: 函数名 评估]	准确率 (Accuracy)	STREETS SOUTHWARE	0.00%	0.00%	4.00%	1.00%	1.00%
所 50011 ⁹⁵⁸	准确率 (Accuracy)	泰祖郡 50011998	0.00%	0.00%	100.00%	100.00%	100.00%
[级别 2: 参数名	宏观平均- 精确率 (Macro- Avg Precision)	#1810 S0011998	0.00%	0.00%	100.00%	100.00%	100.00%
评估 (当函数名 正确时)]	宏观平均- 召回率 (Macro- Avg Recall)	#18215 S0011990	0.00%	0.00%	100.00%	100.00%	100.00%
	宏观平均- F1 (Macro- Avg F1)	##### 5001199B	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	#### 5001199B	0.00%	0.00%	0.00%	0.00%	0.00%

		(Accuracy	_{month} 50011988		FRIN 50011988		FRIN 50011998	#1818 S1011998
		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	₹RE 50011988	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	学冠目 50011998 学冠目 50011998	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
新福縣 50011998	- NO 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	时延 (秒)	-w c0011998	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		李祖朝 50011999	The second secon	TOTAL TOTAL STATE OF THE STATE	The second secon	THE STATE OF THE S	THE REPORT OF THE PARTY OF THE
[级别 0: 整体 DSL评估]	格式正确 率 (Format Correctne ss)	#15EM 50011990	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)	#FEER 50011999	0.00%	0.00%	15.00%	5.00%	5.00%
	東阳暦 5001199		李冠群 50011998		東西縣 50011998		李冠群 50011998

<mark>宽松</mark> 综合		李冠群 50011998				李冠群 50011998	
准确率							
(函数名							- 1
+非"无"		011998				1998	
表现 经		李冠群 50011				李冠群 500 12	
闭域值)							3
9738 50011993		9-70-FE 50011998		9/108K 50011993		9-738 50011998	
198							
(Accuracy		0.00%	0.00%	15.00%	5.00%	5.00%	1
955E 50011393		李冠群 50011998				李冠群 50011998	
准确率							
(Accuracy		0.00%	0.00%	100.00%	100.00%	100.00%	-
), 188 50013993		李冠群 50011998				李冠群 50011998	
· 安侧亚杓	093		8				
							等
-0011998		0.00%	0 00%	100 00%	100 00%	100 00%	
· ·		0.00 /0	0.00 /0	100.00 /0	100.00 /0	100.00 /0	
. (30131011)		es 50011992		mer 84 500 11958		一号拼 50011938	
宏观平均-		- 10 mm				Skiller	
召回率							
(Macro-		0.00%	0.00%	100.00%	100.00%	100.00%	
Avg		李冠群 500113-5				李冠群 50011995	
Recall)							
宏观平均-		1998	,			-011998	
F1		0.000/	0.0007	100.000/	100.000/	100.000/	
(Macro-		0.00%	0.00%	100.00%	100.00%	100.00%	
Avg F1)							
河域会 粉		3-1051 <0011998		9-13-14 cn011998		李四周 20011398	
		0.00%	0.00%	0.00%	0.00%	0.00%	- 19
Inccuracy		李冠群 50011998				李冠群 50011998	
9			۶۵۸.		. A0A.		
开域参数							等
值-编辑距		李元群 50011998				李冠群 50011998	
离Avg.		0.00	0.00	41.53	47.20	47.20	
Fuzz							
Ratio		msa # 50011998				TO THE SOULE OF THE SOURCE OF	
开试参数		0.00 / 0.00	0.00./	45 77 /	42 16 /	42 16 /	
98							ガ
ı≞-∧vg.		0.00	0.00 /	10.03 /	10.01 /	10.01/	
	准(中) 本) 本) 本) 本) 本) 本) 本) 本) 本) 本	准确率名+非"活构+非"词(加率(Accuracy) 准值。不可以,是在的人类的,并不可以,是在的人类的,并不可以,是在的人类的,并不可以,是一个人类的,并且一个人类的,并且一个人,并且一个人,并且一个人,并且一个人类的,并且一个人,并且一个人,并且一个人,并且一个人,并且一个人,并且一个人,并且一个人,并且一个人,也可以是一个一个人,也可以是一个人,也可以是一个一个人,也可以是一个人,也可以是一个人,也可以是一个人,也可以是一个人,也可以是一个人,也可以是一个人,也可以是一个人,也可以是一个人,也可以是一个人,也可以是一个人,也可以是一个一个,也可以是一个一个人,也可以是一个一个一个,也可以是一个一个一个一个,也可以是一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个	准确率 (函数名 +非 "无" 参数结构 +非 "无" 闭域值) 准确率 (Accuracy) 宏观平均- 精确率 (Macro- Avg Precision) 宏观平均- 召回率 (Macro- Avg Recall) 宏观平均- 日 (Macro- Avg Recall) 宏观平均- 日 (Macro- Avg Recall) テ1 (Macro- Avg F1) 対域参数 値-准确率 (Accuracy) 开域参数 値-知要 (Accuracy) カ0.00% の.00%	准确率 (函数名 +非 "无" 参数结构 +非 "无" 所域值) 0.00% 0.00% 准确率 (Accuracy) 0.00% 0.00% 法确率 (Accuracy) 0.00% 0.00% 宏观平均- 相确率 (Macro- Avg Precision) 0.00% 0.00% 宏观平均- 日(Macro- Avg F1) 0.00% 0.00% 局域参数 值-准确率 (Accuracy) 0.00% 0.00% 开域参数 值-编辑距 高Avg. Fuzz Ratio 0.00 / 0.00 0.00 / 0.00 开域参数 值-Avg. 0.00 / 0.00 0.00 / 0.00	准确率 (函数名 +非 "无" 参数结构 +非 "无" 例域值) 准确率 (Accuracy) 0.00% 0.00% 15.00% 准确率 (Accuracy) 0.00% 0.00% 100.00% 宏观平均-精确率 (Macro-Avg Precision) 宏观平均-F1 (Macro-Avg Recall) 宏观平均-F1 (Macro-Avg F1) 別域参数 値・准确率 (Accuracy) 0.00% 0.00% 0.00% 100.00% 非确率 (Accuracy) 0.00% 0.00% 100.00% 10	准确率 (函数名 +非 "无" 参数结构 +非 "无" 所球値)	准備率 (函数名 +非 "无" 参数结构 +非 "无" 的域值) / 推備率 (Accuracy) 0.00% 0.00% 15.00% 5.00% 5.00%

\$188 50011998	BLEU- 1/2/3/4			0.00 /	9.75 / 2.67	15.71 / 11.90	15.71 / 11.90
	开域参数 值-Avg. ROUGE- 1/2/L	\$100 Jan	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
9/8EH 5001	时延 (秒)	季冠帽 50011998	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。
- 在<tools></tools>标签中为你提供了详细的工具列表。 2

3

4 <tools>

```
{"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。