**线上数据闭环-标注规范**

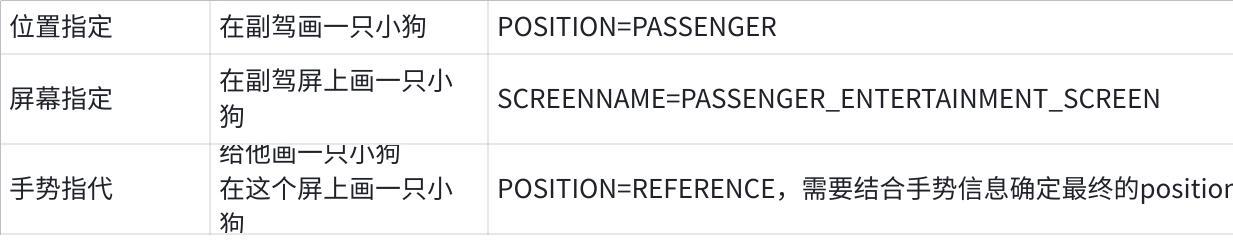
**参考资料**

[API标注文档](https://li.feishu.cn/docx/Vog6dN0U9oefLMxv4cdcoeoZnZe)

[API范围梳理](https://li.feishu.cn/docx/RGdhdykqQon4paxoXkEcRW1dnbL)

**标注规范**

* 通用问答[API范围梳理](https://li.feishu.cn/docx/RGdhdykqQon4paxoXkEcRW1dnbL)
* Category: 【人物】【公司】【影视】【文学】【音乐】【历史】【地理】【**世界之最】【体育】【车牌查询】【文学推荐】【其他问答】【地理通用】【日历】@辛洪生**
* 出游灵感
* category【国家】【城市】【景点】【出游美食】【旅游】【城市景点推荐】@常凌[OTA 5.1 出游灵感评测集 预期API 标注标准说明](https://li.feishu.cn/docx/HsbCdgihwopbYkxTKRScC3qpnzg)
* 汽车问答@汪洋[汽车问答NLU-API多轮对话标注规范](https://li.feishu.cn/docx/GKqNdXiZtoyOxDxhKTScLmWmnbb)
* 媒体搜索@常凌[媒体搜索数据优化标注说明](https://li.feishu.cn/docx/T0tBdXfnRorqjjxUbkrcHwjMnR0)[媒体搜索API标注&槽位对齐更新](https://li.feishu.cn/docx/CpqZdZVcpo6k5vxguX9cv4jBnch)
* 绘画大师@张崇宇[绘画大师-参数说明与指标计算规则（语义）](https://li.feishu.cn/docx/UqJtdfGiSok53pxPdTrcU52vnHg) [绘画大师多轮测试集数据需求与规则](https://li.feishu.cn/docx/MUSRdKWuDo4qoNxwWUncPrsendh) 、位置/屏幕/手势可参考需求文档：[SS\_Max\_理想同学\_绘画大师\_需求文档\_001](https://li.feishu.cn/docx/IPYqdYI9YoOWAsxwXWscL8omnMd)
* 绘画单轮，APINAME，KEYPHRASE，CATEGORY，STYLE，POSITION/SCREENNAME
* 绘画多轮：keyphrase、category、style等的信息继承API的标注
* 绘画位置指定POSITION和屏幕指定SCREENNAME），手势指代（POSITION）



**点击图片可查看完整电子表格**

**重点case标注规范**

1. 对于当前query属于应该被拒掉的query，API结果标成 reject

|  |  |  |
| --- | --- | --- |
| 上文 | 当前query | API |
| [{"user": "声音开到60", "li": "好的，帮你把媒体音量调到60啦。"}] | 哪一家呢？ | reject |
| [{"user": "声音调大", "li": "搞定。"}] | 可以了 | reject |

1. 对技能型的query，闲聊的数据，API应该标成 None

|  |  |  |
| --- | --- | --- |
| 上文 | 当前query | API |
|  | 帮我写一首诗 | None |
|  | 讲个笑话吧 | None |
|  | 分析/预测一下明天的股票走势 | None |
|  | 二零二四年起名字哪几个字比较好  我的妈妈离婚且再婚我该怎么办  什么东西越洗越脏脑筋急转弯 | None  None |

1. ASR识别错误的case，标注 ASR错

|  |  |  |
| --- | --- | --- |
| 上文 | 当前query | API |
|  | 为什么快点跑得快？ | ASR错 |
|  | 北京是不是（语音截断的数据） | ASR错 |

1. 对于应该拒识的数据，标 reject

|  |  |  |
| --- | --- | --- |
| 上文 | 当前query | API |
|  | 停电 | reject |

1. 对于属于车控的case标注成 TOD （一句话中有多个指令的也标TOD）

|  |  |  |
| --- | --- | --- |
| 上文 | 当前query | API |
|  | 商场/公寓/加油站在哪个区/省/位置 | TOD |
|  | 打开/取消/后退 | TOD |
|  | xx到xx的距离 | TOD |
|  | 今天现在车外空气质量怎么样//台州元旦节一月一号会下雨吗 | TOD |
|  | 问广场/商场/公寓等的所属城市 | TOD |



1. 人设的case标注成 人设



1. 数学计算类的标成 mathQA

|  |  |  |
| --- | --- | --- |
| 上文 | 当前query | API |
|  | 1+1等于几 | MathQA |

1. 对于只说歌曲名称的词语，只有热榜的歌曲才会走TOD，否则拒识。

|  |  |  |
| --- | --- | --- |
| 上文 | 当前query | API |
|  | 青花瓷 | TOD |
|  | 忠孝东路走九遍 | reject |

1. xx开车/高铁到xxx多久/多远 包括问火车票/机票的 这种数据在备注列备注上 出行助手5.3新增“交通出行，
2. 诗词/歌词的反向查询

|  |  |  |
| --- | --- | --- |
| 上文 | 当前query | tag |
|  | **总有一天我会比你早死去因为我的眼睛比你大十四岁**这是哪一首歌 | **总有一天我会比你早死去因为我的眼睛比你大十四岁&歌词查询** |
|  | 不问苍生问鬼神出自哪首诗 | **不问苍生问鬼神**&**诗词查询** |
|  | 我本意心向明月下一句是什么 | 我本意心向明月&下一句 |

**数据标注规则**

[API标注规范](https://li.feishu.cn/docx/RGdhdykqQon4paxoXkEcRW1dnbL)

送标数据只标一个CATEGORY

|  |  |  |
| --- | --- | --- |
| query | api |  |
| **白朴天净沙秋**原文是什么 | [{'apiname': 'QASearch', 'category': '文学', 'query': '白朴天净沙秋原文的内容是什么', 'tag': '白朴天净沙秋&原文'}] | 诗词名称→原文 |
| 读一下古诗**古朗月行** | {'apiname': 'QASearch', 'category': '文学', 'query': '读一下古诗，古朗月行。', 'tag': '古诗&古朗月行'} |  |
| 曹丕的**七步诗**的内容 | {'apiname': 'QASearch', 'category': '人物', 'query': '曹丕的七步诗的内容', 'tag': '曹丕&七步诗&内容'} | 主要是事件相关典故 |
| **贺知章**的**咏柳**这首诗的内容是什么 | {'apiname': 'QASearch', 'category': '人物', 'query': '贺知章的咏柳这首诗的内容', 'tag': '贺知章&咏柳&内容'} | bing：诗词内容+赏析 |
| **总有一天我会比你早死去因为我的眼睛比你大十四岁**这是哪一首歌 | [{'apiname': 'QASearch', 'category': '音乐', 'query': '总有一天我会比你早死去，因为我的眼睛比你大14岁，这是哪一首歌', 'tag': '总有一天我会比你早死去因为我的眼睛比你大14岁&歌词查询'}] | 歌词内容→歌曲    音乐问答里：**大约有10/82=12%的query**，关于歌词反向查询的 |
| **我喝过你喝过的水**在哪个歌词里哪个歌里面有底线 | {'apiname': 'QASearch', 'category': '音乐', 'query': '我喝过，你喝过的水在哪个歌词里面', 'tag': ''我喝过你喝过的水在&歌词查询'} |  |
| 【歌词】是什么歌 | 春去春会来花谢花会再开是哪首歌 | 来自产品补充 |
| 歌词里面有我曾经还是那个少年是什么歌 |
| 刚买了iphone五又买了iphone六是什么歌 |
| 【歌词】是哪首歌的歌词 | 红尘一笑浪滔滔是哪首歌的歌词 |
| 我在人民广场吃着炸鸡是哪一首歌的歌词 |
| 是不是有首歌歌词里有【歌词】 | 是不是有首歌歌词里有爸爸妈妈给我的不少不多 |
| 有没有一个歌歌词有我们都已经长大 |
| 有首歌里【歌词】是什么歌 | 有一首歌里我真的陪他聊到黎明是什么歌 |
| 有首歌里唱转身离开你有话说不出来是什么歌 |
| 标TOD | 问天气预报，天气查询的数据 |  |
| 其他问答 | 摩托车/电动车/三轮车 |  |
| 诗词的下一句，上一句。以及诗词查询，要标文学 | 天街小雨润如酥的上一句 CATEGORY=文学，TAG 天街小雨润如酥&上一句  天街小雨润如酥 出自哪首诗？ CATEGORY=文学，TAG 天街小雨润如酥&诗词查询 | 注;query中不可以出现诗词查询/诗词出处， |
| 歌词的上一句，下一句，以及歌词查询，要标音乐 | 歌词也就是是是非非的上一句是什么 CATEGORY=音乐，TAG 也就是是是非非&上一句  也就是是是非非的出自哪首歌？ CATEGORY=音乐，TAG 也就是是是非非&歌词查询 |  |
| 数字计算类的也标mathqa | 七年前妈妈的年龄是儿子的6倍，儿子今年12岁，妈妈今年多少岁？ |  |

**scheme**

**诗词**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Category** | **意图类别** | **意图** | **TAG** | **优先级** | **是否应走搜索** | **示例query** |
| 诗词 | 信息获取 | 诗词内容获取 | 根据题目 | P0 | 是 | 古诗春江花月夜  搜索古诗闻王昌龄左迁龙标遥有此寄 |
| 根据作者 | P0 | 是 | 给我放首李白的古诗吧  说一首王维的诗 |
| 根据内容 | P0 | 是 | 我本意心向明月下一句是什么  十年生死两茫茫的全诗 |
| 根据字词 | P0 | 是 | 含有过字的古诗  哪些诗句里面含有风 |
| 信息查询 | 诗作属性查询 | 诗句释文 | P0 | 是 | 我本将心向明月奈何明月照沟渠什么意思 |
| 文献书目/题目 | P0 | 是 | 不问苍生问鬼神出自哪首诗 |
| 作者/朝代 | P0 | 是 | 饮湖上初晴后雨的作者还有朝代是什么 |
| 创作时间/作者背景 | P0 | 是 |  |
| 体裁/表达思想/艺术特点 | P0 | 是 | 晚春这首诗表达了什么 |
| 信息查询 | 诗词之最 |  | P1 | 是 | 苏轼写的写过最好的诗是哪一首 （人物）  世界上最难的十首古诗 （世界之最）  最难的十首古诗 （文学） |
| 生成 | 写诗 | 季节景观 |  | 否 | 你写一首关于冬天的诗 |
| 藏头/尾诗 |  | 否 | 创作一个藏尾诗 |
| 关系主题 |  | 否 |  |
| 续写/仿写 | 关于诗词风格的仿写 |  |  | 否 |  |

1. 诗词内容获取：
2. 部分内容获取判断判断是否为上下一句，如果是则优先思必驰，否的话优先大模型，大模型作为兜底
3. 上下一句：“十年生死两茫茫的下一句” 文学
4. 非上下一句：“蜀道难的第一段” 文学
5. 写生/仿写/续写走大模型 这个走None 朗读/背诵/诗词属性查询/诗词之最走文学

背一下凉州词===

背一首诗===

背一首黄鹤楼

背诵，朗读古诗的，tag需要统一标准 凉州词&全文

1. 世界之最定义，明显的世界上最\*\*\*，世界最早的，国家最\*\*，国家第一高/最高的建筑，（最高的小区这种不算世界之最）
2. 媒体搜索-影视分类下 电影在映/待映的信息，标注到槽位 【PUBLISHSTATE】中，槽位值归一化为【在映 / 待映】
3. 新闻目前只包括时间(最近/今天等) 地点(北京/上海) 主题(娱乐/财经)，不涉及具体人或者事件，包含有具体人名/事件/物品等的新闻查询，需要标成实效性相关的；
4. 水利局，城建局，公安局，派出所，机构 CATEGORY=公司

**按照语义进行区分，媒体/影视 分类下的人物ACTORS/DIRECTOR（演员/导演）**

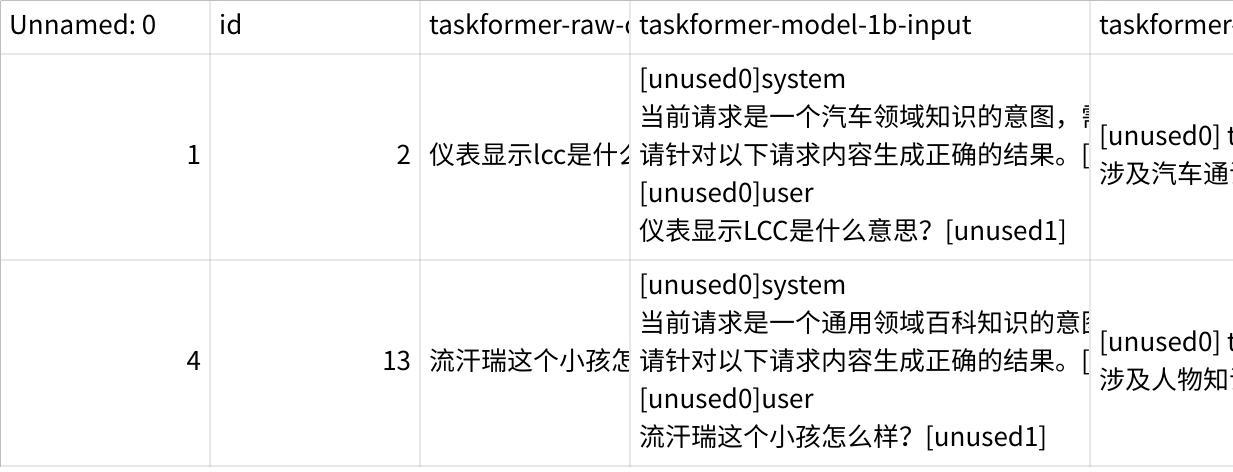
**新闻话题下的数据 他问最近就写最近，他问最新就写最新**

**只有影视推荐的，最近=最新，其他的都不用替换**

**只有新闻的时间写在PUBLISHTIME，音乐媒体都是RANKTYPE**

**送标数据格式**

1. 送标数据会提供baichuan的API和MindGPT的API结果供标注同学参考，具体参考数据格式如下：



**点击图片可查看完整电子表格**