



第十五届全国大学生服务外包创新创业大赛



长安陌上

基于Qwen大模型的语料库问答平台

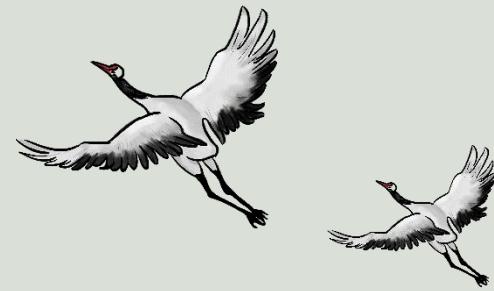


INFJ 团队

目 录



- 壹 • 项目背景
- 贰 • 产品介绍
- 叁 • 技术路线
- 肆 • 语料库构建
- 伍 • 大模型微调
- 陆 • 项目创新
- 柒 • 商业模式



项目背景



中国国内旅游人次及增长率（2020–2023）



自2020年初爆发新冠疫情以来，全球范围内的**旅游业**受到了**严重冲击**。中国旅游业在2023年展现出了明显的复苏迹象，呈现出了强劲的**增长势头**。

The figure consists of three stacked screenshots of the Chinese Government website (www.gov.cn).
Top screenshot: Headline - '以文旅高质量发展满足人民群众日益增长的需求——文化和旅游部部长孙业礼“部长通道”回应文旅热点'.
Middle screenshot: Headline - '今年暑期国内旅游人数超十八亿人次——文旅市场复苏提速'.
Bottom screenshot: Headline - '国家旅游局关于实施“旅游+互联网”行动计划的通知'.
All screenshots show the official government logo and navigation bar.

政府提出了一系列支持旅游行业**发展的政策和举措。**同时也积极推动**人工智能技术**的发展，并将其列为国家战略之一，为智能问答系统等人工智能技术在**旅游领域**的应用提供了**政策支持和指导**。

文旅行业痛点



文旅信息复杂



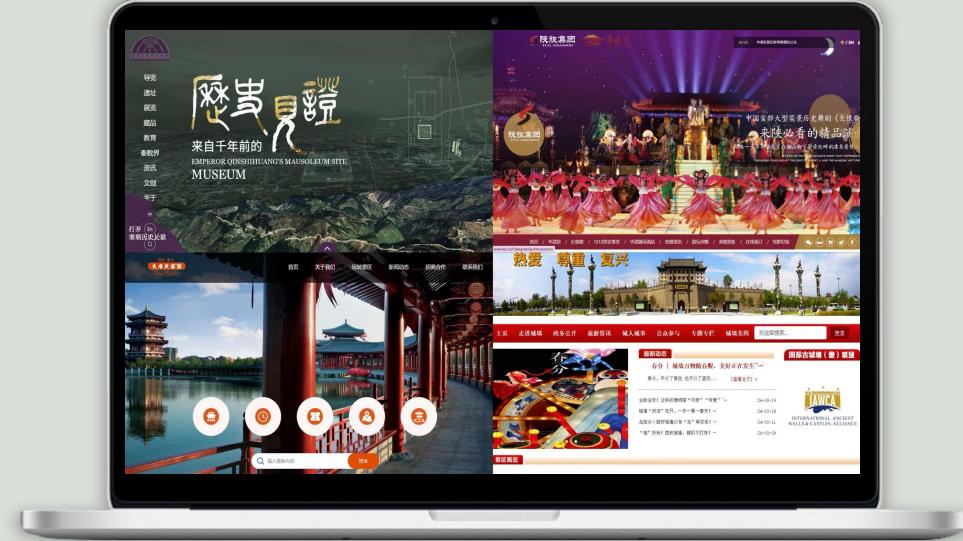
信息种类繁杂
(景点、攻略、评论等)



文旅信息分散



信息获取渠道多



针对性不强



无法针对个人问题提出建议

大模型痛点



LLM有**处理复杂任务的能力**和**强大的语言理解能力**，可以为文旅行业赋能。



但目前，**通用型LLM**在特定领域**缺乏足够的领域专业知识**。



而且市面上的LLM系统均存在一定的**幻觉**，从而导致**输出不准确**。

需要制作一个基于**大模型的文旅语料库**问答平台。

项目亮点

个性化攻略定制



提供**个性化旅游攻略**，涵盖闺蜜游、亲子游、偏人文或风景等，打造独特、难忘的旅游体验。

项目亮点

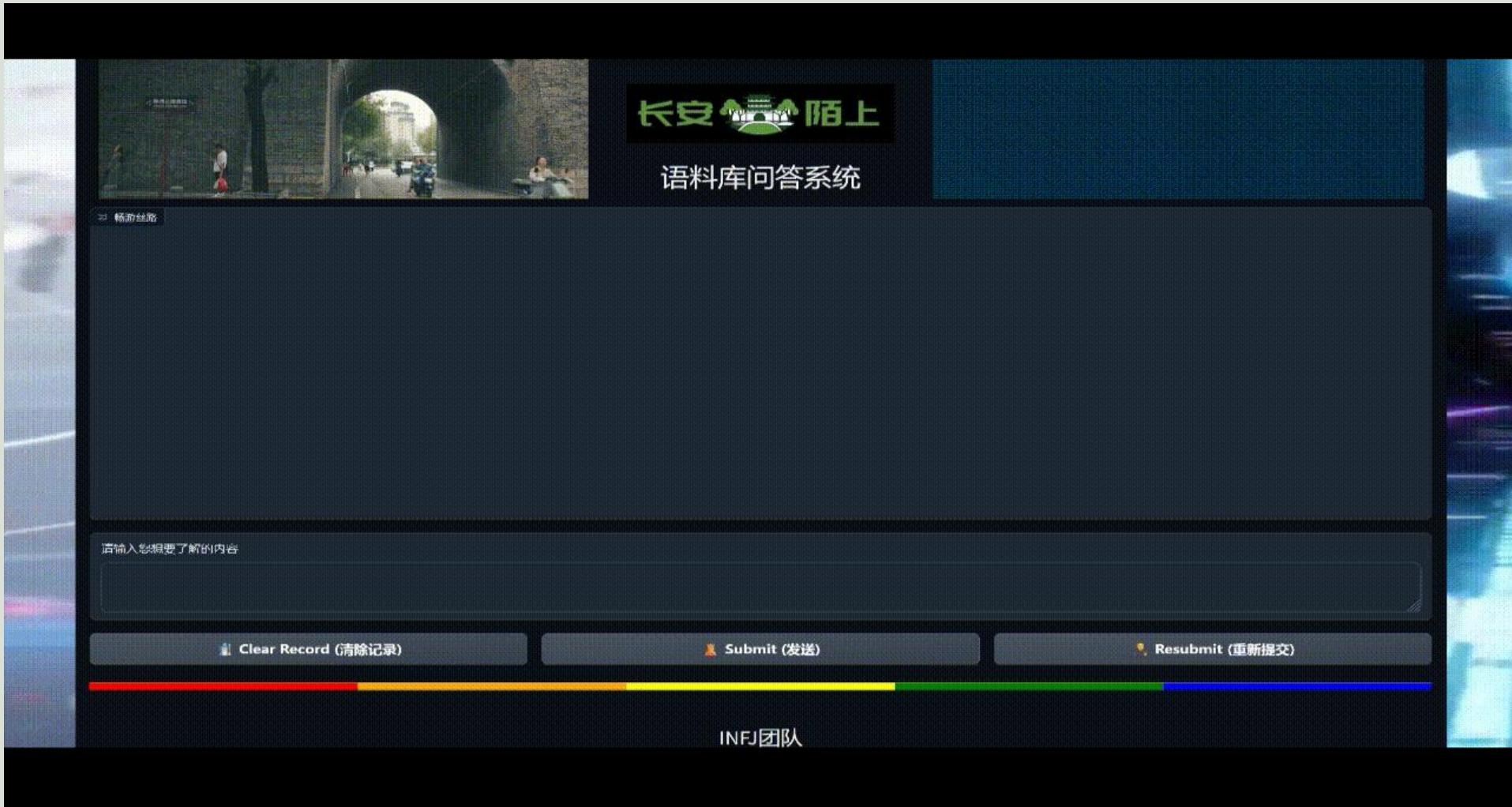
大数据情感分析



通过**大数据情感分析**，综合酒店、餐厅评价的优劣内容，提供更全面、客观的评价服务。

项目亮点

多轮问答—动态同步



通过将**多轮问答**和**动态同步**相结合，用户不仅能够连续对话，还可随时获取最新的旅游信息。

技术路线



增加模型知识广度



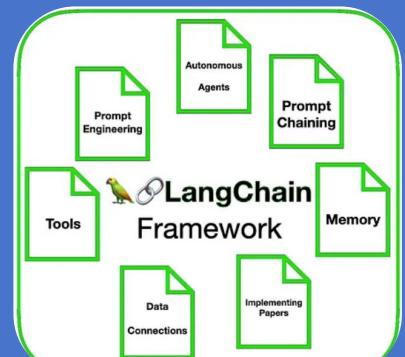
收集到的数据

提升模型问答性能

本地知识库



LangChain构建本地知识库



长安陌上LLM



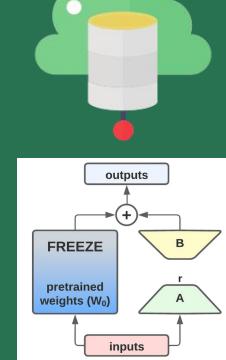
Qwen-7b-chat
大语言模型

大模型微调

语料库构建



LoRA微调



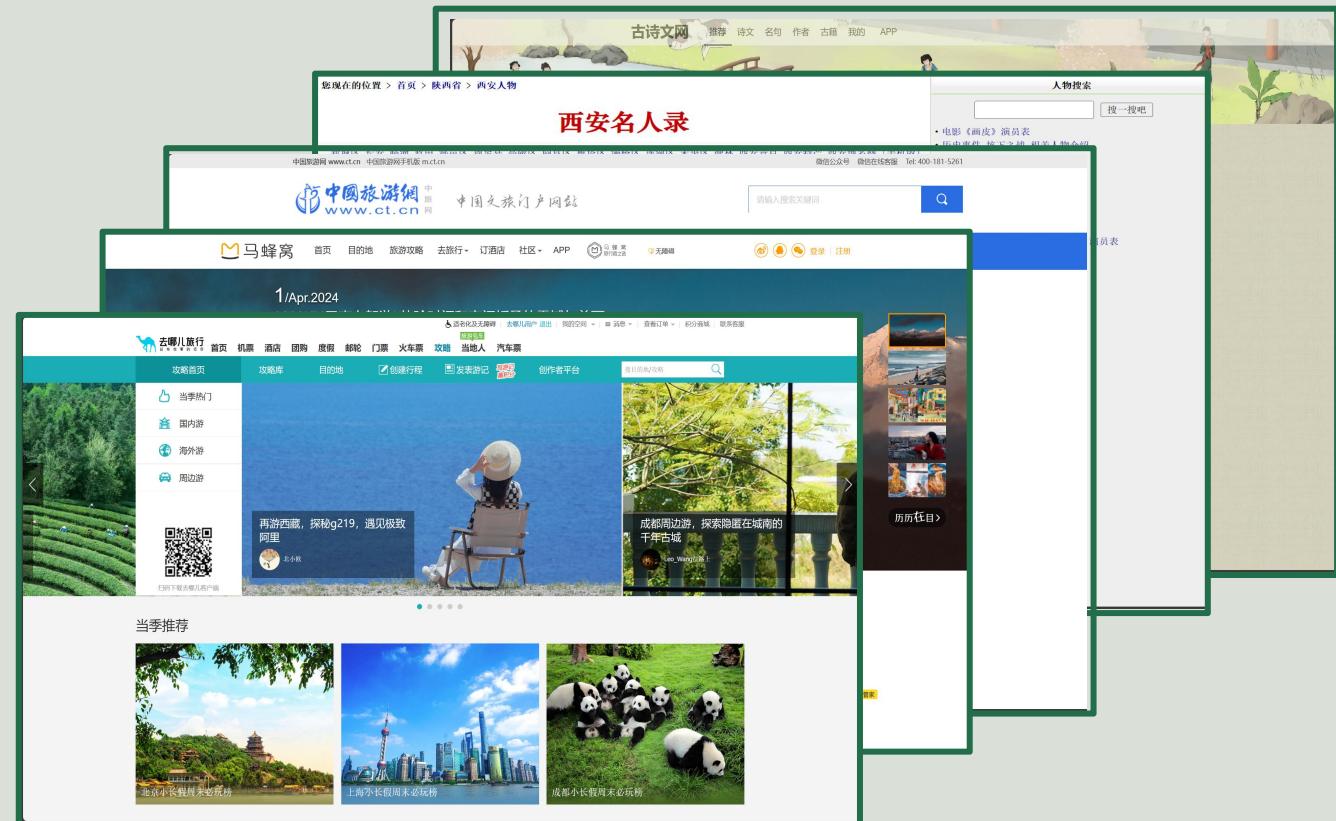
微调

生成的答案

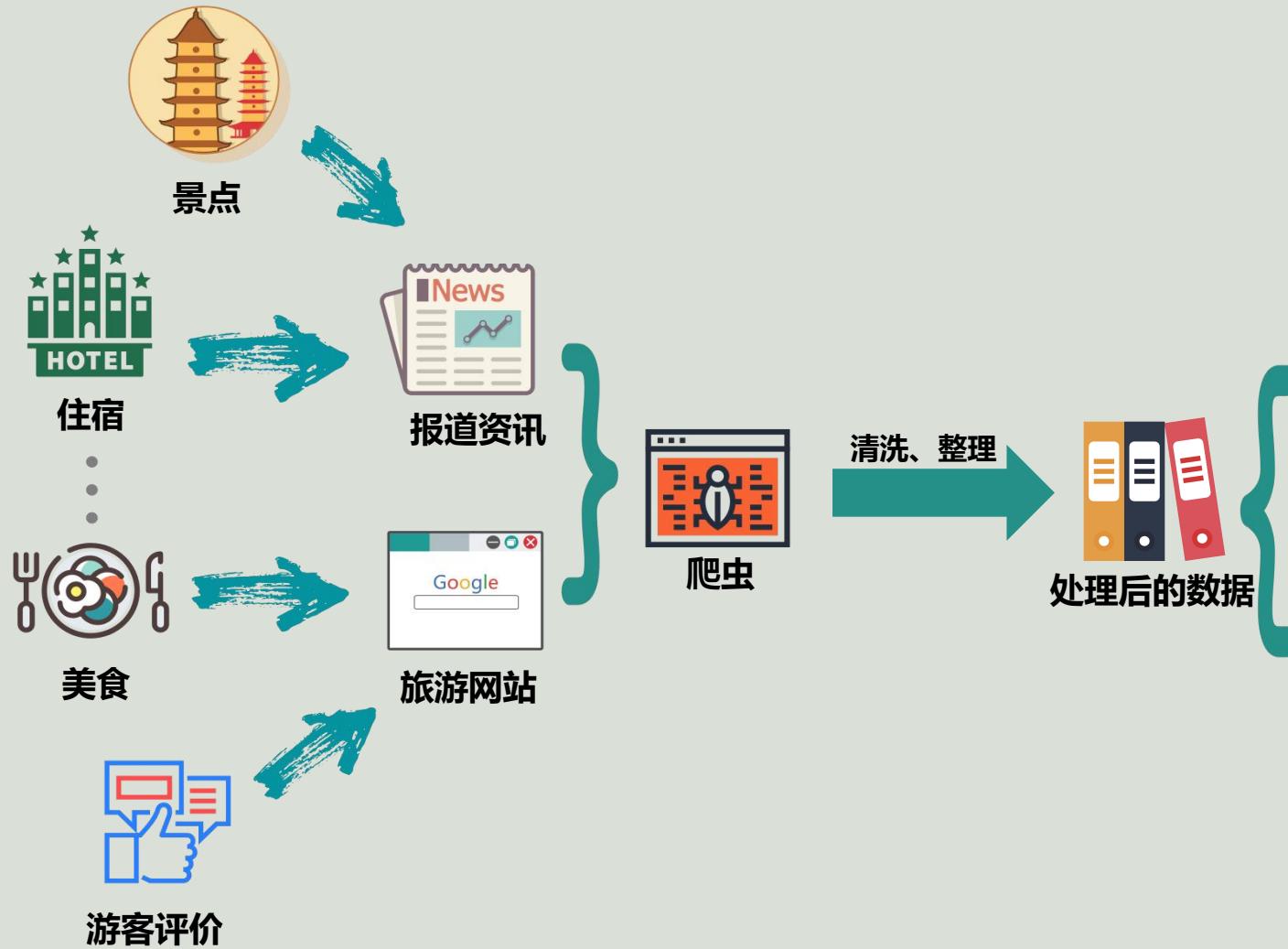
数据来源



种类	来源
景点信息	去哪儿旅游
特色美食	去哪儿旅游
美食店铺信息+评论	美团、大众点评
酒店信息+评论	美团、大众点评
出行信息	携程旅游
旅行攻略	小红书、去哪儿旅游
城市历史	百度百科
城市名人	中国名人网
城市古诗名言	古诗词网



数据处理过程



LangChain知识库

改进self-instruct——技术创新



Prompt编写原则——技术创新



①角色定义
精确，与完成任务所需背景知识匹配

②任务目标描述
简洁明了，避免使用模糊或歧义的词汇

⑤输出要求
明确输出格式要求，可再次强调重要的任务要求

③参考案例
具有代表性，分别代表不同类型

④任务具体要求思维链、步骤
多用枚举，少用否定句，如流程较长应考虑拆分为多个任务



语料库构建——Prompt案例



角色和目标

→ 这是文字整合器。根据用户输入的文本中整理出网友们对{hotel_name}的评价。

任务具体要求

评价包括以下指标：酒店卫生、酒店服务、酒店地理位置、酒店配置、酒店星级、酒店价格、酒店餐饮等。

对象可以有多个。若没有该项则可以不输出。

输出完毕后结束，不要生成新的用户输入，不要新增内容。回答要准确，不需要加工总结，只需要还原现实。

参考案例

示例模版

...

[

{

"instruction": "根据游客们的评论，{hotel_name}怎么样，请给我一个综合评价？",
"input": "",
"output": "问题的回答"

}

]

...

输出要求

→ 请根据以下文本，严格按照示例模版格式输出内容。

用户输入

→ 用户输入：{}

语料库构建——问答对生成



477 西安天泽文化酒店(浐灞国际会展中心店)
478 璞晶酒店(高新西北工业大学地铁站店)
479 西安高新半导体产业园亚朵酒店
480 地建南山温泉酒店
481 西安摩登·爱琴海酒店式公寓
482 西安唐城宾馆(小寨吉祥村地铁站店)
483 漠梓源酒店(西安沣东新城管委会店)
484 西安南山北民宿
485 西安安居巷叁号酒店(钟楼永宁门店)
486 西安大唐博相府酒店(大雁塔大唐不夜城店)
487 西安富临酒店(华清池兵马俑店)
488 西安高新唐延路亚朵酒店
489 终南·景莊温泉度假酒店
490 西安名都国际酒店
491 如家酒店·neo(西安钟楼北大街地铁站店)
492 栖玥酒店(西安大雁塔小寨地铁站店)
493 星辰酒店(西安钟鼓楼回民街店)
494 城市便捷酒店(西安湖北大夏凤城九路店)
495 博雅臻选酒店(西安北客站店)
496 加雅酒店(西安钟鼓楼五路口地铁站店)
497 全季酒店(西安钟楼店)
498 Y酒店(西安高铁北站凤城十二路地铁站店)
499 宜必思尚品酒店(陕西省人民医院莘雁店)

环境优美，地理位置好，交通方便，房间舒适，卫生服务员很热情！这次入住体验非常满意！酒店的服务态度超级好，员工们始终很专业。亚朵是最舒适的酒店，卫生整体很干净，服务很到位，亚朵的卫生太不干净了，水吧台玻璃上的灰都能写字了！这个价钱真的很超值。房间很干净整洁，酒店的所有员工都很热情周到。尤其是前台宾馆在当地多年来小有名气，周边餐饮丰富，前台服务热情周到！房间干净明亮，设施齐全，特别是前台王丽美女服务很周到！极好的入住体验，民宿开下山下的村子里，来之前还担心卫生问题，酒店位于书院门，比较古风，出游比较方便，离南门和钟楼比较近。虽然酒店是2015年装修，家具、房间设施有些许老旧。但是这家准备翻越秦岭但是进大城市的话进出都很花时间，所以选择这家。房间干净整洁，有地暖，住着舒服，地理位置也可以，出行方便。环境非常好，服务也不错，适合情侣度假！2. 深秋入住，红枫挂牌4星级酒店，2022年新装修的，配2份早餐，空调24小时已开启。工作人员非常热情，房间干净整洁，地理位置近地铁站和回民街。服务态度特别好，办理入住的时候很快，房间卫生也很干净，酒店位置离钟楼地铁站很近，但在居民区里所以很安静，不会影响休息。酒店的整体感觉非常不错，房间干净整洁，设施也很完善。员工服务特别好，就在五路口地铁口附近去哪里，就在五路口地铁口。服务特别热情，酒店工作人员会特别nice的说话。设施很齐全，周边环境很不错很多吃的，可是我赶时间没去吃到。早餐也不错，酒店的服务人员态度很好，有求必应，耐心的解答，停在方便。



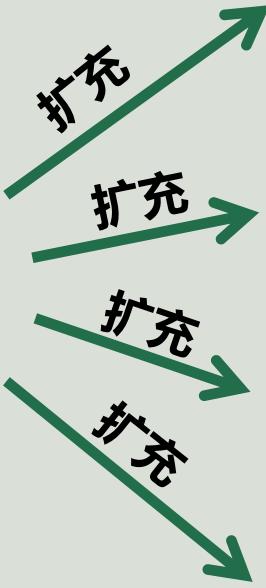
```
[  
 {  
 "instruction": "根据游客们的评论，你觉得24H color酒店(西安大雁塔西影电影博物馆店)怎么样，请给我一个综合评价？",  
 "input": "",  
 "output": "根据游客们的评论，24H color酒店在西安大雁塔西影电影博物馆店的位置十分便利，"  
 "房间宽敞整洁，环境优美，提供贴心的服务和小点心。"  
 "客人们对酒店的设计风格、服务热情度和早餐品质都给予了积极评价。"  
 "此外，酒店还受到好评的地方包括咖啡店的工业风格、房间设施的完善以及服务细节的体贴。"  
 "然而，有部分客人提到停车费较高这一点。"  
 "总体来说，24H color酒店在西安大雁塔西影电影博物馆店获得了多方好评，是一家性价比较高且值得推荐的酒店。  
 }  
,  
 [  
 {  
 "instruction": "根据游客们的评论，你觉得24Hcolor酒店(西安大寨路三五一一店)怎么样，请给我一个综合评价？",  
 "input": "",  
 "output": "24Hcolor酒店(西安大寨路三五一一店)在游客评价中表现出艺术感强、"  
 "服务细心周到、环境优美、设施完善的特点。"  
 "客人们对酒店的空间大小、服务态度、早餐品质、咖啡厅设施等方面都给予了积极评价。"  
 "部分客人提到酒店距离景点较远、卫生间设计有些特殊，"  
 "但总体来说，酒店在价格、服务和环境方面都得到了客人的认可，是一家性价比较高、适合休闲度假的酒店。  
 }  
 ]
```

酒店名称

评论数据

原始数据由**改进的self-instruct技术**得出问答对

西安贾三清真灌汤包子馆的交通指南？



西安贾三清真灌汤包子馆如何才能到达？

西安贾三清真灌汤包子馆应该怎样才能过去？

前往西安贾三清真灌汤包子馆的途径都有那些？

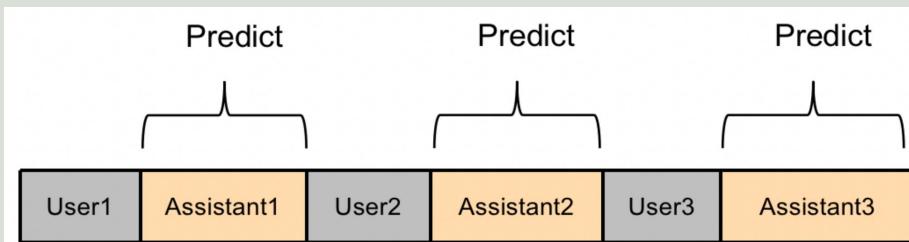
前往西安贾三清真灌汤包子馆应该乘坐什么交通工具？

通过数据增强技术，扩充了数据集，增加了模型的**泛化能力**，减少**模型幻觉**。

```
[  
  {  
    "instruction": "我想制定一个适合四人同学一起的西安4日游玩路线，可以帮我吗？",  
    "input": "",  
    "output": "当然可以！在西安旅游四天的行程中，你可以安排第一天参观兵马俑，第二天游览华清池和大雁塔，第三天去钟楼和鼓楼，第四天去秦始皇陵。建议你选择舒适的交通工具，注意防晒，准备好足够的水和食物。",  
  },  
  {  
    "instruction": "根据你制定的攻略，我在出行时需要注意什么？",  
    "input": "",  
    "output": "在出行时需要注意西安的气候变化，准备好合适的衣物；要注意安全，保护好个人财物；在景点游览时要遵守相关规定，不要乱扔垃圾。",  
  },  
  {  
    "instruction": "有推荐的美食店或美食吗？",  
    "input": "",  
    "output": "在西安可以尝试当地的特色美食，比如羊肉泡馍、肉夹馍、凉皮等。推荐的美食店有老孙家羊肉泡馍、回民街等。",  
  }  
]
```

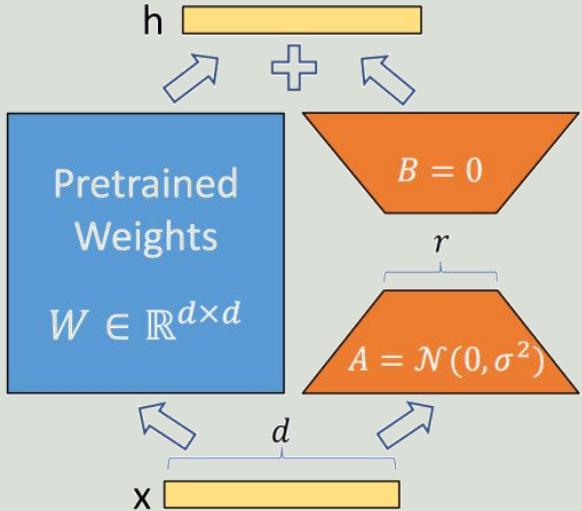


```
inputs = <user1> <assistant1> <user2> <assistant2> <user3> <assistant3>  
labels = <-100> <assistant1> <-100> <assistant2> <-100> <assistant3>
```



Firefly方法构造多轮数据
充分利用**所有**回复信息，也
不存在重复计算，非常高效。

LoRA微调Qwen

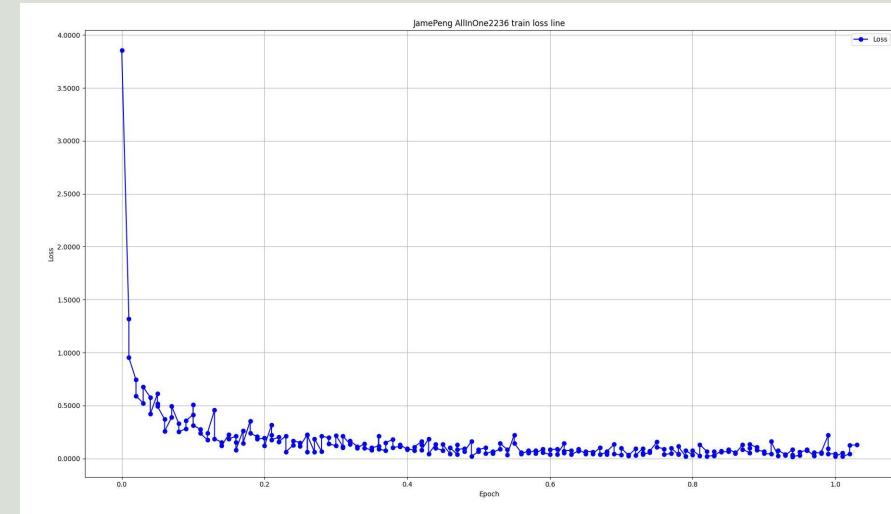


```
python finetune.py \
--model_name_or_path $MODEL \
--data_path $DATA \
--bf16 True \
--output_dir output_qwen \
--num_train_epochs 5 \
--per_device_train_batch_size 2 \
--per_device_eval_batch_size 1 \
--gradient_accumulation_steps 8 \
--evaluation_strategy "no" \
--save_strategy "steps" \
--save_steps 1000 \
--save_total_limit 10 \
--learning_rate 3e-4 \
--weight_decay 0.1 \
--adam_beta2 0.95 \
--warmup_ratio 0.01 \
--lr_scheduler_type "cosine" \
--logging_steps 1 \
--report_to "none" \
--model_max_length 1000 \
--gradient_checkpointing \
--use_lora
```

LoRA大模型微调技术

微调参数

微调效果



产品效果对比



通过数据分析给我御品轩的平均消费金额

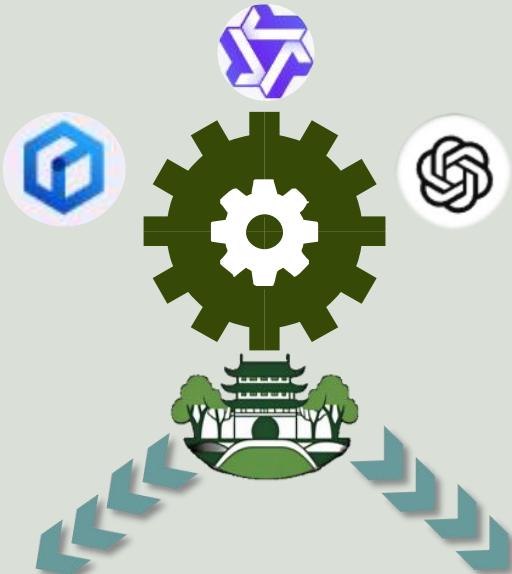
为了提供准确的数据，我需要访问最近一段时间御品轩的具体销售数据或者查阅相关数据库。如果能提供具体日期范围或数据量，我可以帮你计算平均消费金额。请告诉我时间段是多久，一般是一周、一个月还是全年数据？如果没有具体信息，只能给出估算。

例如（假设你提供了月度数据）：
“御品轩2022年1月份至12月份的整体平均消费金额为xx元/次”

给我提供一下最新的华山游玩的数据时间

很遗憾，我无法提供最新的华山游玩数据时间。由于旅游数据和政策会随时间变化，为了获取最准确的信息，建议您直接访问华山景区的官方网站、关注相关旅游平台或联系景区客服。这样，您可以获取到最新的开放时间、门票价格、游客接待量以及其他相关数据。同时，也请留意景区可能因天气、维护等原因而调整开放时间或发布临时通知，以确保您的行程安排得当。

重新生成



御品轩(高新店)在顾客眼中的评分是多少？

4.6/5分

通过数据分析给我御品轩的平均消费金额

给我提供一下最新的华山游玩的开放时间

平均消费金额：34元

你
御品轩(高新店)在顾客眼中的评分是多少？

ChatGPT

很抱歉，我无法提供当前时间的实时信息，包括特定餐厅的顾客评分。你可以通过访问餐饮评价平台或者搜索引擎，查找御品轩（高新店）的评价和评分。这些平台通常会提供用户对餐厅的评价和打分，以及文字评论，可以帮助你更好地了解餐厅的口碑和服务质量。

我想和我的闺蜜一起去西安游玩4天3夜，给我游玩攻略

为你和闺蜜设计的西安游玩攻略如下：

第一天：抵达西安后，24小时接站，送住酒店，全天自由活动。

第二天：参观唐华清宫、秦始皇陵兵马俑，晚上去大唐不夜城。

第三天：登西岳华山，体验奇险山的壮丽景色。

第四天：游览明城墙、大雁塔广场、钟鼓楼广场，最后品尝回民街美食。

这个行程涵盖了西安的历史文化景点和自然风光，适合与闺蜜一起畅游。

功能创新：

个性化生成旅游攻略
实时扩充知识库
大数据情感分析



收集大量文旅数据，通过AIGC技术实现个性化旅游攻略的生成



LangChain框架构建可以实时扩充的知识库，Agent应用落地



通过大数据情感分析，得出具体餐厅、酒店较客观、更全面的评价。



技术创新：

改进self-instruct技术构造语料库
自创Prompt模版编写Prompt语句
对语料库数据进行数据增强



使用ChatGPT3.5 API自动构建语料库，大大节省人工成本。

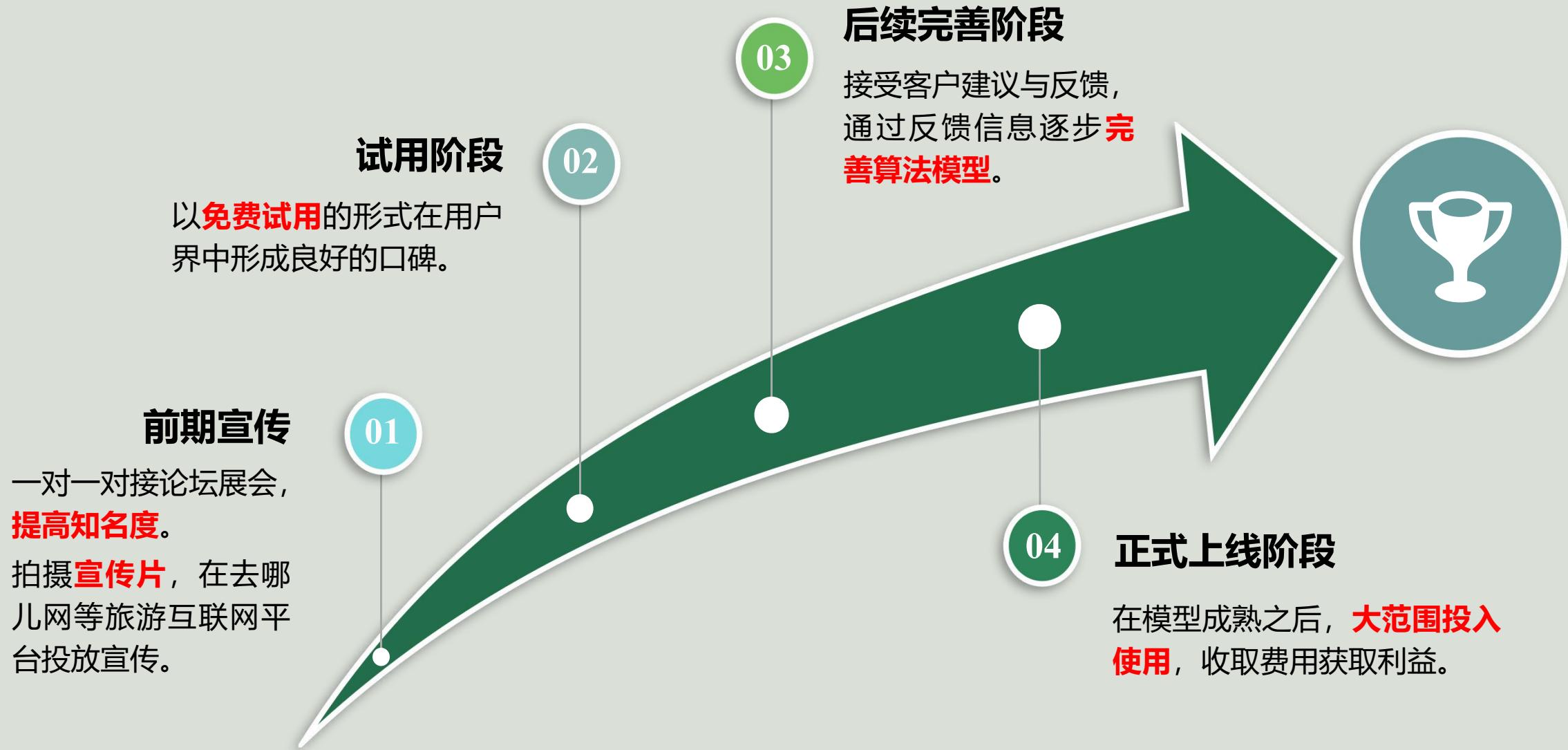


自研Prompt设计原则，自建Prompt模版，提高语料生成质量



通过数据增强，减少模型幻觉







INFJ 团队

詩詞會

大安 陌上

壯志西行追古踪
歲在甲申

狐烟大漠夕阳中
駕鈴古道絲綢路

駕鈴古道絲綢路
胡馬猶聞唐漢風

駕鈴古道絲綢路
胡馬猶聞唐漢風