



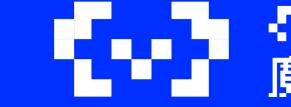
智谱·AI



marsoft | 安硕信息



北京交通大学
BEIJING JIAOTONG UNIVERSITY



ModelScope
魔搭社区



阿里云

SPM 2023

ChatGLM 金融大模型挑战赛

馒头科技



项目答辩说明（必填）

赛道	创意应用赛道/模型开发赛道
赛题	<i>ChatGLM</i> 金融大模型问答系统
项目进展程度	新项目/进展中项目

① 团队成员介绍

② 任务简介

③ 方案介绍

- 数据处理
- 大模型Prompt设计
- 微调方案

④ 总结

团队成员介绍

”

● **陈朝才**，毕业于中国科学技术大学，现就职于福建省厦门市 Manteia 数据科技有限公司。

- 在公司负责放疗产品线的 AI 相关研究，致力于开发先进的医学影像处理技术和放射治疗技术。公司开发的自动勾画产品已经成功应用于国内外上千家医院，目前，公司成功研发 Mozi 自适应放疗计划系统，实现了卡脖子产品的国产化，并已通过美国 FDA 和我国 NMPA 的认证。

● **宋智刚**，毕业于成都信息工程大学，现就职于江西金融发展集团旗下国家高新技术企业深圳京发科技控股公司。

- 公司致力于通过科技、大数据风控及与担保联动，打造“科技+风控+增信”一体化赋能服务能力，为实体场景和金融机构提供数字化服务，目前已服务个人客户超过 1500 万，资金交易总额超过 1300 亿，合作机构超过 200 家。

● **郑锦鹏**，毕业于广东外语外贸大学，目前担任广州市龙程电子有限公司的市场总监。

- 作为国内首批将 VR 技术与实体娱乐项目相结合的公司之一，我们在 VR 领域取得了一系列令人瞩目的创新成果。我们公司的产品在娱乐休闲场所得到广泛应用，而 VR 技术也在医疗、教育、培训、旅游、房地产等多个领域发挥着作用。我们以全方位的研发实力，愿意为不同行业的定制需求提供解决方案。我期待通过这次 AI 大模型比赛，进一步拓展创新思维，探索 AI+ 品牌 IP, AI+VR 的可能，为未来的市场营销和技术应用做技术铺垫。



任务简介

以ChatGLM2-6B模型为中心制作一个问答系统，回答用户的金融相关的问题

➤ 初级：数据基本查询（40分）

- 参赛者需要利用提供的ChatGLM2-6B开源模型和上市公司年报原始数据，并以此为基础创建信息问答系统。系统需能够解决基本查询，如：某公司2021年的研发费用是多少？等问题。

➤ 中级：数据统计分析查询（30分）

- 在初级阶段的基础上，参赛者需要进行金融数据的统计分析和关联指标查询。系统需基于各类指标，提供问题和答案，如：某公司2021年研发费用增长率为多少？等问题。

➤ 高级：开放性问题（30分）

- 如：某公司2021年主要研发项目是否涉及国家创新领域，如新能源技术、人工智能等？

任务简介 - 评测指标

”

$$\begin{cases} \max_{\text{similar}} (\text{sentence1}, \text{sentence2}, \text{sentence3}), & \text{无基础信息及关键词} \\ 0.25 + 0.25 + \max_{\text{similar}} (\text{sentence1}, \text{sentence2}, \text{sentence3}) * 0.5, & \text{基础信息正确, 关键词正确} \\ 0.25 + 0 + \max_{\text{similar}} (\text{sentence1}, \text{sentence2}, \text{sentence3}) * 0.5, & \text{基础信息正确, 关键词错误} \\ 0, & \text{基础信息错误} \end{cases}$$

任务简介 - 评测指标

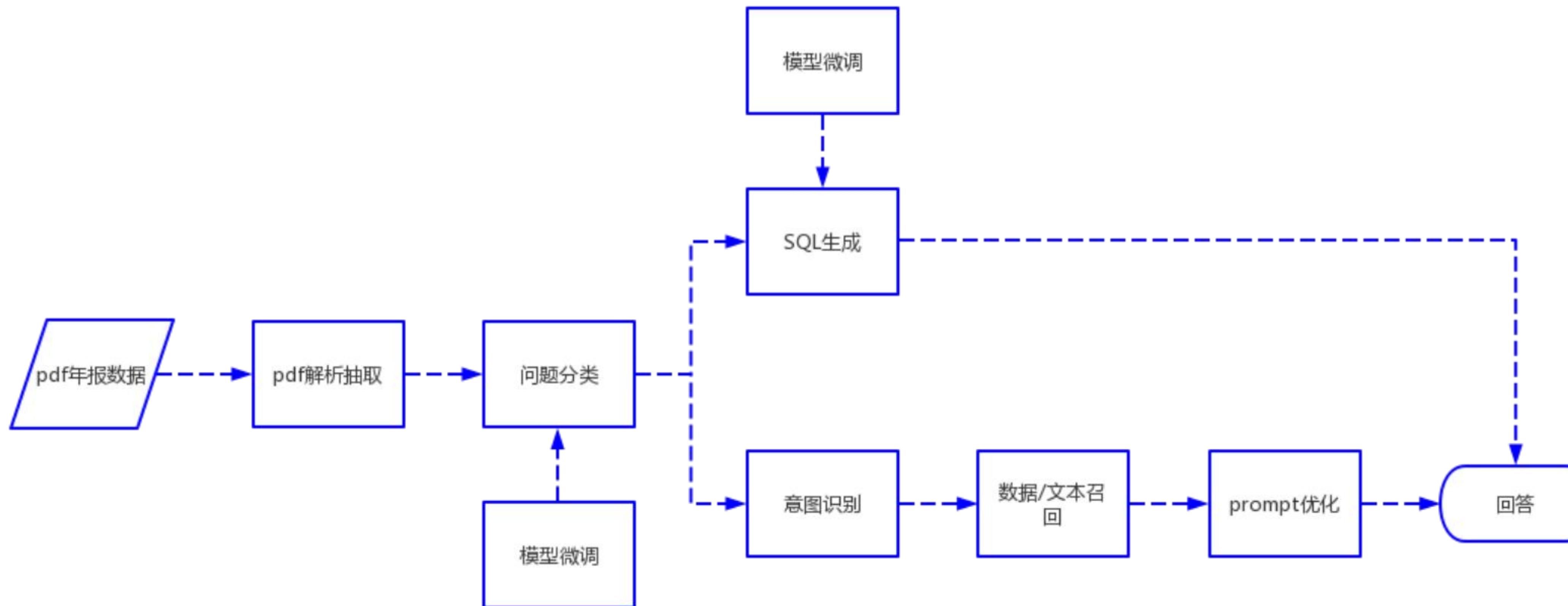
➤ **问题:** 2019年中国工商银行财务费用是多少元?

- **关键词:** 财务费用、2019
- **prompt:** "财务费用": "12345678.9元"
- **答案:** 12345678.9元
- **示例答案:** ["2019年中国工商银行财务费用是12345678.9元。", "2019年工商银行财务费用是12345678.9元。", "中国工商银行2019年的财务费用是12345678.9元。"]

➤ **评测计算示例:**

- **答案一:** 工商银行2019年财务费用是12345678.9元。
- **most similar sentences:**
 - ✓ 2019年工商银行财务费用是12345678.9元。 (Score: 0.9915)
 - ✓ 中国工商银行2019年的财务费用是12345678.9元。 (Score: 0.9820)
 - ✓ 2019年中国工商银行财务费用是12345678.9元。 (Score: 0.9720)
- **评分:** $0.25+0.25+0.9915*0.5=0.9958$ 分。
- **评分解释:** prompt_answer正确、包含所有key_word、相似度最高0.9915。

方案介绍 - 整体流程



PDF解析与信息抽取

解析与抽取步骤：

■ pdf文本抽取

- 官方提供的脚本

■ 页面召回

- 根据报表名称设置关键词找到对应的页

■ 表格识别

- camelot-py基于图像识别的表格提取

■ 信息过滤

- 非合并报表

- 调整报表

- 母公司报表

```
row_keywords = ['在职员工', '职工人数', '专业构成', '离退休职',  
    '生产人员', '销售人员', '技术人员',  
    '行政人员', '管理人员', '业务人员',  
    '教育程度', '硕士', '本科', '大专', '研究生',  
    '专科']
```

```
ei_tables = extract_table_for_rows(key, int(1.5*len(row_keywords)),  
    min_match_number=0.3*len(row_keywords),  
    required_line_keywords=['员工情况', '员工的数量', '员工数量', '专业构成',  
    '离退休职工人数', '员工教育结构'],  
    invalid_line_keywords=[],  
    required_post_keywords=row_keywords,  
    invalid_pre_keywords=[],  
    invalid_post_keywords=[],  
    prefix_pages=1, post_pages=1)
```

```
invalid_keywords = ['其他单位', '董事', '经理', '审议']  
filtered_tables = filter_tables(ei_tables, invalid_keywords)
```

PDF解析与信息抽取

1、合并资产负债表

编制单位：上海安硕信息技术股份有限公司

86

[View Details](#) | [Edit](#) | [Delete](#)

86

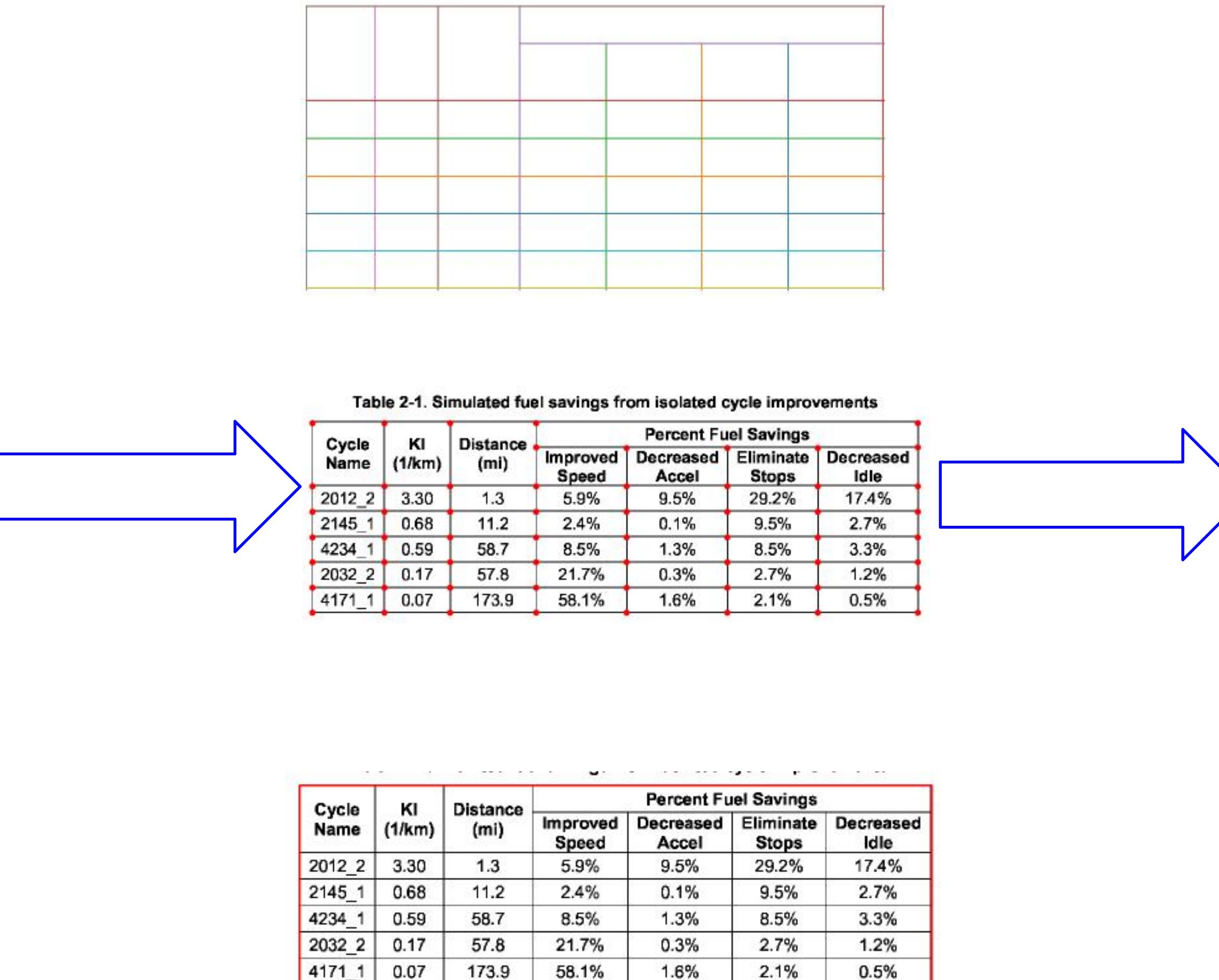
项目	2021年12月31日	2020年12月31日
流动资产：		
货币资金	125,060,806.04	212,466,694.89
结算备付金		
拆出资金		
交易性金融资产	80,000,000.00	172,500,000.00

3、合并利润表

单位：元

项目	2021 年度	2020 年度
一、营业总收入	755,179,963.25	660,948,843.66
其中：营业收入	755,179,963.25	660,948,843.66
利息收入		
已赚保费		
手续费及佣金收入		
二、营业总成本	735,370,091.25	594,534,618.83
其中：营业成本	499,503,404.13	392,458,263.98
利息支出		
手续费及佣金支出		
退保金		

92

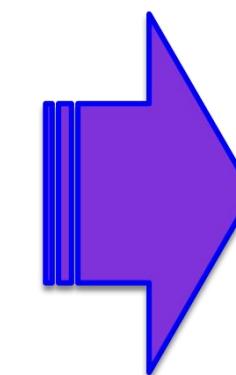


```
"cbs_info": [
    "page|87\n",
    "项目|2021 年 12 月 31 日|2020 年 12 月 31 日\n",
    "流动资产: ||\n",
    "货币资金|125,060,806.04|212,466,694.89\n",
    "结算备付金||\n",
    "拆出资金||\n",
    "交易性金融资产|80,000,000.00|172,500,000.00\n",
    "衍生金融资产||\n",
    "应收票据|172,710.00|390,257.00\n",
    "应收账款|72,293,737.60|68,592,443.21\n",
    "应收款项融资||\n",
    "预付款项|346,895.66|437,577.51\n",
```

问题分类

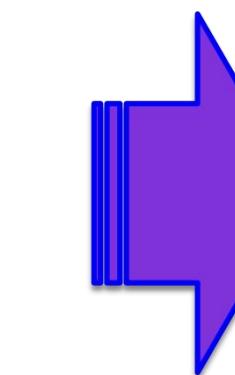
基于规则

- ✓ 是否包含公司名称 /年份
- ✓ 能否匹配到报表字段
- ✓ 能否匹配到公式
- ✓ 特殊词汇



大模型Prompt

请问“{}”是属于下面哪个类别的问题?
A: 基本信息查询,例如证券信息、股票简称、股票代码、外文名称、法定代表人、注册地址、办公地址、公司网址、电子邮箱等
B: 公司员工人数统计,例如员工人数、员工专业、员工教育程度等
C: 财务相关,例如金额、费用、资产、收入等
D: 以上都不是
例如:
1. XXXX的费用收入是多少元?
输出: C
2. XX公司法定代表人是谁?
输出: A
3. 请简要介绍分析XX公司的XXX情况。
输出: D
4. XX公司硕士人数是什么?
输出: B
你只需要回答编号,不要回答其他内容.



P-Tuning

PRE_SEQ_LEN=512
LR=2e-2
max_source_length 512
max_target_length 128

问题分类 - 模型微调

字母分类编号	分类选项名称	分类描述
A	公司基本信息	可以通过检索【公司基本信息表】来获取结果
B	公司员工信息	可以通过检索【公司员工信息表】来获取结果
C	财务报表相关内容	可以通过检索【财务三大报表】来获取结果
D	计算题	需要根据问题类型检索计算因子来完成计算
E	统计题	需要根据问题类型分析条件检索来获取结果
F	开放性问题	根据问题关键词来检索全文相关匹配来回答问题

问题分类 - 模型微调

`prompt_classify_question = ""`

请问 "{}" 是属于下面哪个类别的问题?

- A: 公司基本信息,包含股票简称,公司名称,外文名称,法定代表人,注册地址,办公地址,公司网址网站,电子信箱等.
- B: 公司员工信息,包含员工人数,员工专业,员工类别,员工教育程度等.
- C: 财务报表相关内容,包含资产负债表,现金流量表,利润表中存在的字段,包括费用,资产,金额,收入等.
- D: 计算题,无法从年报中直接获得,需要根据计算公式获得,包括增长率,率,比率,比重,占比等.
- E: 统计题,需要从题目获取检索条件,在数据集/数据库中进行检索、过滤、排序后获得结果.
- F: 开放性问题,包括介绍情况,介绍方法,分析情况,分析影响,什么是XXX.

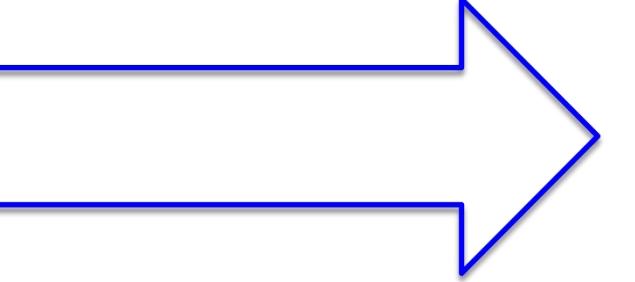
你只需要回答字母编号,不要回答字母编号及选项文本外的其他内容.

`".format(question_text)`

问题分类 - 模型微调

构建微调训练集（900+数量的标记数据）

```
[{"id": 0, "question_prompt": "\n    请问“能否根据2020年金宇生物技术股份有限公司的年报，给我简要介绍一下报告期内公司的社会责任工作情况？”， "question": "能否根据2020年金宇生物技术股份有限公司的年报，给我简要介绍一下报告期内公司的社会责任工作情况？", "query": "F"}, {"id": 1, "question_prompt": "\n    请问“请根据江化微2019年的年报，简要介绍报告期内公司主要销售客户的客户集中度情况，请根据江化微2019年的年报，简要介绍报告期内公司主要销售客户的客户集中度情况，并结合同行业情况进行分析”， "question": "请根据江化微2019年的年报，简要介绍报告期内公司主要销售客户的客户集中度情况，并结合同行业情况进行分析", "query": "F"}, {"id": 2, "question_prompt": "\n    请问“2019年四方科技电子信箱是什么？”是属于下面哪个类别的问题?\n    A: 公司基", "question": "2019年四方科技电子信箱是什么?", "query": "A"}, {"id": 3, "question_prompt": "\n    请问“研发费用对公司的技术创新和竞争优势有何影响？”是属于下面哪个类别的问题?\n    A: 公司基", "question": "研发费用对公司的技术创新和竞争优势有何影响?", "query": "F"}, {"id": 4, "question_prompt": "\n    请问“康希诺生物股份公司在2020年的资产负债比率具体是多少，需要保留至小数点后两位？”， "question": "康希诺生物股份公司在2020年的资产负债比率具体是多少，需要保留至小数点后两位?", "query": "D"}, {"id": 5, "question_prompt": "\n    请问“苏州科达在2020年的博士及以上人员数量是多少？”是属于下面哪个类别的问题?\n    A: 公司基", "question": "苏州科达在2020年的博士及以上人员数量是多少?", "query": "B"}, {"id": 6, "question_prompt": "\n    请问“河北金牛化工股份有限公司2020年法定代表人对比2019年是否相同？”是属于下面哪个", "question": "河北金牛化工股份有限公司2020年法定代表人对比2019年是否相同?", "query": "D"}]
```



模型微调参数

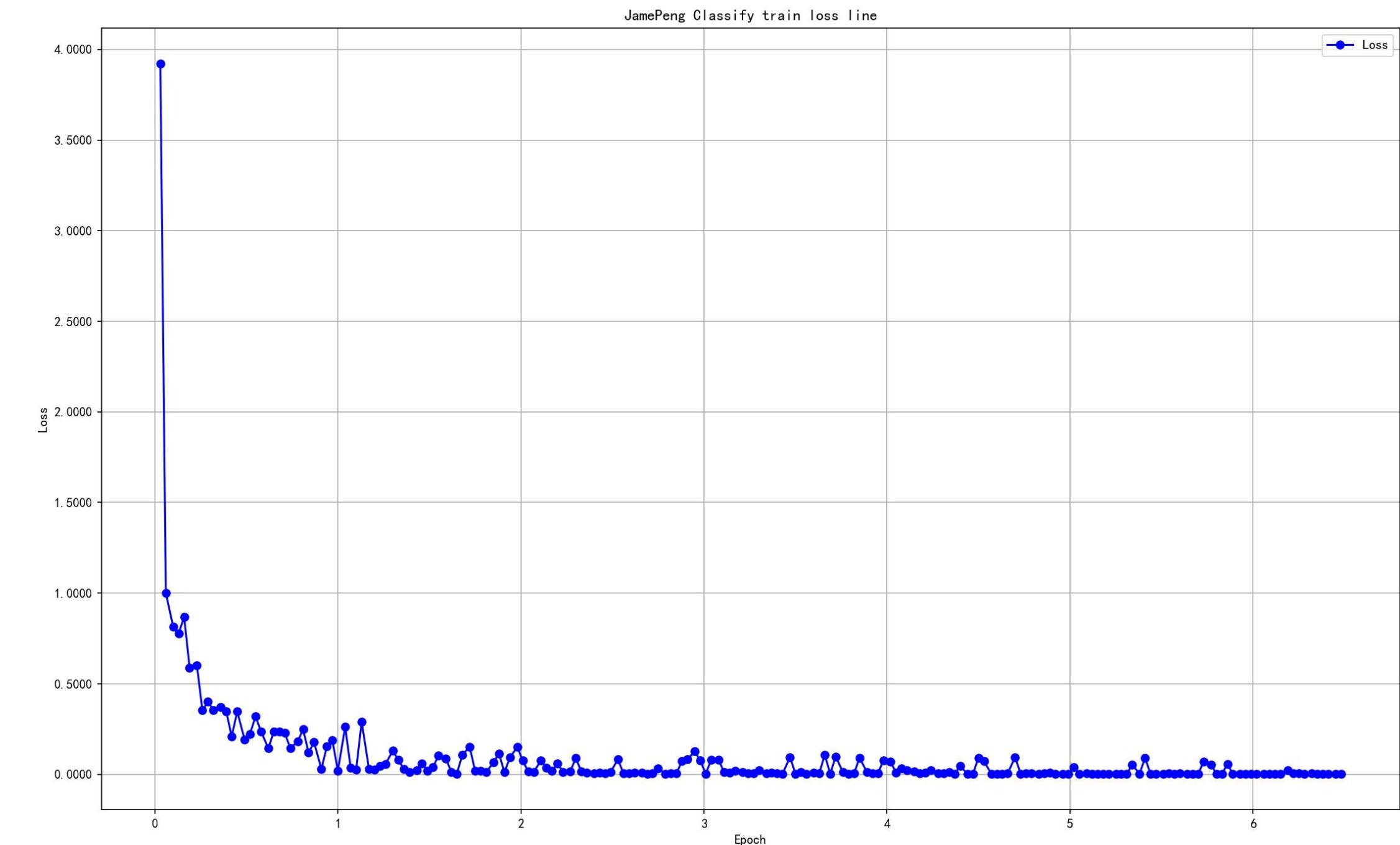
```
PRE_SEQ_LEN=512
LR=2e-2
RUNNING_STEP=400
NUM_GPUS=1
CUDA_VISIBLE_DEVICES=1 torchrun --standalone --nnodes=1 --nproc-per-node=$NUM_GPUS main.py \
--do_train \
--train_file Fin_train/train.json \
--validation_file Fin_train/dev.json \
--preprocessing_num_workers 10 \
--prompt_column question_prompt \
--response_column query \
--overwrite_cache \
--model_name_or_path /home/jamepeng/git_projects/chatglm2-6b-model \
--output_dir output/JamePeng_Ptuning_Classify/Fin-Train-chatglm2-6b-pt-$PRE_SEQ_LEN-$LR \
--overwrite_output_dir \
--max_source_length 512 \
--max_target_length 128 \
--per_device_train_batch_size 1 \
--per_device_eval_batch_size 1 \
--gradient_accumulation_steps 16 \
--predict_with_generate \
--max_steps $RUNNING_STEP \
--logging_steps 2 \
--save_steps $RUNNING_STEP \
--learning_rate $LR \
--pre_seq_len $PRE_SEQ_LEN \
```

问题分类 - 模型微调

模型微调：ChatGLM2对选择题有一定的理解能力，默认有一定的分类能力，通过微调来增强分类能力

- 最早期的一次微调，一开始的效果比较糟糕，最后发现是训练字段小于传入的模版+问题的字符串，导致训练过程中的传入字段丢失截断。loss一直在0.20下不去，距离第一轮训练目标0.10以下还有很大差距，实用情况比较糟糕。
- 通过统计问题和答案的长度，最后调整了max_source_length参数符合max(问题+答案)的条件，确认了max_source_length为512，max_target_length为128，使用ptuning对chatglm2-6b进行微调训练，step为250，训练结果的loss到了0.09，基本可用级别。
- 为了获取更好的效果，尝试对prompt不断进行调整，从一开始的只有选项的题目，到有一定信息的选项描述（高频的关键字等），同时对训练step加大到400，对微调训练有很大的帮助，loss的下降比较明显，最优平均值达到0.01以下，拥有比较优秀的分类能力。
- 尝试调低Learning rate可以获得更快的训练收敛效益。

```
[{"epoch": 6.28, "learning_rate": 0.0006, "loss": 0.0025, "step": 388}, {"epoch": 6.32, "learning_rate": 0.0005, "loss": 0.0047, "step": 390}, {"epoch": 6.35, "learning_rate": 0.0004, "loss": 0.0022, "step": 392}, {"epoch": 6.38, "learning_rate": 0.0003, "loss": 0.0032, "step": 394}, {"epoch": 6.41, "learning_rate": 0.0002, "loss": 0.0024, "step": 396}, {"epoch": 6.45, "learning_rate": 0.0001, "loss": 0.0028, "step": 398}, {"epoch": 6.48, "learning_rate": 0.0, "loss": 0.0023, "step": 400}]
```



问题分类 - 结果示例

山东玻纤集团股份有限公司2020年的资产负债比率保留两位小数是多少?



D

请提供2021年南通江天化学股份有限公司其他流动资产的详细数据。



C

请具体描述一下2019年景旺电子主要控股参股公司分析的情况。



F

2019年负债总金额第7高的上市公司为?



E

通威股份有限公司2019年的外文名称为?



A

SQL生成

*ChatGLM2*本身有一定的*NL2SQL*能力，但是难以在项目中稳定输出有效正确的*SQL*格式的语句。

线上的*GLM-130B*有良好的*NL2SQL*的性能，所以我们判断应该可以透过微调训练来增强*ChatGLM2-6B*的*SQL*能力

请根据以下用户输入，输出sql代码。\\n用户输入：“2019年哪家公司的总负债最高？”

用户的输入是：“2019 年哪家公司的总负债最高？”为了回答这个问题，我们需要在数据库中查找 2019 年各公司的总负债信息，并找出负债最高的那家公司。以下是对应的 SQL 代码：

```
sql
1 SELECT company_name, total_debt
2 FROM companies
3 WHERE year = 2019
4 ORDER BY total_debt DESC
5 LIMIT 1;
```

在此代码中，`companies` 是包含公司名称和总负债信息的表，`year` 是表示负债年度的字段，`total_debt` 是表示总负债的字段。代码首先筛选出 2019 年的数据，然后按总负债降序排列，最后取出排在前面的第一家公司的信息。

以上内容为 AI 生成，不代表开发者立场，请勿删除或修改本标记

Zero-shot *GLM-130B*的*NL2SQL*结果



SQL生成

尝试一：

早期尝试用公共的NL2SQL训练集（2019年NL2SQL大赛的公开标注数据集），大概40000多道题目（包含电影、书籍、评价、金融等等五花八门的题目类型）。结果发现，不仅没有能够很好对题目进行SQL生成（loss在0.36-0.50摇摆下不去），而且丧失了基本的对话能力，自认为是题目里出现过的张三【后续重点】。

尝试二：

再次尝试公共NL2SQL训练集，这次加上了比赛题目的题型，希望强化SQL后可以对比赛题目也能有比较好的表现。训练后发现，小规模的标注60-100道题难以对原来较大规模的题目造成影响，而且发现原来的训练集虽然大，但其实也是有很多问题会对本次项目有负面影响，最后放弃掉公共的NL2SQL的训练集。他山之石，难以攻玉！



SQL生成 - 训练集生成

根据前面两个尝试结果，我们认为要想稳定生产准确有效SQL语言，是需要根据实际的任务情况和方向来设定训练目标，最后决定自己来标注训练集，为了加快效率，使用*ChatGPT/Bard*来协助生成该写出不同提问模板，再透过字段的随机条件填充。

构建针对其模版格式的问题训练集，根据比赛早期提供的四个模版扩充了24种不同问法，并且强化了除了查询以外的SQL语句，如排序，输出范围，统计计数，条件求和，单字段检索，多字段检索，多字段检索多字段，字段的过滤等，不同比较级的语言描述（高低，大小，多少，否定词+理想，中文数字和阿拉伯数字）。【为了应对无日志调优情况下，SQL语句广谱性挑战】

第一轮扩充到了200多道个人校对的高质量NL2SQL训练集。C榜前扩充到了1000多道NL2SQL训练集。

SQL生成 - 训练集生成

数据构造

构建
问答模板

```
{"question": "[年份]年哪家公司的[指标]最高?", "answer": "select 公司全称 from company_table where 年份 = '[年份]' and [指标] is not null order by [指标] desc limit 1"}
```

变量
随机填充

```
{"question": "2022年哪家公司的购建固定资产最高?", "answer": "select 公司全称 from company_table where 年份 = '2022' and 购建固定资产 is not null order by 购建固定资产 desc limit 1"}
```

大模型
问题改写

```
{"question": "2022年哪家公司的购建固定资产最高?", "answer": "select 公司全称 from company_table where 年份 = '2022' and 购建固定资产 is not null order by 购建固定资产 desc limit 1"} {"question": "2022年哪家公司购建固定资产的资金投入最多?", "answer": "select 公司全称 from company_table where 年份 = '2022' and 购建固定资产 is not null order by 购建固定资产 desc limit 1"} {"question": "在2022年，哪家公司的固定资产购建支出居于首位?", "answer": "select 公司全称 from company_table where 年份 = '2022' and 购建固定资产 is not null order by 购建固定资产 desc limit 1"}
```

...

Prompt替换

你是一名**Mysql数据库开发人员**，你精通Mysql数据库的sql代码编写，你需要根据已知的表名、字段名和用户输入的问题编写sql代码

已知表名：

company_table**已知字段名**: [公司全称、年份、经营活动现金流入小计、公司的中文简称、固定资产...]
注意对问题中的中文数字 (xx亿、xx千万、xx万)

进行阿拉伯**数字转换**，如：
一个亿、一亿需转换为 100000000，一千万需转换为 10000000 要求sql代码中的字段名必须是已知字段名，不得新增字段名**示例模板**: """"用户输入：2019年哪家公司的负债合计最高?
sql如下：```sql select 公司全称 from company_table order by 负债合计 desc limit 1```""""

请根据以下用户输入，**输出sql代码**。用户输入： {}

P-Tuning

PRE_SEQ_LEN=128
LR=2e-2
max_source_length
2200
max_target_length 300
quantization_bit 8

NL2SQL- 结果示例

1 2019年负债总金额第7高的上市公司为?

- SELECT 公司全称 FROM company_table WHERE 年份 = '2019' AND 负债合计 IS NOT NULL ORDER BY 负债合计 DESC LIMIT 1 OFFSET 6

2 2020年其他流动资产最高的前三家上市公司是哪些家?

- SELECT 公司全称 FROM company_table WHERE 年份 = '2020' AND 其他流动资产 IS NOT NULL ORDER BY 其他流动资产 DESC LIMIT 3

3 2020年营业收入最高的7家并且曾经在武汉注册的上市公司是? 金额是?

- SELECT 公司全称, 营业总收入 FROM company_table WHERE 年份 = '2020' AND 注册地址 LIKE '%武汉%' AND 营业总收入 IS NOT NULL ORDER BY 营业总收入 DESC LIMIT 7

4 哪家上市公司2019年总负债最低?

- SELECT 公司全称 FROM company_table WHERE 年份 = '2019' AND 负债合计 IS NOT NULL ORDER BY 负债合计 ASC LIMIT 1

5 哪家在上海注册的上市公司, 2020年营业收入最高? 金额是?

- SELECT 公司全称, 营业收入 FROM company_table WHERE 年份 = '2020' AND 注册地址 LIKE '%上海%' AND 营业收入 IS NOT NULL ORDER BY 营业收入 DESC LIMIT 1

意图识别 - Prompt提取

Prompt 1

对于文本“{}”,其中重要的关键词是什么?你只需要回答关键词,不要回答其他内容.

Prompt 2

- 你的任务是做信息抽取,从下面的问题中提取出最重要的几个关键词:{}
- 注意:
关键词之间用'、'分隔
关键词不要重复
关键词数量不超过3个

意图识别 - Prompt提取

角色和目标

这是文字提取器，你要从用户输入的文本中提取关键词

任务具体要求

关键词是指：问题最终指向的词语，通常是名词或句子的宾语，通常出现在公司名称或时间状语后面

如：净利润、社会责任工作、企业名称、固定资产、外文名称、注册地址、财务费用、长期借款、短期借款、资产及负债、收回投资收到的现金、净利润率、企业研发经费与利润比值、企业研发经费与营业收入比值、研发人员占职工人数比例、企业硕士及以上人员占职工人数比例、企业研发经费占费用比例、收回投资所收到的现金、关键审计事项、法人代表、负债总金额、总负债、无。
对象可以有多个。没有写“无”。

输出完毕后结束，不要生成新的用户输入，不要新增内容

示例模板：

"""

用户输入：能否根据的年报，给我简要介绍一下报告期内公司的社会责任工作情况？

关键词1:社会责任工作

用户输入：其他非流动金融资产第十一高的上市公司是哪家？

关键词1:其他非流动金融资产

用户输入：研发人员占职工人数比例是多少?请。？

关键词1:研发人员占职工人数比例

用户输入：办公地址是什么？

关键词1:办公地址

用户输入：法定代表人对比是否相同？

关键词1:法定代表人

用户输入：根据的年报，请简要介绍报告期内公司主要销售客户的客户集中度如何？请结合同行业情况简要分析。？

关键词1:客户集中度

用户输入：在的现金比率是多少？请保留至小数点后两位。

关键词1:现金比率

用户输入：研发费用和财务费用分别是多少元？

关键词1:研发费用

关键词2:财务费用

"""

参考案例

输出要求

请根据以下文本，严格按照示例模板格式输出内容。

用户输入

用户输入：{}

意图识别 - 模型微调

01

数据集构造：模型输出 + 人工修正

```
{  
    "id": 0,  
    "question_prompt": "\n        请帮我从以下句子中提取关键词。这些关键词是句子中最重要、最能概括句子主题的词汇。通过这些关键词，你可以更好地理解句子的内容。你只需要回答文本中的关键词，不要回答其他内容.\n        用户输入：\n        \"能否根据2020年金宇生物技术股份有限公司的年报，给我简要介绍一下报告期内公司的社会责任工作情况？\"",  
    "question": "能否根据2020年金宇生物技术股份有限公司的年报，给我简要介绍一下报告期内公司的社会责任工作情况？",  
    "query": "社会责任工作情况"  
},  
{  
    "id": 1,  
    "question_prompt": "\n        请帮我从以下句子中提取关键词。这些关键词是句子中最重要、最能概括句子主题的词汇。通过这些关键词，你可以更好地理解句子的内容。你只需要回答文本中的关键词，不要回答其他内容.\n        用户输入：\n        \"请根据江化微2019年的年报，简要介绍报告期内公司主要销售客户的客户集中度情况，并结合同行业情况进行分析。\"",  
    "question": "请根据江化微2019年的年报，简要介绍报告期内公司主要销售客户的客户集中度情况，并结合同行业情况进行分析。",  
    "query": "主要销售客户集中度情况"  
},  
{  
    "id": 2,  
    "question_prompt": "\n        请帮我从以下句子中提取关键词。这些关键词是句子中最重要、最能概括句子主题的词汇。通过这些关键词，你可以更好地理解句子的内容。你只需要回答文本中的关键词，不要回答其他内容.\n        用户输入：\n        \"2019年四方科技电子信箱是什么？\"",  
    "question": "2019年四方科技电子信箱是什么？",  
    "query": "电子信箱"  
},  
{  
    "id": 3,  
    "question_prompt": "\n        请帮我从以下句子中提取关键词。这些关键词是句子中最重要、最能概括句子主题的词汇。通过这些关键词，你可以更好地理解句子的内容。你只需要回答文本中的关键词，不要回答其他内容.\n        用户输入：\n        \"研发费用对公司的技术创新和竞争优势有何影响？\"",  
    "question": "研发费用对公司的技术创新和竞争优势有何影响？",  
    "query": "研发费用,技术创新,竞争优势,影响"  
},  
{  
    "id": 4,  
    "question_prompt": "\n        请帮我从以下句子中提取关键词。这些关键词是句子中最重要、最能概括句子主题的词汇。通过这些关键词，你可以更好地理解句子的内容。你只需要回答文本中的关键词，不要回答其他内容.\n        用户输入：\n        \"康希诺生物股份公司在2020年的资产负债比率具体是多少，需要保留至小数点后两位？\"",  
    "question": "康希诺生物股份公司在2020年的资产负债比率具体是多少，需要保留至小数点后两位？",  
    "query": "资产负债比率"}
```

02

模型Prompt:

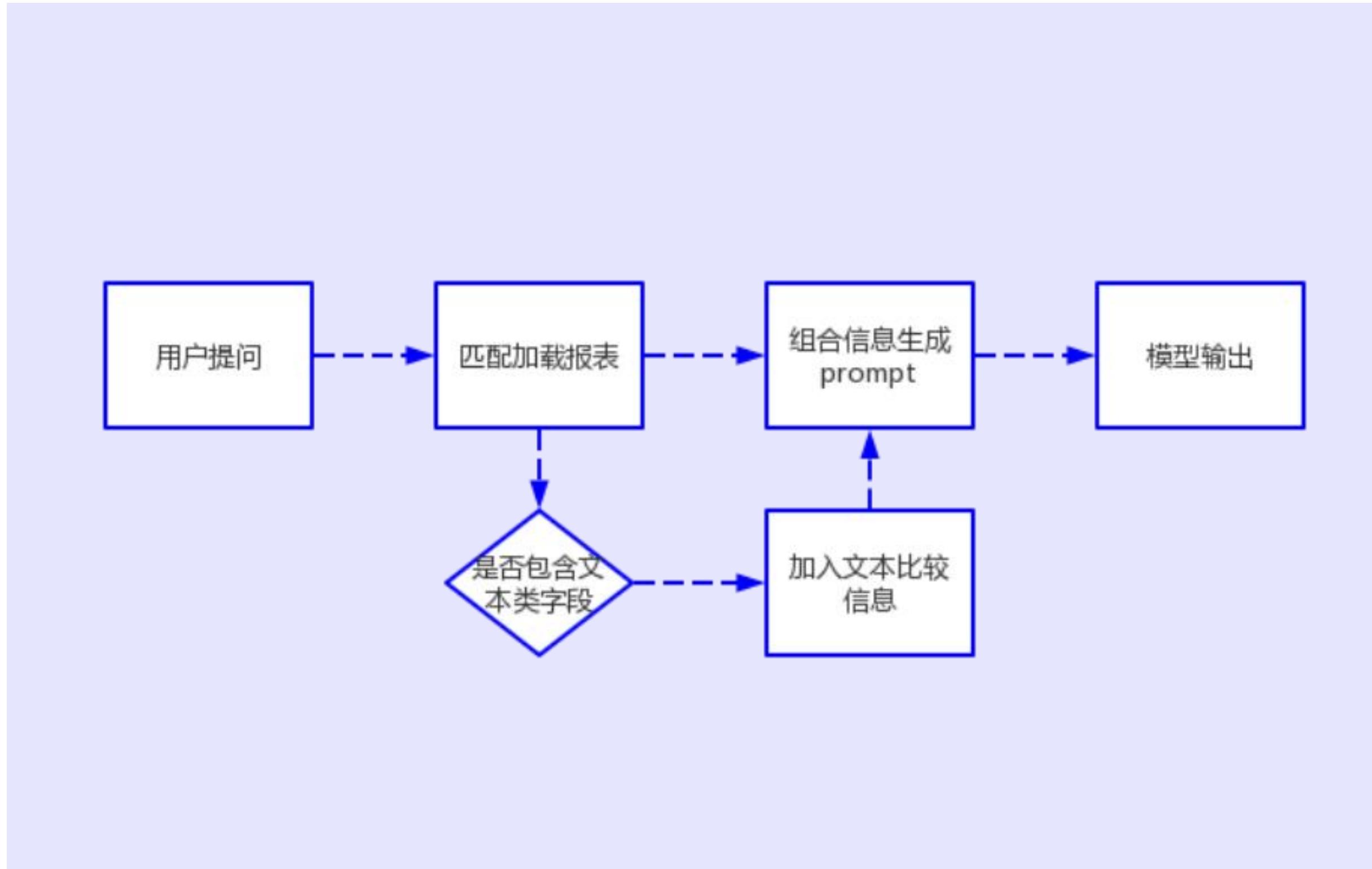
请帮我从以下句子中提取关键词。这些关键词是句子中最重要、最能概括句子主题的词汇。通过这些关键词，你可以更好地理解句子的内容。你只需要回答文本中的关键词，不要回答其他内容。

用户输入：

意图识别 - 结果示例

- ① Question: 请告诉我山东华鲁恒升化工股份有限公司2020年的技术人员人数的具体数值
 - Keywords: 技术人员人数
- ② Question: 请简要介绍2019年成都银河磁体股份有限公司面临退市情况。
 - Keywords: 面临退市情况
- ③ Question: 2020年湖南九典制药股份有限公司应收款项融资是多少元?
 - Keywords: 应收款项融资
- ④ Question: 在2020年的时候，厦工股份流动负债比率为多少?
 - Keywords: 流动负债比率
- ⑤ Question: 2020年广西博世科环保科技股份有限公司投资收益和净利润分别是多少元?
 - Keywords: '投资收益', '净利润'

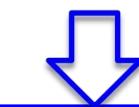
问题回答 - type1问题处理流程



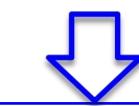
- 1 提取提问中的公司名称/年份信息
- 2 根据公司名称/年份/意图匹配报表表格字段
- 3 若匹配到文本类字段与多个年份，则加入文本比较信息
- 4 组合匹配到的字段，生成prompt
 - ✓ 文本类字段将字段值加双引号
 - ✓ 非文本类字段将字段名称加双引号
 - ✓ 人数类字段结果为XXX有XXX人，其他字段结果为XXX是XXX
 - ✓ 金额类字段结果为XXX是XXX元
- 5 prompt给模型产生输出

问题回答 - type1问题回答示例

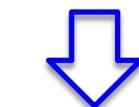
问：请提供2020年昊华化工科技股份有限公司研发人员人数的详细数据。



提取到的公司名称/年份以及意图词：昊华化工科技股份有限公司/2020/研发人员人数



匹配到的报表字段：研发人员数量占公司总人数的比例（%）/公司研发人员的数量



已知昊华化工科技股份有限公司（简称：昊华科技，证券代码：600378）2020年的资料如下：
“研发人员数量占公司总人数的比例（%）”是36.22人，“公司研发人员的数量”是2491人。

请回答问题：请提供2020年昊华化工科技股份有限公司研发人员人数的详细数据。

注意你的回答应该按照以下要求：

1. 你回答的格式应该是：2020年昊华化工科技股份有限公司的xxxx是xxxx。
2. 你只需要回答问题相关的内容，不要回答无关内容。
3. 你不需要进行计算。
4. 你的回答只能来源于提供的资料。



模型回答：2020年昊华化工科技股份有限公司的研发人员数量是2491人

问题回答 - type2问题处理流程

01

匹配问题中是否出现比值类关键字

02

找到比值计算公式

03

根据公式生成问题链

04

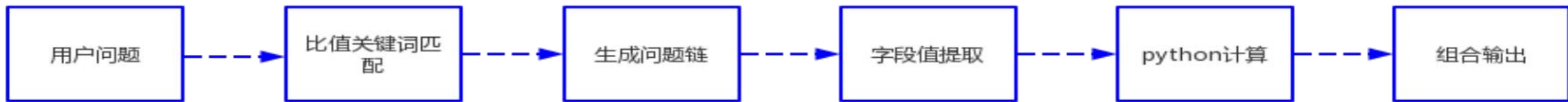
通过问题链回答解析公式中字段对应的值

05

通过python计算得到结果

06

组合问题链结果以及计算结果得到回答

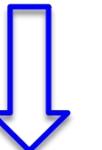


问题回答 - type2问题回答示例

用户问题: 在2019年的财务数据中, 易华录营业利润率是多少? 保留两位小数



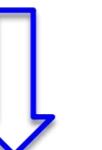
匹配计算公式: 营业利润率=营业利润/营业收入



问题链:
2019年易华录的营业利润是多少元?
2019年易华录的营业收入是多少元?



提取回答进行计算: $527509725.56 / 3743903561.80$



最终回答:
易华录2019年的营业利润是527509725.56元。易华录2019年的营业收入是3743903561.80元。根据公式, 营业利润率=营业利润/营业收入得出结果0.14(14.09%)

问题回答 - type2问题处理流程

01

识别提问中的
公司名称/年
份/意图关键
词

02

根据关键词和
问题分别召回
top3相关文
本

03

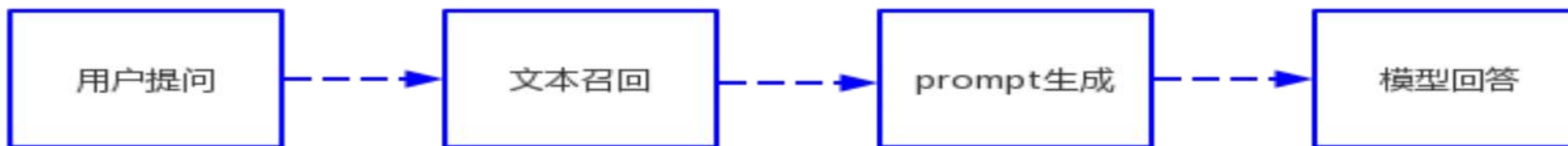
找到匹配长度
最大的文本

04

组合生成
prompt

05

模型输出



问题回答 - type3问题回答示例

你需要阅读理解年报的片段来真实详细完整的回答用户的提问。

下面是年报内容格式的一些说明：

1. 片段由标题和正文内容组成。
2. 片段中的标题通常以中文数字如一二三四五或阿拉伯数字12345开始。
3. “**√适用**”表示该项内容公司存在该事项，“**√不适用**”表示公司不存在该事项。
4. “**√是**”表示该项是或者有，“**√否**”表示该项不是或者没有。

*****弘元绿色能源股份有限公司2019年年报*****

-----片段:1-----

九、破产重整相关事项

√不适用

十、重大诉讼、仲裁事项

□本年度公司有重大诉讼、仲裁事项√本年度公司无重大诉讼、仲裁事项

十一、上市公司及其董事、监事、高级管理人员、控股股东、实际控制人、收购人处罚及整改情况

√不适用

十二、报告期内公司及其控股股东、实际控制人诚信状况的说明

√不适用

十三、公司股权激励计划、员工持股计划或其他员工激励措施的情况及其影响

(一)相关激励事项已在临时公告披露且后续实施无进展或变化的

√适用

事项概述 查询索引

(二)临时公告未披露或有后续进展的激励情况

股权激励情况

√不适用

其他说明

√不适用

员工持股计划情况

√不适用

其他激励措施

√不适用

十四、重大关联交易

(一)与日常经营相关的关联交易

1、已在临时公告披露且后续实施无进展或变化的事项

√不适用

2、已在临时公告披露，但有后续实施的进展或变化的事项

√不适用

问：根据弘元绿色能源股份有限公司2019年的年报数据，能否简要介绍公司报告期内破产重整相关事项的详情。

2023-09-16 16:24:59.733 | INFO | generate_answer_with_classify:generate_answer:371 - Answer length 46

2023-09-16 16:24:59.734 | INFO | generate_answer_with_classify:generate_answer:372 | 根据弘元绿色能源股份有限公司2019年的年报数据，公司在报告期内没有涉及到破产重整相关事项。

问题回答 - 一些测试结果

问：安硕信息在2021年的主要销售客户和主要供应商情况

你需要阅读理解年报的片段来真实详细完整的回答用户的提问。
下面是年报内容格式的一些说明：
1. 片段由标题和正文内容组成。
1. 片段中的标题通常以中文数字如一二三四五或阿拉伯数字12345开始。
2. 片段中的表格采用制表符‘’分隔。
3. “√适用”表示该项内容公司存在该事项，“√不适用”表示公司不存在该事项。
4. “√是”表示该项是或者有，“√否”表示该项不是或者没有。

*****安硕信息2021年年报*****

片段:1-----
...
(8) 主要销售客户和主要供应商情况
公司主要销售客户情况
前五名客户合计销售金额(元) 195,383,881.28
前五名客户合计销售金额占年度销售总额比例 25.87%
前五名客户销售额中关联方销售额占年度销售总额比例 0.00%
公司前5大客户资料
序号 客户名称 销售额(元) 占年度销售总额比例
1 客户一 62,974,407.57 8.34%
2 客户二 59,449,455.44 7.87%
3 客户三 35,764,528.30 4.74%
4 客户四 22,431,141.06 2.97%
5 客户五 14,764,348.91 1.96%
合计 195,383,881.28 25.87%
主要客户其他情况说明
√不适用
公司主要供应商情况
前五名供应商合计采购金额(元) 17,369,457.27
前五名供应商合计采购金额占年度采购总额比例 40.18%
前五名供应商采购额中关联方采购额占年度采购总额比例 0.00%
公司前5名供应商资料
序号 供应商名称 采购额(元) 占年度采购总额比例
1 供应商一 5,150,763.97 11.92%
2 供应商二 4,660,778.60 10.78%
3 供应商三 2,920,588.50 6.76%
4 供应商四 2,478,685.43 5.73%
5 供应商五 2,158,640.77 4.99%
合计 17,369,457.27 40.18%
主要供应商其他情况说明
√不适用
3、费用
单位：元
2021年 2020年 同比增减 重大变动说明
销售费用 31,447,117.7 326,601,491.38 18.22%
管理费用 67,947,047.6 759,258,169.56 14.66%
财务费用 1,659,870.6 0506,016.23 228.03% 主要系报告期内利息支出增加所致。
研发费用 129,007,518.3 2110,687,052.49 16.55%
4、研发投入
√适用
主要研发项目名称 项目目的 项目进展 拟达到的目标 预计对公司未来发展的影响
基础平台及A3Web应用框架研发 1、顺应云原生技术发展趋势，保证基础平台和应用框架技术先进性，持续迭代优化；2、在信创背景下，增强产品的国产化适配能力。完成新版本研发 1、满足信创认证（国产硬件、操作系统、数据库、浏览器）2、对A3应用系统开发提供全面支撑，提高程序质量和开发效率 1.完成信创认证，适配更多的国产软件，使其更加安全高效，有利于公司拓展金融客户。

2023-09-17 14:54:58.634 | INFO | generate_answer_with_classify:generate_answer:372 - 根据安硕信息2021年年报的片段，公司在2021年的主要销售客户包括前五名客户合计销售金额为195,383,881.28元，占年度销售总额比例为25.87%。前五名客户销售额中关联方销售额占年度销售总额比例为0.00%。公司前5大客户资料如下：

- 客户一：前5名客户合计销售金额为62,974,407.57元，占年度销售总额比例为8.34%。
- 客户二：前5名客户合计销售金额为59,449,455.44元，占年度销售总额比例为7.87%。
- 客户三：前5名客户合计销售金额为35,764,528.30元，占年度销售总额比例为4.74%。
- 客户四：前5名客户合计销售金额为22,431,141.06元，占年度销售总额比例为2.97%。
- 客户五：前5名客户合计销售金额为14,764,348.91元，占年度销售总额比例为1.96%。

此外，安硕信息在2021年的主要供应商包括前五名供应商合计采购金额为17,369,457.27元，占年度采购总额比例为40.18%。前五名供应商采购额中关联方采购额占年度采购总额比例为0.00%。公司前5名供应商资料如下：

- 供应商一：前5名供应商合计采购金额为5,150,763.97元，占年度采购总额比例为11.92%。
- 供应商二：前5名供应商合计采购金额为4,660,778.60元，占年度采购总额比例为10.78%。
- 供应商三：前5名供应商合计采购金额为2,920,588.50元，占年度采购总额比例为6.76%。
- 供应商四：前5名供应商合计采购金额为2,478,685.43元，占年度采购总额比例为5.73%。
- 供应商五：前5名供应商合计采购金额为2,158,640.77元，占年度采购总额比例为4.99%。

问题回答 - 一些测试结果

问：请介绍2021年安硕信息的审计意见

你需要阅读理解年报的片段来真实详细完整的回答用户的提问。
下面是年报内容格式的一些说明：
1. 片段由标题和正文内容组成。
2. 片段中的标题通常以中文数字如一二三四五或阿拉伯数字12345开始。
3. 片段中的表格采用制表符‘分隔。
4. “/适用”表示该项内容公司存在该事项，“/不适用”表示公司不存在该事项。
5. “/是”表示该项是或者有，“/否”表示该项不是或者没有。

*****安硕信息2021年年报*****

片段:1-----
审计意见类型 标准的无保留意见
审计报告签署日期 2022年04月26日
审计机构名称 大信会计师事务所(特殊普通合伙)
审计报告文号 大信审字[2022]第4-00399号
注册会计师姓名 张静娟、陈丽华
审计报告正文
上海安硕信息技术股份有限公司全体股东：
二、审计意见
我们审计了上海安硕信息技术股份有限公司（以下简称“贵公司”）的财务报表，包括2021年12月31日的合并及母公司资产负债表、2021年度的合并及母公司利润表、合并及母公司现金流量表、合并及母公司股东权益变动表，以及财务报表附注。我们认为，后附的财务报表在所有重大方面按照企业会计准则的规定编制，公允反映了贵公司2021年12月31日的合并及母公司财务状况以及2021年度的合并及母公司经营成果和现金流量。
三、形成审计意见的基础
我们按照中国注册会计师审计准则的规定执行了审计工作。审计报告的“注册会计师对财务报表审计的责任”部分进一步阐述了我们在这些准则下的责任。按照中国注册会计师职业道德守则，我们独立于贵公司，并履行了职业道德方面的其他责任。
我们相信，我们获取的审计证据是充分、适当的，为发表审计意见提供了基础。
三、关键审计事项
关键审计事项是我们根据职业判断，认为对本期财务报表审计最为重要的事项。这些事项的应对以对财务报表整体进行审计并形成审计意见为背景，我们不对这些事项单独发表意见。
(一) 应收账款坏账准备
1. 事项描述
关于应收账款预期信用损失的确定方法会计政策请参阅附注三、(十一)；关于应收账款账面余额及预期信用损失请参阅财务报表附注五、(四)。
截止2021年12月31日，贵公司财务报表所示应收账款账面余额为9,833.67万元，坏账准备为2,694.30万元，账面价值为7,129.37万元。账面价值占总资产的比例为8.75%。贵公司管理层在确定应收账款预计可收回金额时需要评估客户的信用情况，参照历史损失经验及对未来经济的预测。因预期信用损失涉及管理层运用重大会计估计和判断，我们将应收账款坏账准备做为关键审计事项。
2. 审计应对
我们针对应收账款坏账准备执行的主要审计程序包括：
(1) 对贵公司信用政策及应收账款管理相关内部控制的设计和运行有效性进行了评估和测试；
(2) 分析贵公司应收账款预期信用损失会计估计的合理性，包括确定应收账款组合的依据、金额重大的判断、单独计提坏账准备的判断等；
(3) 分析计算贵公司资产负债表日预期信用损失金额与应收账款余额之间的比率，并与同行业预期信用损失计提比例进行比较，分析应收账款预期信用损失计提是否充分；
(4) 获取贵公司预期信用损失计提表，检查计提方法是否按照预期信用损失的计提政策执行；重新计算预期信用损失计提金额是否准确；
(5) 通过分析贵公司应收账款的账龄、客户信誉情况和客户的历史回款情况，并执行应收账款函证程序及检查期后回款情况等，评价应收账款预期信用损失计提的合理性。
(二) 收入确认
1. 事项描述
贵公司主营业务为客户提供软件产品定制服务和技术开发服务等。关于收入确认政策请参阅财务报表附注三、(二十三)；关于营业收入发生额请参阅财务报表附注五、(三十三)。
由于收入是重要的财务指标之一，存在管理层为了达到特定目标或期望而操纵收入确认时点的固有风险，因此我们将收入确认识别为关键审计事项。
2. 审计应对
我们针对收入执行的相关审计程序如下：
(1) 对照《企业会计准则第14号-收入》规定的收入确认基本原则，分析贵公司收入确认时点是否满足收入确认条件；
(2) 将贵公司收入确认政策与同行业上市公司进行对比分析；
(3) 了解、测试销售及收款相关的内部控制制度。

2023-09-17 14:54:47.366 | INFO | generate_answer_with_classify:generate_answer:372 - 根据年报内容，安硕信息2021年度的审计意见为“标准的无保留意见”。这意味着审计员对公司的财务报表整体表示满意，公司的财务报表在所有重大方面都按照企业会计准则的规定编制，公允反映了公司2021年12月31日的合并及母公司财务状况以及2021年度的合并及母公司经营成果和现金流量。审计员认为，后附的财务报表在所有重大方面都符合中国注册会计师审计准则的规定，并为发表审计意见提供了基础。

问题回答 - 一些测试结果

问：安硕信息在2021年的公司股权激励计划是什么样的？

你需要阅读理解年报的片段来真实详细完整的回答用户的提问。

下面是年报内容格式的一些说明：

1. 片段由标题和正文内容组成。
1. 片段中的标题通常以中文数字如一二三四五或阿拉伯数字12345开始。
2. 片段中的表格采用制表符“\t”分隔。
3. “√适用”表示该项内容公司存在该事项，“√不适用”表示公司不存在该事项。
4. “√是”表示该项是或者有，“√否”表示该项不是或者没有。

*****安硕信息2021年年报*****

-----片段:1-----

十三、公司股权激励计划、员工持股计划或其他员工激励措施的实施情况

√适用

1、股权激励

1、公司于2021年11月29日召开的第四届董事会第十次会议及第四届监事会第十次会议，分别审议通过了《关于公司2020年限制性股票激励计划第一类限制性股票第一个解除限售期解除限售条件成就的议案》和《关于调整2020年限制性股票激励计划回购价格的议案》，为符合解除限售条件的激励对象共计412人，可解除限售的限制性股票数量为999,050股。同时，根据《上海安硕信息技术股份有限公司2020年限制性股票激励计划（草案）》的有关规定及公司2020年第二次临时股东大会的授权，公司董事会对限制性股票的回购价格进行调整，第一类限制性股票的回购价格由10.60元/股调整为10.40元/股。

2、公司于2021年12月15日召开的2021年第二次临时股东大会已审议通过《关于回购注销2020年限制性股票激励计划部分限

制性股票的议案》，本次回购注销激励对象已获授但尚未解除限售的限制性股票合计213,975股，涉及人数52人。其中的36

名激励对象已辞职，不再具备激励对象资格，16名激励对象因个人层面绩效考核要求不符合全部解除限售的条件。本次部

分限制性股票回购注销事宜完成后，公司总股本将由141,009,750股减少至140,795,775股。

3、公司将回购注销10名已离职激励对象尚未解除限售的第一类限制性股票，鉴于公司2021年业绩考核未达标，公司将回购注销第二个解除限售期的第一类限制性股票。共计回购注销第一类限制性股票1,034,72

5股。公司将作废49名已离职激励对象尚未归属的第二类限制性股票，鉴于公司2021年业绩考核未达标，公司将作废第二个归属期的第二类限制性股票。共计作废第二类限制性股票1,193,600股。

4、关于限制性股票激励计划，结合公司目前的经营情况，在本资产负债表日时点合理估计未来合并净利润难以完成股权激励年度目标，因此资产负债表日不计提第三期相关股权激励成本费用支出。

董事、高级管理人员获得的股权激励情况

√适用

单位：股

职务	年初持有股票期权数量	报告期新授予股票期权数量	报告期内可行权股数	报告期内已行权股数	报告期内已行权股数行权价格（元/股）	期末持有股票	报告期
末市价	期初持有限制性股票数量	报告期新授予限制性股票数量	限制性股票的	本期已解锁			

-----片段:2-----

根据《上海安硕信息技术股份有限公司2020限制性股票激励计划（草案）》（以下简称“《激励计划》”）相关规定以及2020年第二次临时股东大会的授权，鉴于公司本激励计划授予的激励对象中36人因个人原因已离职，不再具备激励对象资格，16人因个人年度绩效考核未达标，公司向上述52名激励对象回购部分/全部已获授但尚未解除限售的限制性股票。

本次回购股份的种类为股权激励限售股。本次回购注销的限制性股票合计213,975股，占股权激励计划所涉及的标的股票的比例为5.99%，占回购前公司总股本的比例为0.15%，回购价格为10.40元/股，回购金额为2,225,340.00元。

经中国证券登记结算有限责任公司深圳分公司审核确认，公司本次限制性股票回购注销事宜已于2022年4月6日办理完成。本次回购注销完成后，公司总股本将由141,009,750股变更为140,795,775股。

大信会计师事务所（特殊普通合伙）于2022年03月08日出具了“大信验字[2022]第4-00011号”验资报告，对公司截止2022

年1月31日减少注册资本及股本的情况进行了审验，审验结果为：截止2022年1月31日止，变更后的注册资本为人民币140,795,775元，股本为人民币140,795,775元。

十六、其他重要事项

1、前期会计差错更正

(1)追溯重述法

单位：元

处理程序	受影响的各个比较期间报表项目名称	累积影响数
会计差错更正的内容		
(2)未来适用法		
会计差错更正的内容	批准程序	采用未来适用法的原因
2、债务重组		
3、资产置换		

2023-09-17 14:54:44.290 | INFO | generate_answer_with_classify:generate_answer:372 - 安硕信息在2021年的公司股权激励计划包括以下内容：

1. 股权激励计划

公司于2021年11月29日召开的第四届董事会第十次会议及第四届监事会第十次会议，分别审议通过了《关于公司2020年限制性股票激励计划第一类限制性股票第一个解除限售期解除限售条件成就的议案》和《关于调整2020年限制性股票激励计划回购价格的议案》，为符合解除限售条件的激励对象共计412人，可解除限售的限制性股票数量为999,050股。同时，根据《上海安硕信息技术股份有限公司2020年限制性股票激励计划（草案）》的有关规定及公司2020年第二次临时股东大会的授权，公司董事会对限制性股票的回购价格进行调整，第一类限制性股票的回购价格由10.60元/股调整为10.40元/股。

2. 员工持股计划

公司还实施了员工持股计划，以鼓励公司员工积极进取，提高公司业绩。该计划的具体实施情况在年报中没有详细说明。

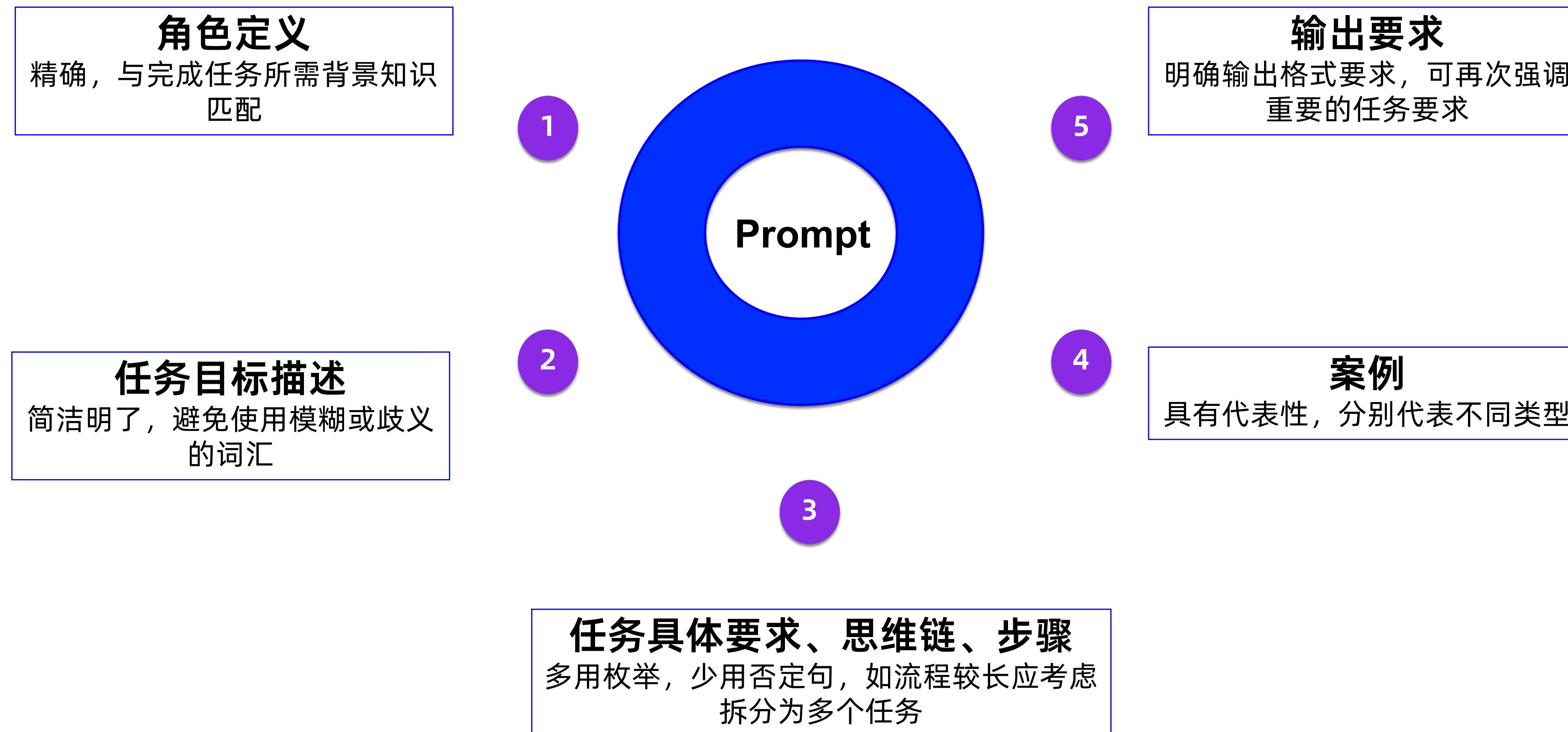
需要注意的是，上述股权激励计划已经进行了多次调整和修改，具体实施情况可能与本次年报有所不同。

问题回答 - 性能

	问题数量	总时间	单条处理时间	QPS
分类	1956	258s	0.13s	7.70
SQL生成	200	540s	2.7s	0.37
意图识别	1956	489s	0.25s	4.00

	问题数量	总时间	单条处理时间	QPS
Type1	683	365s	0.53s	1.88
Type2	485	627s	1.29s	0.77
Type3	587	2930s	4.99s	0.20

总结 - Prompt结构设计



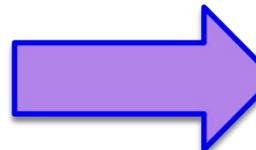
总结 - 模型微调

只用到一个模型，三个强化能力是各自单独的微调权重，根据传入的微调类型自动加载权重

```
# 创建一个枚举类型，不同ptuning分类
class PtuningType(Enum):
    Nothing = 0
    Classify = 1
    Keywords = 2
    NL2SQL = 3
```

```
ChatGLM_Ptuning.  
__init__(self, ptuning_type: PtuningType)
```

```
# 卸载掉已经装在权重的模型
def unload_model(self):
    del self.model
    del self.tokenizer
    torch.cuda.empty_cache()
```



不加载微调权重，使用预训练模型，保持问答能力

```
ChatGLM_Ptuning(PtuningType.Nothing)
```

加载问题分类微调权重，强化用户提问分类能力

```
ChatGLM_Ptuning(PtuningType.Classify)
```

加载关键词提取微调权重，强化意图识别能力

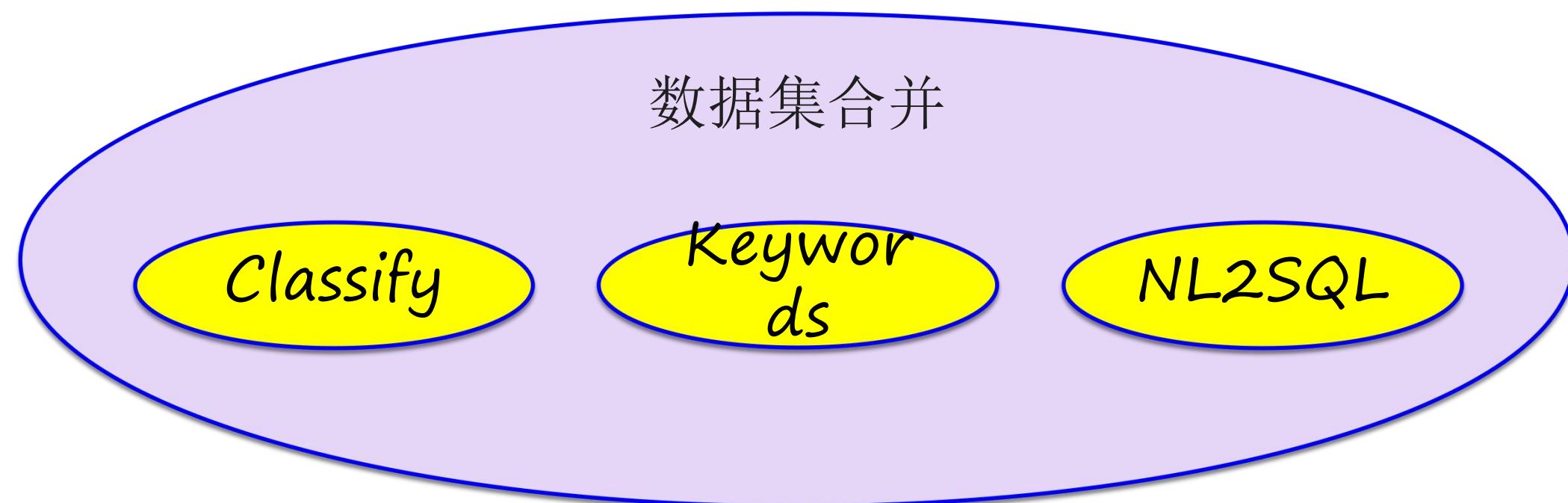
```
ChatGLM_Ptuning(PtuningType.Keywords)
```

加载NL2SQL微调权重，强化SQL语言生成能力

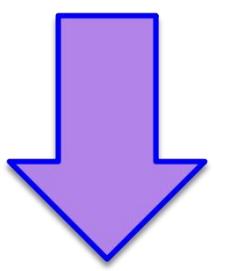
```
ChatGLM_Ptuning(PtuningType.NL2SQL)
```

总结 - 模型微调

只用到一个模型，项目后期趋于稳定，合并训练集微调权重，只需要加载一次权重即可获得3种能力



Ptuning v2



All In One

```
# 创建一个枚举类型，不同ptuning分类
class PtuningType(Enum):
    Nothing = 0
    AllInOne = 1
```

```
ChatGLM_Ptuning.  
__init__(self, ptuning_type: PtuningType)
```

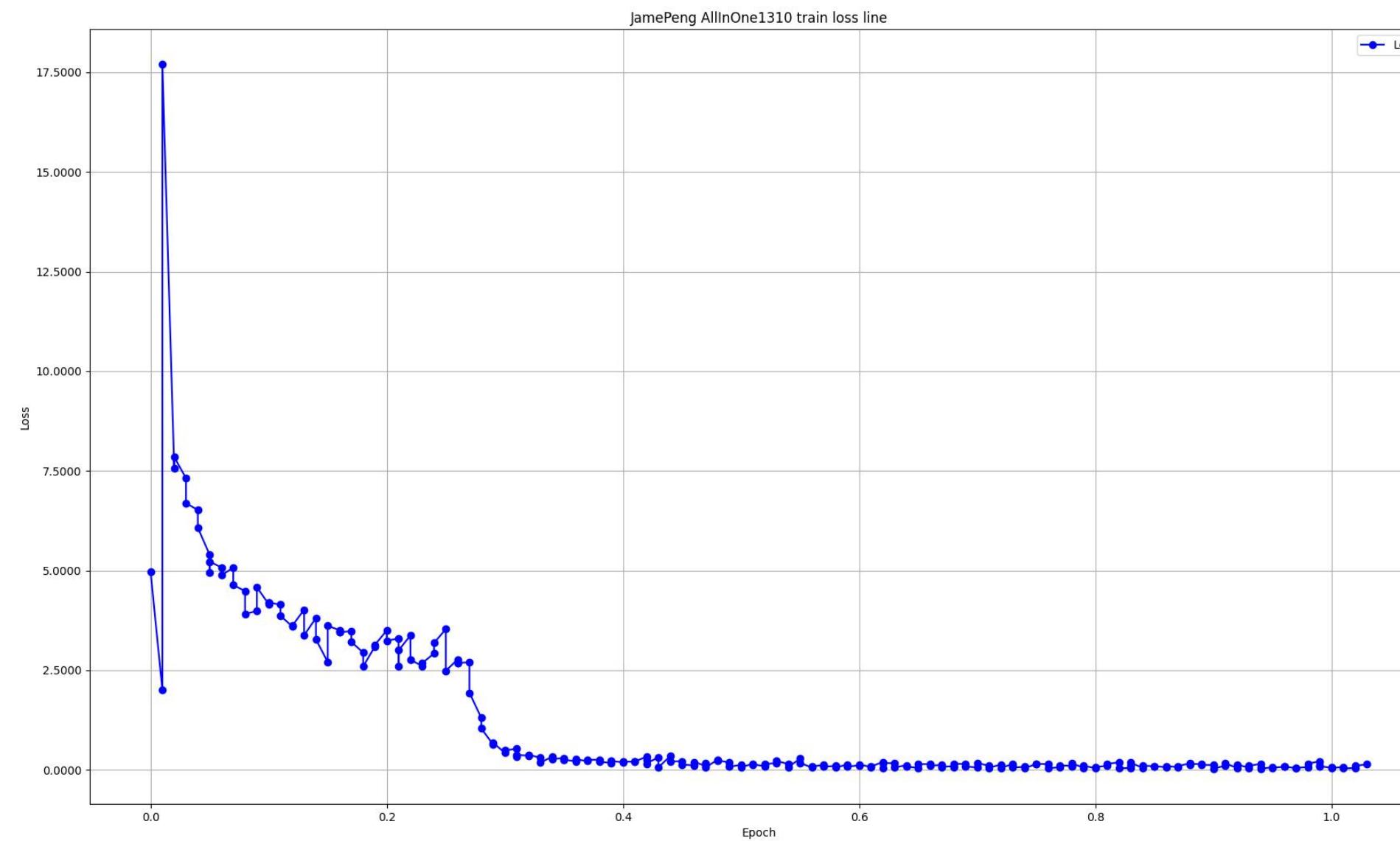
不加载微调权重，使用预训练模型，保持问答能力

```
ChatGLM_Ptuning(PtuningType.Nothing)
```

加载AllInOne微调权重，强化三合一微调的能力

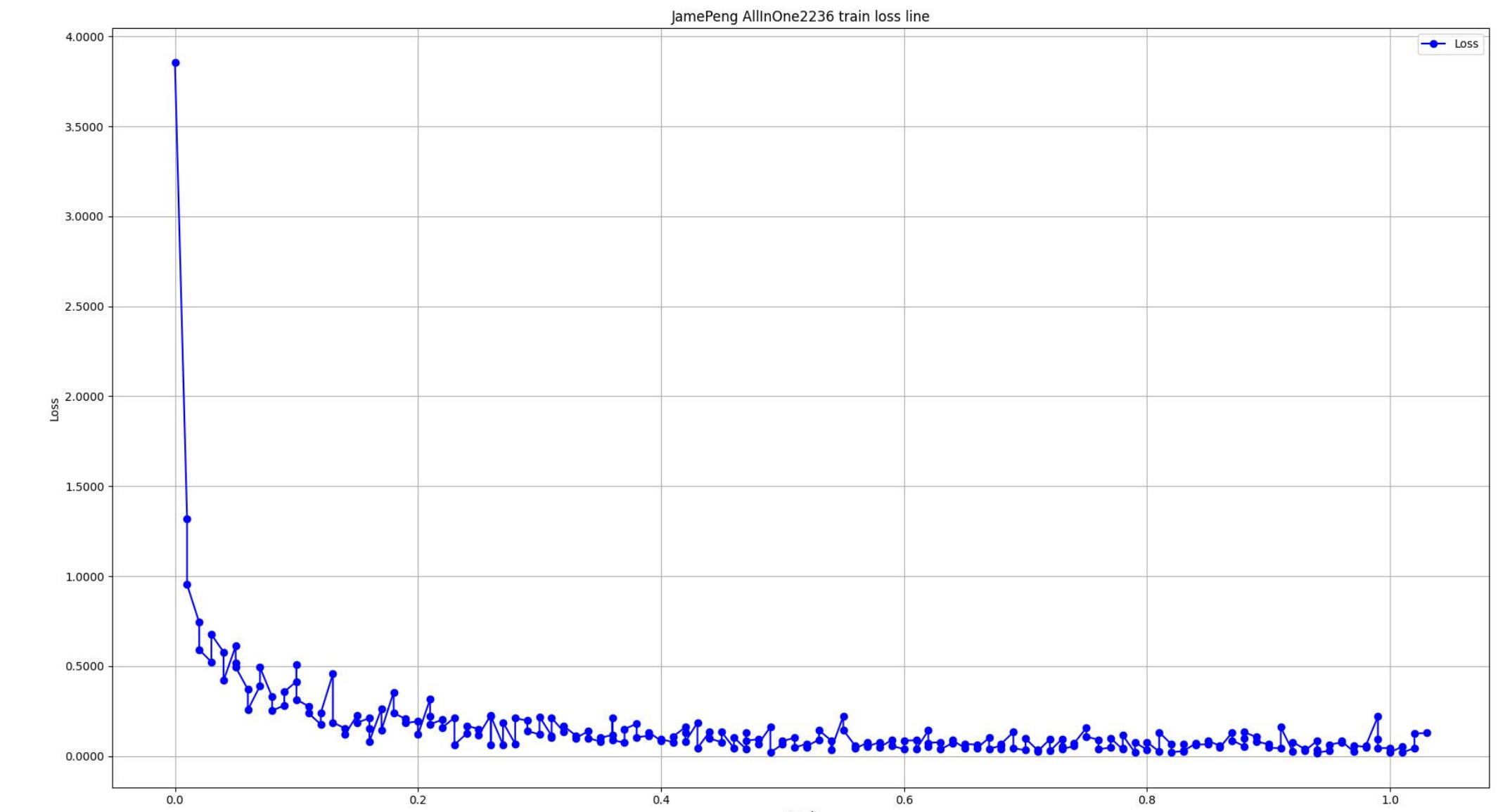
```
ChatGLM_Ptuning(PtuningType. AllInOne)
```

总结 - 模型微调



1310方案: 【权重大小379MB】

- 1、直接合并数据集，训练最大长度2100
- 2、学习率为 $2e-2$
- 3、训练STEP 450
- 4、Loss 0.03-0.13波动
- 5、学习率到0.007 需要STEP 292



2236方案: 【权重大小169MB】

- 1、精简SQL数据集后合并，训练最大长度1024
- 2、学习率为 $1e-2$
- 3、训练STEP 450
- 4、Loss 0.01-0.12波动
- 5、学习率到0.007 需要STEP 132

总结 - 优化方向

- 准确性

- 提高召回准确率, 匹配/语义
- 增强模型指令遵循的能力, 简单指令/复杂链式指令/其他指令
- 提高模型自身的知识能力, 知识库/知识图谱

- 速度

- 更精准的召回, 减少需要处理的 token数
- 大小模型协同, 多模型协同
- 工程方面, cache/量化

- 通用性

- 检索再回答, 搜索引擎?
- 直接回答, 微调/全量训练

总结 - 大模型

- 大模型的泛化处理能力, 可以适应各种NLP任务, 文本分类/意图识别/实体提取/SQL生成等等
 - 指令识别能力
 - 示例学习能力
- 大模型的可微调性, 性能优异的模型基座, 少量数据微调能得到非常不错的效果
 - P-Tuning



总结 - 其他

01

比赛中的团队合作

- 数据处理
- 模型训练
- 流程处理

02

开源社区的力量

- 智谱ChatGLM系列模型
- 安硕
- 魔搭

03

大模型+医疗

- 医疗辅助诊断
- 医疗知识库构建
- 患者咨询

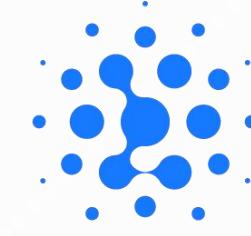
总结 - 成绩历程

复赛C榜		复赛B榜		复赛A榜		初赛			
排名	参与者	组织	score	type1Score	type2Score	type3-1Score	type3-2Score		
1 ↑1	众安硕眼探企	上海安硕信息技术股份有限公司	82.20	91.00	76.00	80.00	85.00		
2 ↑1	众种菜小队	浙江大学	75.20	79.00	69.00	75.00	89.00		
3 ↑1	众金融智能	人大-北大-广联达	74.20	67.00	77.00	72.00	89.00		
4 ↑1	众SodaAI	厦门苏哒智能	73.30	64.00	73.00	80.00	89.00		
5 ↑1	众南哪都队	南京大学	73.20	63.00	75.00	77.00	89.00		
6 ↑1	众nsddd	中国科学院	72.80	66.00	69.00	82.00	90.00		
6 ↑34	众Tablsabaopilong	上海交通大学	72.80	66.00	69.00	82.00	90.00		
8	众馒头科技	Manteia/京发科技/龙程电子	72.40	69.00	66.00	82.00	89.00		
9	众结婚买房代代韭菜	结婚买房代代韭菜	71.90	58.00	73.00	82.00	89.00		
10	众chatglm反卷总局	腾讯	71.80	67.00	67.00	80.00	89.00		

复赛C榜		复赛B榜		复赛A榜		初赛			
排名	参与者	组织	score	type1Score	type2Score	type3-1Score	type3-2Score		
1	众馒头科技	Manteia/京发科技/龙程电子	88.58	92.70	89.57	80.09	89.20		
2	众安硕眼探企	上海安硕信息技术股份有限公司	88.52	93.19	92.02	74.14	89.31		
3	众流宝真人	杭州大道一以科技有限公司	85.93	87.36	89.03	76.24	88.68		
4	众chatglm反卷总局	腾讯	85.84	86.79	89.45	75.64	88.99		
5	众南哪都队	南京大学	84.29	88.35	85.98	72.61	88.68		
6	众吃辣子	饺子研究院	82.20	77.09	87.82	74.65	90.13		
7	众IntelliSeekers		81.76	82.17	87.43	65.90	89.62		
8	众龙盈战队	龙盈智达(北京)科技有限公司	81.71	83.37	81.54	75.94	88.94		
9	众Tablsabaopilong	上海交通大学	80.16	87.86	74.80	74.56	89.70		
10	众nsddd	中国科学院	79.95	87.49	74.37	74.83	89.87		

复赛C榜		复赛B榜		复赛A榜		初赛			
排名	参与者	组织	score	type1Score	type2Score	type3-1Score	type3-2Score		
1	众馒头科技	Manteia/京发科技/龙程电子	83.07	91.34	77.24	79.58	88.54		
2	众chatglm反卷总局	腾讯	79.35	80.98	73.74	83.36	88.86		
3	众种菜小队	浙江大学	78.78	85.76	67.42	85.52	89.78		
4	众结婚买房代代韭菜	结婚买房代代韭菜	78.04	81.92	69.34	84.18	88.93		
5	众nsddd	中国科学院	75.87	80.45	65.30	83.30	89.58		
6	众南哪都队	南京大学	75.81	71.74	73.42	80.42	88.33		
7	众馒头科技	Manteia/京发科技/龙程电子	75.62	83.11	63.40	82.13	89.01		
8	众龙盈战队	龙盈智达(北京)科技有限公司	74.28	68.52	71.01	82.07	89.08		
9	众偏要为空	吾道科技	72.85	72.98	64.58	81.24	88.71		
10	众Tablsabaopilong	上海交通大学	72.59	70.81	64.27	83.19	90.04		

复赛C榜		复赛B榜		复赛A榜		初赛			
排名	参与者	组织	score	type1Score	type2Score	type3-1Score	type3-2Score		
1	众馒头科技	Manteia/京发科技/龙程电子	88.62	91.64	91.90	77.06	89.54		
2	众安硕眼探企	上海安硕信息技术股份有限公司	88.33	92.97	93.09	71.31	89.44		
3	众流宝真人	杭州大道一以科技有限公司	87.82	91.35	90.89	75.89	88.79		
4	众chatglm反卷总局	腾讯	87.23	87.33	92.55	75.22	89.64		
5	众吃辣子	饺子研究院	86.32	83.73	92.93	75.15	89.97		
6	众nsddd	中国科学院	85.89	85.14	91.30	74.24	89.83		
7	众龙盈战队	龙盈智达(北京)科技有限公司	85.85	86.36	88.67	77.61	89.50		
8	众Tablsabaopilong	上海交通大学	85.04	86.02	88.69	73.77	89.97		
9	众结婚买房代代韭菜	结婚买房代代韭菜	84.60	82.45	90.02	74.56	89.48		
10	众南哪都队	南京大学	84.38	87.32	87.75	70.80	89.24		



智谱·AI



| 安硕信息



北京交通大学
BEIJING JIAOTONG UNIVERSITY



ModelScope
魔搭社区



Thanks