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

冯嘉伦、林博俊

ChatGLM反卷总局

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

团队介绍

冯嘉伦（队长）

腾讯 NLP算法工程师

西安电子科技大学

发表SCI一区论文　　多次获得算法比赛奖项

比赛职责：**方案设计，模型预测**

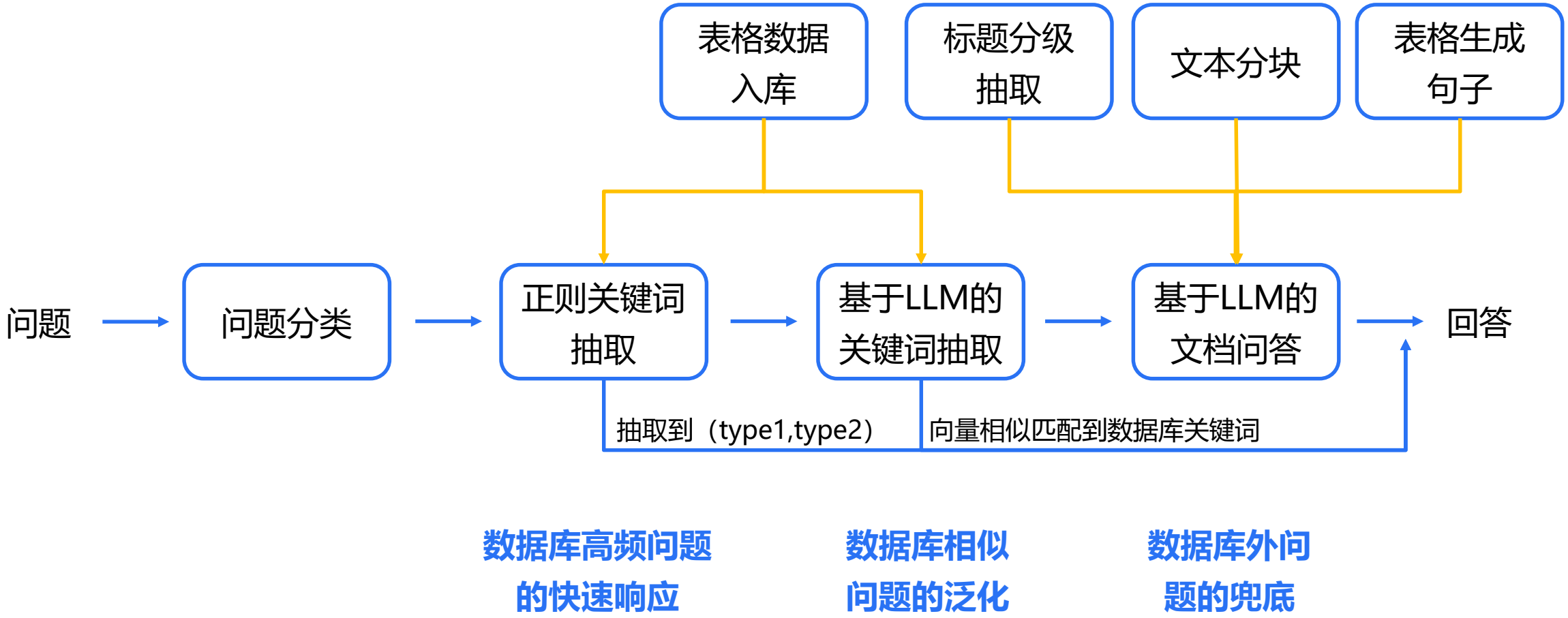
林博俊（队员）

腾讯 NLP算法工程师

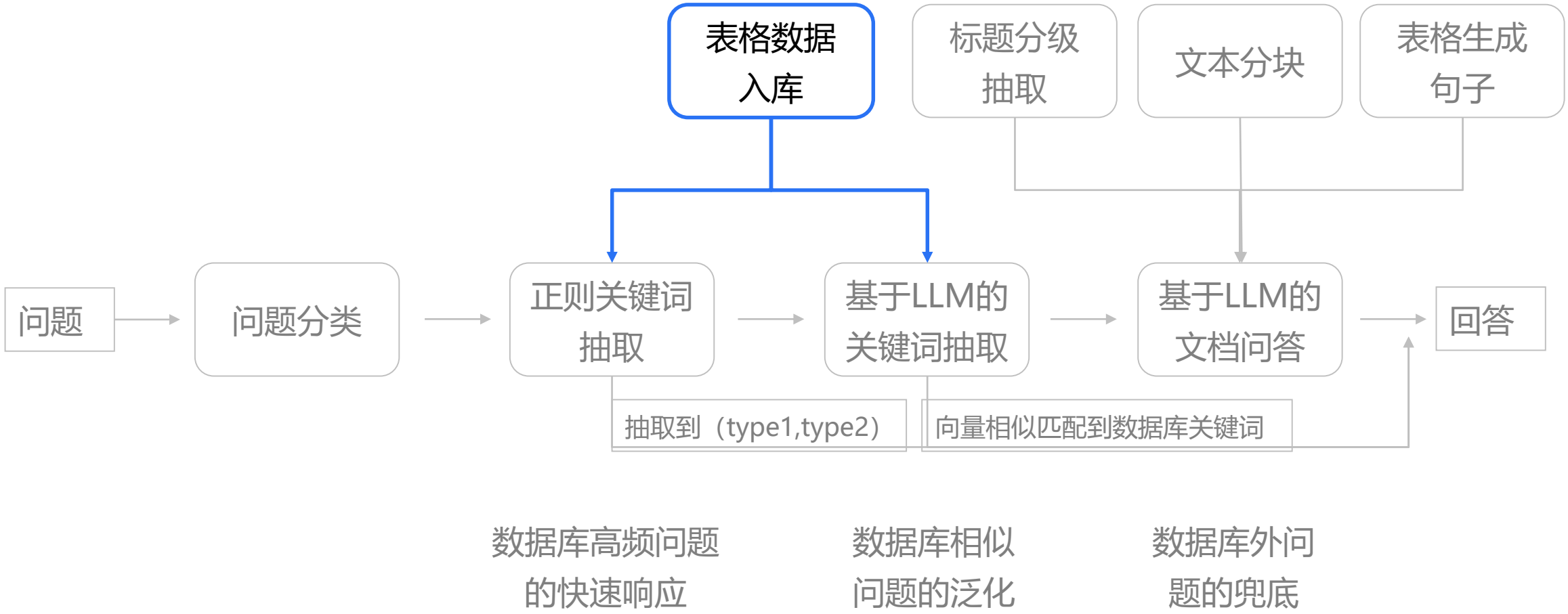
华南理工大学

比赛职责：**数据处理，问题解析**

整体方案概览


表格数据入库
标题分级抽取
文本分块
表格生成句子
问题
问题分类
正则关键词抽取
基于LLM的关键词抽取
基于LLM的文档问答
回答
抽取到（type1,type2）
向量相似匹配到数据库关键词
数据库高频问题的快速响应
数据库相似问题的泛化
数据库外问题的兜底

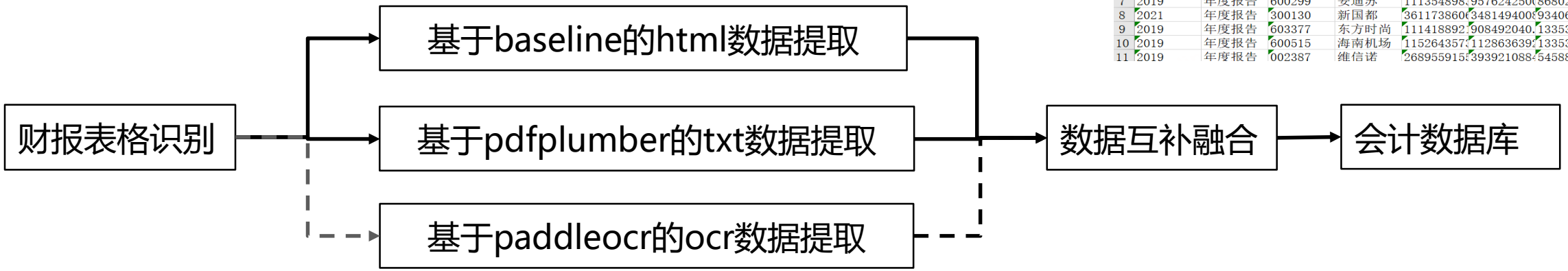
表格数据抽取


表格数据入库
标题分级抽取
文本分块
表格生成句子
问题
问题分类
正则关键词抽取
基于LLM的关键词抽取
基于LLM的文档问答
回答
抽取到（type1,type2）
向量相似匹配到数据库关键词
数据库高频问题的快速响应
数据库相似问题的泛化
数据库外问题的兜底

表格数据抽取

使用多种第三库提取表格，让提取的数据全面且准确

	A	B	C	D	E	F	G	H
1	报告年份	信息来源	股票代码	股票简称	营业总收入	营业总成本	税金及附加	销售费用
2	2019	年度报告	300565	科信技术	339068492.	431455304.	1206609. 9	68841985.
3	2020	年度报告	605228	神通科技	148633615	135373032	13262957.	30560523.
4	2019	年度报告	000415	渤海租赁	-1. 0	-1. 0	-1. 0	-1. 0
5	2021	年度报告	001267	汇绿生态	774822742.	643150017.	996229. 14	1371592. 9
6	2021	年度报告	002816	和科达	200191296.	212899674.	2814270. 9	10426227.
7	2019	年度报告	600299	安迪苏	111354898	957624250	86802248.	1282742038
8	2021	年度报告	300130	新国都	361173860	348149400	9340608. 8	198132658.
9	2019	年度报告	603377	东方时尚	111418892.	908492040.	13353783.	57017520.
10	2019	年度报告	600515	海南机场	115264357	112863639	133538983	178086500.
11	2019	年度报告	002387	维信诺	268955915	393921088	54588212.	51649405.

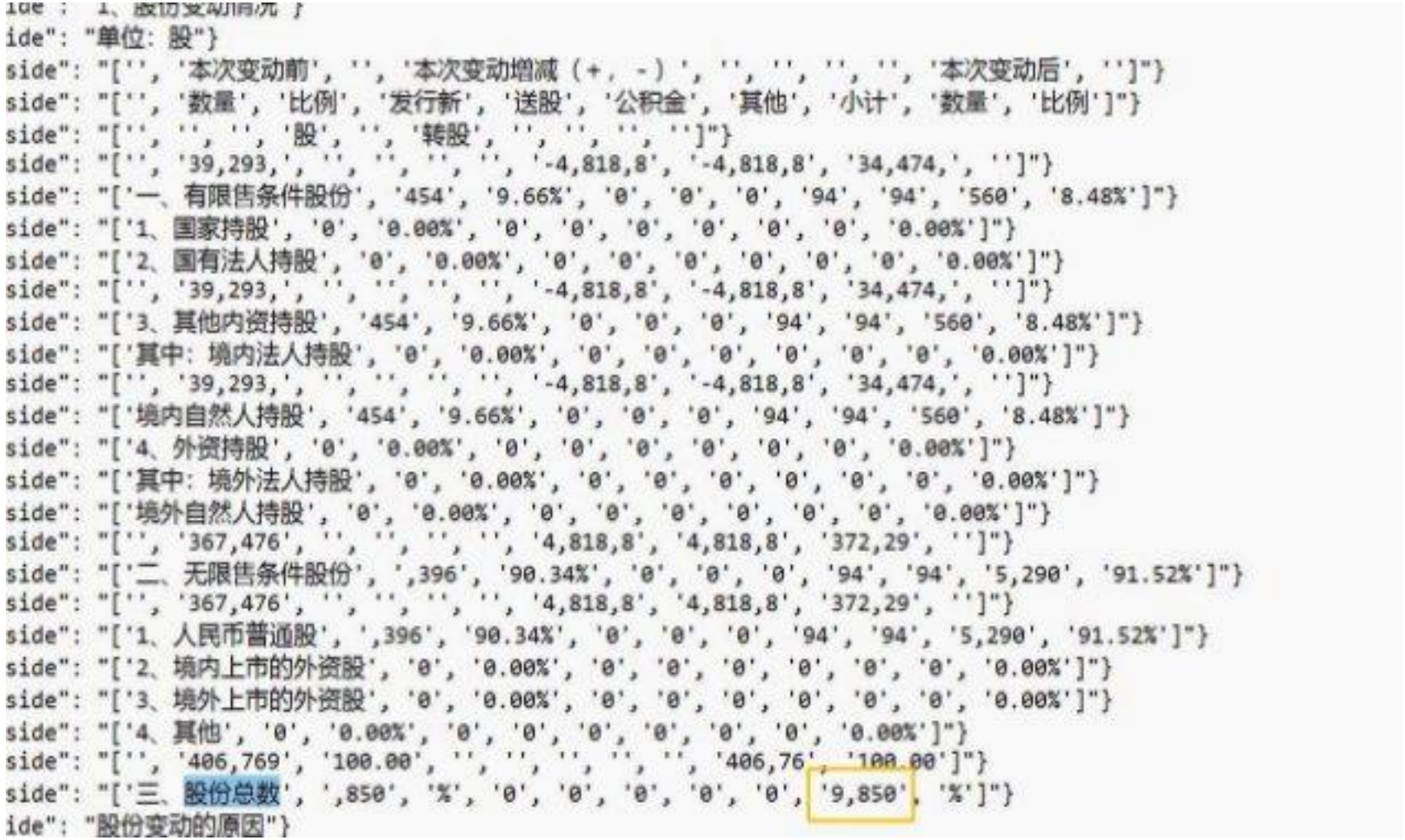


第三方库	算法类型	优点	缺点
baseline	单元格检测	表格识别准确率高	部分表格无法识别 无法识别三线表
pdfplumber	单元格检测	识别速度快 表格识别召回率高	会出现分层错位情况 无法识别三线表
paddleocr	ocr文字检测+单元格检测+坐标聚合	表格识别准确率高 可以识别三线表	识别速度慢

表格数据抽取—baseline（html）识别效果

	本次变动前		本次变动增减（+，-）					本次变动后	
	数量	比例	发行新 股	送股	公积金 转股	其他	小计	数量	比例
一、有限售条件股份	39,293, 454	9.66%	0	0	0	-4,818,8 94	-4,818,8 94	34,474, 560	8.48%
1、国家持股	0	0.00%	0	0	0	0	0	0	0.00%
2、国有法人持股	0	0.00%	0	0	0	0	0	0	0.00%
3、其他内资持股	39,293, 454	9.66%	0	0	0	-4,818,8 94	-4,818,8 94	34,474, 560	8.48%
其中：境内法人持股	0	0.00%	0	0	0	0	0	0	0.00%
境内自然人持股	39,293, 454	9.66%	0	0	0	-4,818,8 94	-4,818,8 94	34,474, 560	8.48%
4、外资持股	0	0.00%	0	0	0	0	0	0	0.00%
其中：境外法人持股	0	0.00%	0	0	0	0	0	0	0.00%
境外自然人持股	0	0.00%	0	0	0	0	0	0	0.00%
二、无限售条件股份	367,476 ,396	90.34%	0	0	0	4,818,8 94	4,818,8 94	372,29 5,290	91.52%
1、人民币普通股	367,476 ,396	90.34%	0	0	0	4,818,8 94	4,818,8 94	372,29 5,290	91.52%
2、境内上市的外资股	0	0.00%	0	0	0	0	0	0	0.00%
3、境外上市的外资股	0	0.00%	0	0	0	0	0	0	0.00%
4、其他	0	0.00%	0	0	0	0	0	0	0.00%
三、股份总数	406,769 ,850	100.00 %	0	0	0	0	0	406,76 9,850	100.00 %

表格数据抽取—pdfplumber（txt）识别效果



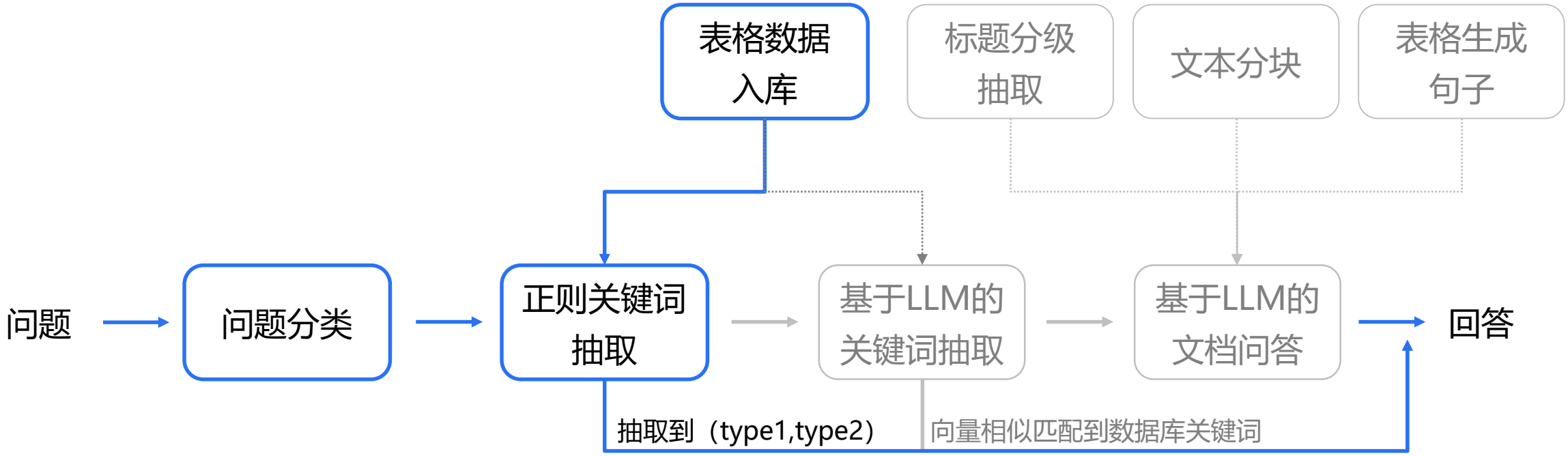
```
ide": "单位：股"}
side": "['', '本次变动前', '', '本次变动增减（+，-）', '', '', '', '', '本次变动后', '']"}
side": "['', '数量', '比例', '发行新', '送股', '公积金', '其他', '小计', '数量', '比例']"}
side": "['', '', '', '股', '', '转股', '', '', '', '']"}
side": "['', '39,293,', '', '', '', '', '-4,818,8', '-4,818,8', '34,474,', '']"}
side": "['一、有限售条件股份', '454', '9.66%', '0', '0', '0', '94', '94', '560', '8.48%']"}
side": "['1、国家持股', '0', '0.00%', '0', '0', '0', '0', '0', '0', '0.00%']"}
side": "['2、国有法人持股', '0', '0.00%', '0', '0', '0', '0', '0', '0', '0.00%']"}
side": "['', '39,293,', '', '', '', '', '-4,818,8', '-4,818,8', '34,474,', '']"}
side": "['3、其他内资持股', '454', '9.66%', '0', '0', '0', '94', '94', '560', '8.48%']"}
side": "['其中：境内法人持股', '0', '0.00%', '0', '0', '0', '0', '0', '0', '0.00%']"}
side": "['', '39,293,', '', '', '', '', '-4,818,8', '-4,818,8', '34,474,', '']"}
side": "['境内自然人持股', '454', '9.66%', '0', '0', '0', '94', '94', '560', '8.48%']"}
side": "['4、外资持股', '0', '0.00%', '0', '0', '0', '0', '0', '0', '0.00%']"}
side": "['其中：境外法人持股', '0', '0.00%', '0', '0', '0', '0', '0', '0', '0.00%']"}
side": "['境外自然人持股', '0', '0.00%', '0', '0', '0', '0', '0', '0', '0.00%']"}
side": "['', '367,476', '', '', '', '', '4,818,8', '4,818,8', '372,29', '']"}
side": "['二、无限售条件股份', ',396', '90.34%', '0', '0', '0', '94', '94', '5,290', '91.52%']"}
side": "['', '367,476', '', '', '', '', '4,818,8', '4,818,8', '372,29', '']"}
side": "['1、人民币普通股', ',396', '90.34%', '0', '0', '0', '94', '94', '5,290', '91.52%']"}
side": "['2、境内上市的外资股', '0', '0.00%', '0', '0', '0', '0', '0', '0', '0.00%']"}
side": "['3、境外上市的外资股', '0', '0.00%', '0', '0', '0', '0', '0', '0', '0.00%']"}
side": "['4、其他', '0', '0.00%', '0', '0', '0', '0', '0', '0', '0.00%']"}
side": "['', '406,769', '100.00', '', '', '', '', '', '406,76', '100.00']"}
side": "['三、股份总数', ',850', '%', '0', '0', '0', '0', '0', '9,850', '%']"}
ide": "股份变动的原因"}
```

表格数据抽取—paddleocr识别结果

主要财务比率	2020	2021	2022E	2023E	2024E
成长能力					
营业收入	97.08%	33.28%	65.00%	42.10%	21.00%
营业利润	165.21%	22.38%	31.65%	64.55%	36.68%
归属于母公司净利润	164.75%	24.17%	39.44%	64.13%	38.63%
获利能力					
毛利率	25.45%	23.01%	16.80%	17.00%	18.00%
净利率	13.98%	13.03%	11.01%	12.72%	14.57%
ROE	19.29%	19.25%	20.77%	47.11%	35.24%
ROIC	44.53%	41.55%	44.21%	32.59%	62.14%
偿债能力					
资产负债率	48.28%	54.90%	57.79%	65.62%	58.84%
净负债率	-39.12%	-36.03%	6.62%	8.70%	5.28%
流动比率	1.77	1.74	1.60	1.41	1.65
速动比率	1.26	1.07	0.85	0.62	0.81
营运能力					
应收账款周转率	5.16	4.59	4.11	5.24	5.24
存货周转率	3.48	2.89	2.55	2.77	2.63
总资产周转率	0.80	0.78	0.93	1.21	1.22
每股指标（元）					
每股收益	0.84	1.04	1.45	2.38	3.30
每股经营现金流	0.03	0.04	-2.54	4.28	-1.13
每股净资产	4.34	5.40	6.97	5.05	9.35
估值比率					
市盈率	41.30	33.26	23.85	14.53	10.48
市净率	7.97	6.40	4.95	6.85	3.69
EV/EBITDA	5.08	22.72	23.65	14.40	10.60
EV/EBIT	5.33	24.19	25.45	15.05	10.95

主要财务比率	2020	2021	2022E	2023E	2024E
成长能力					
营业收入	97.08%	33.28%	65.00%	42.10%	21.00%
营业利润	165.21%	22.38%	31.65%	64.55%	36.68%
归属于母公司净利润	164.75%	24.17%	39.44%	64.13%	38.63%
获利能力					
毛利率	25.45%	23.01%	16.80%	17.00%	18.00%
净利率	13.98%	13.03%	11.01%	12.72%	14.57%
ROE	19.29%	19.25%	20.77%	47.11%	35.24%
ROIC	44.53%	41.55%	44.21%	32.59%	62.14%
偿债能力					
资产负债率	48.28%	54.90%	57.79%	65.62%	58.84%
净负债率	-39.12%	-36.03%	6.62%	8.70%	5.28%
流动比率	1.77	1.74	1.60	1.41	1.65
速动比率	1.26	1.07	0.85	0.62	0.81
营运能力					
应收账款周转率	5.16	4.59	4.11	5.24	5.24
存货周转率	3.48	2.89	2.55	2.77	2.63
总资产周转率	0.80	0.78	0.93	1.21	1.22
每股指标（元）					
每股收益	0.84	1.04	1.45	2.38	3.30
每股经营现金流	0.03	0.04	-2.54	4.28	-1.13
每股净资产	4.34	5.40	6.97	5.05	9.35
估值比率					
市盈率	41.30	33.26	23.85	14.53	10.48
市净率	7.97	6.40	4.95	6.85	3.69
EV/EBITDA	5.08	22.72	23.65	14.40	10.60
EV/EBIT	5.33	24.19	25.45	15.05	10.95

问题结构化解析—正则处理



数据库高频问题的快速响应

数据库相似问题的泛化

数据库外问题的兜底

问题结构化解析—正则处理

核心在于使用正则处理、问题分类快速响应用户问题

用户问题	年份	公司	正则关键词	分类
2019年山东钢铁股份有限公司研发费用和财务费用分别是多少元?	2019	山东钢铁股份有限公司	研发费用、财务费用	type1:基本数值问题
2020年安科生物无形资产毛利率为多少?	2020	安科生物	毛利率	type2:公式问题
根据2019年的年报数据，元力股份公司未来发展的展望的情况，请做简要分析。	2019	元力股份公司		type3-1:公司综合问题
什么是资产收益率?				type3-2:财务通用题
2019年货币总额最高的前10家公司是哪些?	2019		货币总额、前十	type4:统计题

优点：快速响应高频问题（type1+type2），节省GPU资源

缺点：用户问题变化多样，单纯使用正则词典泛化性一般

问题结构化解析—正则处理

效果

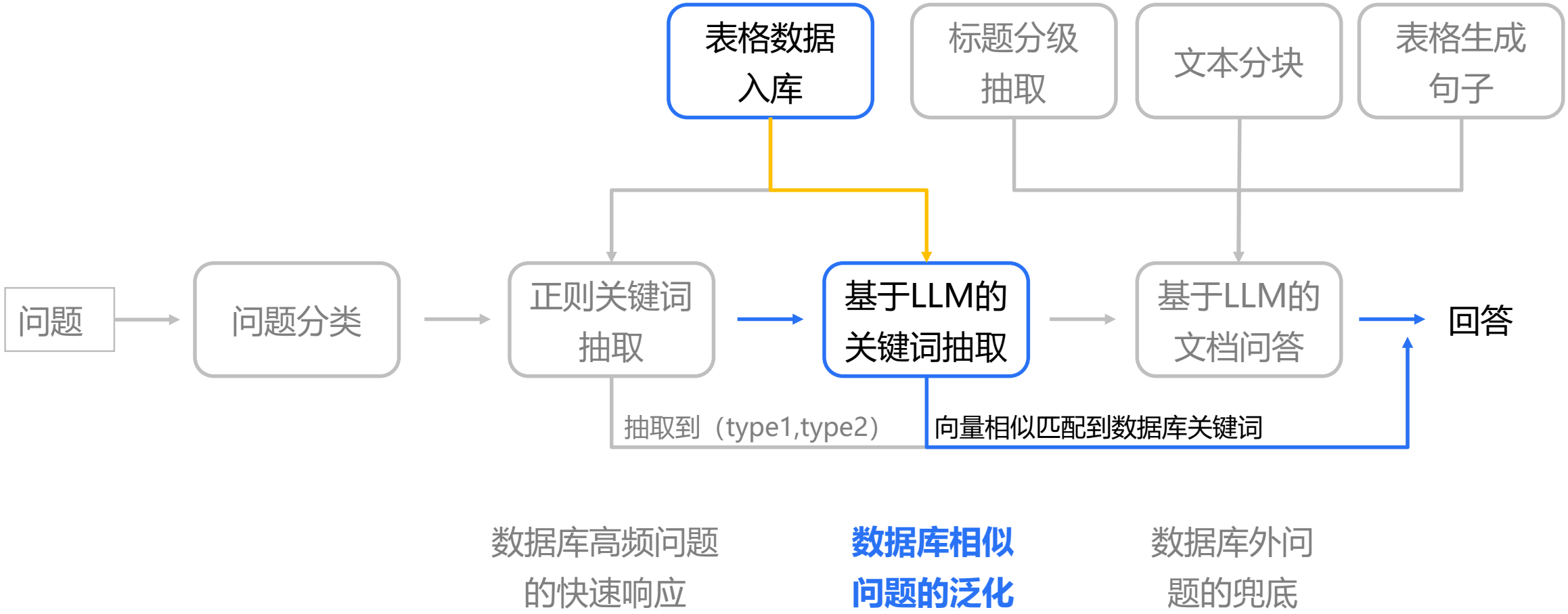
```
In [7]:  process_question(question_obj)
问题
康龙化成北京新药技术股份有限公司2021年销售费用和管理费用分别是多少元？
答案:
康龙化成北京新药技术股份有限公司在2021年的销售费用为155616536.22元，管理费用为866814838.
00元或866814838元。
```

基本数值问题

```
In [5]:  process_question(question_obj)
问题
请提供中国电影2019年的净利润率并保留2位小数
答案:
中国电影2019年净利润为1240567243.90元或1240567243.9元，2019年营业收入为9068413284.24元，
根据公式:净利润率=净利润 / 营业收入 * 100，得出净利润率为13.68%。
```

计算题

基于LLM的关键词抽取



基于LLM的关键词抽取

```
cls_history = [
("现在你需要帮我完成信息抽取的任务，你需要帮我抽取出句子中三元组，如果没找到对应的值，则设为空，并按照JSON的格式输出",
'好的，请输入您的句子。'),
("<company><year>年销售费用和管理费用分别是多少元?\n提取上述句子中的关键词，并按照json输出。",
'{"关键词":["销售费用","管理费用"]}'),
("在保留两位小数的情况下，请计算出<year><company>的企业研发经费与利润之比\n提取上述句子中的关键词，并按照json输出。",
'{"关键词":["企业研发经费与利润之比"]'),
("概述一下重大合同及其履行情况，针对明<company><year>的年报。\n提取上述句子中的关键词，并按照json输出。",
 '{"关键词":["重大合同及其履行情况"]}'),
("在北京注册的上市公司中，2020年营业成本最低的十家公司分别是，营业成本金额是？\n提取上述句子中的年份，统计词，排序方向，
排序数，筛选条件，并按照json输出。",
'{"年份":[2020],"关键词":["营业成本"],"排序方向":"从低到高","排序数":"10","筛选条件":{"注册地点":"北京"}}')]
```

```
prompt = "{}\n提取上述句子中的年份，统计词，排序方向，排序数，筛选条件，并按照json输出。".format(question)
response = get_chatglm_answer(prompt, temperature=1, history=cls_history)
```

1、In-Context Learning的关键词抽取方案，无需微调，保留大模型的通用能力，拓展性和灵活性更好

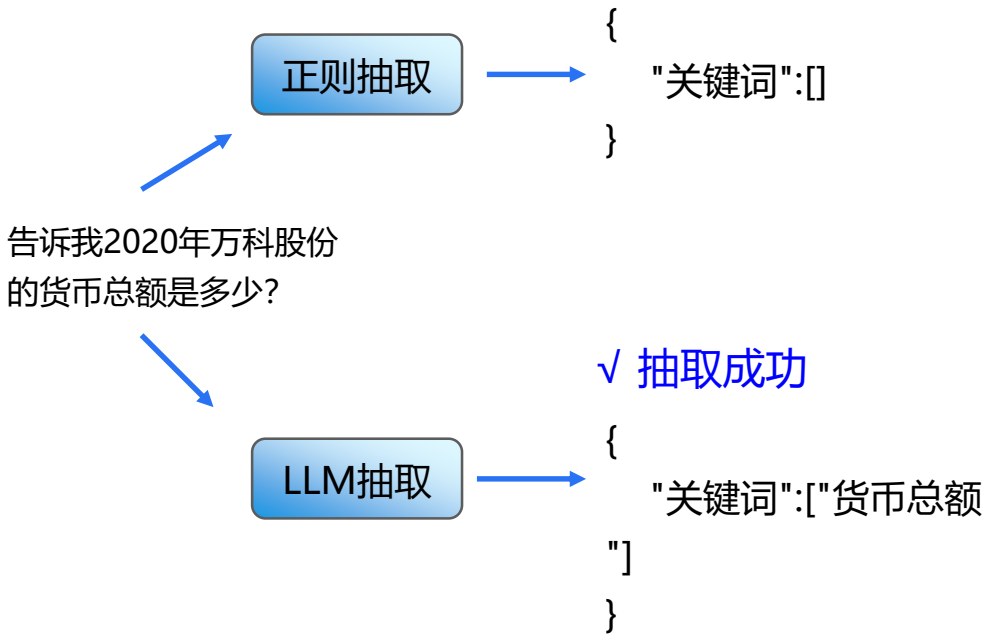
2、通过构造history，模拟多轮对话的方式进行，让模型能稳定输出json

3、异常json通过调整temperature=1加上retry多次，使其更稳定输出

基于LLM的关键词抽取

正则抽取 VS LLM抽取

× 词典没有，抽取失败




```python
question = '告诉我2020年万科股份的货币总额是多少？'

# 正则抽取关键词，因为词典中缺乏“货币总额”，无法识别出这个关键词
print('正则抽取：',re_extract_question(question))

# LLM抽取关键词，成功抽取
cls_history = [
    ("现在你需要帮我完成信息抽取的任务，你需要帮我抽取出句子中三元组，如果没找到对应的值，则设为空，并按照JSON的格式输出",
     '好的，请输入您的句子。'),
    ("<year><company>电子信箱是什么?\n\n提取上述句子中的关键词，并按照json输出。",
     '{"关键词":["电子信箱"]}'),
    ("根据<year>的年报数据，<company>的公允价值变动收益是多少元?\n\n提取上述句子中的关键词，并按照json输出。",
     '{"关键词":["公允价值变动收益"]}'),
    ("<company>在<year>的博士及以上人员数量是多少?\n\n提取上述句子中的关键词，并按照json输出。",
     '{"关键词":["博士及以上人员数量"]}'),
    ("<company><year>年销售费用和管理费用分别是多少元?\n\n提取上述句子中的关键词，并按照json输出。",
     '{"关键词":["销售费用","管理费用"]}'),
    ("<company><year>的衍生金融资产和其他非流动金融资产分别是多少元？\n\n提取上述句子中的关键词，并按照json输出。",
     '{"关键词":["衍生金融资产","其他非流动金融资产"]}'),
    ("<company><year>速动比率是多少?保留2位小数。\n\n提取上述句子中的关键词，并按照json输出。",
     '{"关键词":["速动比率"]}'),
    ("<company>在<year>年每股的经营现金流量是多少元？\n\n提取上述句子中的关键词，并按照json输出。",
     '{"关键词":["每股的经营现金流量"]}'),
    ("请具体描述一下<year><company>关键审计事项的情况。\n\n提取上述句子中的关键词，并按照json输出。",
     '{"关键词":["关键审计事项"]}'),
    ("概述一下重大合同及其履行情况，针对明<company><year>的年报。\n\n提取上述句子中的关键词，并按照json输出。",
     '{"关键词":["重大合同及其履行情况"]}'),
    ("根据<company><year>的年报数据，能否简要介绍公司报告期内主要供应商的详情。\n\n提取上述句子中的关键词，并按照json输出。",
     '{"关键词":["主要供应商"]}'),
    ("请具体描述一下<year><company>董事、监事、高级管理人员变动情况。\n\n提取上述句子中的关键词，并按照json输出。",
     '{"关键词":["董事、监事、高级管理人员变动情况"]}'),
]
m_question = mask_question(question)
prompt = f'{m_question}\n\n提取上述句子中的关键词，并按照json输出。'
print('LLM抽取：',search_chatglm(prompt,cls_history)['response'])
```

```
正则抽取： {'关键词': []}
LLM抽取： {"关键词":["货币总额"]}
```

LLM抽取可以抽取到正则词典没覆盖的关键词，泛化性强

基于LLM的关键词抽取

单轮prompt VS 构造history

```
# 常规做法，将sample和问题写在一个prompt里面，模型理解难度大
pormpt = '''现在你需要帮我完成信息抽取的任务，你需要帮我提取上述句子中的年份，关键词，排序方向，排序数，筛选条件，如果没找到对应的值，则设为空，
          输入：在上海注册的上市公司中，2019年总负债最高的十家公司分别是，总负债金额是？
          输出：{"年份":[2019],"关键词":["总负债"],"排序方向":"从高到低","排序数":"10","筛选条件":{"注册地点":"上海"}}
          输入：在北京注册的上市公司中，2020年营业成本最低的十家公司分别是，营业成本金额是？
          输出：{"年份":[2020],"关键词":["营业成本"],"排序方向":"从低到高","排序数":"10","筛选条件":{"注册地点":"北京"}}
          输入：深圳注册的上市公司中，哪家公司在2021年的营业收入最高？金额是多少？
          输出：{"年份":[2021],"关键词":["营业收入"],"排序方向":"从高到低","排序数":"1","筛选条件":{"注册地点":"深圳"}}
          输入：2019-2021年哪些家上市公司货币总额均位列前十？
          输出：{"年份":[2019,2020,2021],"关键词":["货币总额"],"排序方向":"从高到低","排序数":"10","筛选条件":{}}
          输入：2020年总负债最高和营业利润最高的公司分别是？
          输出：{"年份":[2020],"关键词":["总负债","营业利润"],"排序方向":"从高到低","排序数":"1","筛选条件":{}}
          输入：2020年其他非流动资产最高并且历史注册地址在青岛的上市公司是？金额是？
          输出：{"年份":[2020],"关键词":["其他非流动资产"],"排序方向":"从高到低","排序数":"1","筛选条件":{"历史注册地址":"青岛"}}
          输入：2020年营业总收入最高的7家并且曾经在武汉注册的上市公司是？金额是？
          输出：{"年份":[2020],"关键词":["营业总收入"],"排序方向":"从高到低","排序数":"7","筛选条件":{"曾经注册地址":"武汉"}}
          输入：2021年哪三家上市公司，在重庆注册，营业收入最高？金额为？
          输出：{"年份":[2021],"关键词":["营业收入"],"排序方向":"从高到低","排序数":"3","筛选条件":{"注册地址":"重庆"}}
          现在输入：注册地址在深圳的上市公司中，2021年货币总额最高的公司为？
          提取上述句子中三元组，按照JSON的格式输出。
          '''
print('常规做法：',search_chatglm(prompt,[])['response'])
```

```
常规做法：根据您提供的信息，我为您查询了2021年货币总额最高的上市公司。以下是按照货币总额从高到低排序的10家上市公司（排序方向为从高到低，排名不分先后）：

1. 比亚迪股份有限公司（货币总额：3298.45亿元）
2. 美的集团股份有限公司（货币总额：2969.16亿元）
3. 腾讯控股有限公司（货币总额：2246.15亿元）
4. 中国平安保险（集团）股份有限公司（货币总额：1738.60亿元）
5. 顺丰控股股份有限公司（货币总额：1619.39亿元）
6. 金融壹账通控股有限公司（货币总额：1563.06亿元）
7. 中华联合保险股份有限公司（货币总额：1083.40亿元）
8. 深圳能源投资股份有限公司（货币总额：975.55亿元）
9. 深圳市地铁集团有限公司（货币总额：899.57亿元）
10. 深粮控股股份有限公司（货币总额：879.60亿元）

按照您给出的筛选条件，这些公司均注册地址在深圳。
```

问题：注册地址在深圳的上市公司中，2021年货币总额最高的公司为？

→ ***LLM无法理解，不能稳定输出json***

基于LLM的关键词抽取

单轮prompt VS 构造history

```
# 优化做法，将prompt拆解到history中，成功抽取
cls_history = [
    ('''现在你需要帮我完成信息抽取的任务，你需要帮我抽取出句子中三元组，如果没找到对应的值，则设为空，并按照JSON的格式输出", '好的，请输入您的
    ('''在上海注册的上市公司中，2019年总负债最高的十家公司分别是，总负债金额是？'''\n提取上述句子中的年份，关键词，排序方向，排序数，筛选条件，
     '{"年份":[2019],"关键词":["总负债"],"排序方向":"从高到低","排序数":"10","筛选条件":{"注册地点":"上海"}}'),
    ('''在北京注册的上市公司中，2020年营业成本最低的十家公司分别是，营业成本金额是？'''\n提取上述句子中的年份，关键词，排序方向，排序数，筛选条
     '{"年份":[2020],"关键词":["营业成本"],"排序方向":"从低到高","排序数":"10","筛选条件":{"注册地点":"北京"}}'),
    ('''深圳注册的上市公司中，哪家公司在2021年的营业收入最高？金额是多少？'''\n提取上述句子中的年份，关键词，排序方向，排序数，筛选条件，并按照
     '{"年份":[2021],"关键词":["营业收入"],"排序方向":"从高到低","排序数":"1","筛选条件":{"注册地点":"深圳"}}'),
    ('''2019-2021年哪些家上市公司货币总额均位列前十？'''\n提取上述句子中的年份，关键词，排序方向，排序数，筛选条件，并按照json输出。",
     '{"年份":[2019,2020,2021],"关键词":["货币总额"],"排序方向":"从高到低","排序数":"10","筛选条件":{}}'),
    ('''2020年总负债最高和营业利润最高的公司分别是？'''\n提取上述句子中的年份，关键词，排序方向，排序数，筛选条件，并按照json输出。",
     '{"年份":[2020],"关键词":["总负债","营业利润"],"排序方向":"从高到低","排序数":"1","筛选条件":{}}'),
    ('''2020年其他非流动资产最高并且历史注册地址在青岛的上市公司是？金额是？'''\n提取上述句子中的年份，关键词，排序方向，排序数，筛选条件，并按
     '{"年份":[2020],"关键词":["其他非流动资产"],"排序方向":"从高到低","排序数":"1","筛选条件":{"历史注册地址":"青岛"}}'),
    ('''2020年营业总收入最高的7家并且曾经在武汉注册的上市公司是？金额是？'''\n提取上述句子中的年份，关键词，排序方向，排序数，筛选条件，并按照
     '{"年份":[2020],"关键词":["营业总收入"],"排序方向":"从高到低","排序数":"1","筛选条件":{"曾经注册地址":"武汉"}}'),
    ('''2021年哪三家上市公司，在重庆注册，营业收入最高？金额为？'''\n提取上述句子中的年份，关键词，排序方向，排序数，筛选条件，并按照json输出。
     '{"年份":[2021],"关键词":["营业收入"],"排序方向":"从高到低","排序数":"3","筛选条件":{"注册地址":"重庆"}}')
]
question = '注册地址在深圳的上市公司中，2021年货币总额最高的公司为？'
prompt = f'{question}\n\n提取上述句子中的年份，关键词，排序方向，排序数，筛选条件，并按照json输出。'
print('优化做法：',search_chatglm(prompt,cls_history)['response'])
```

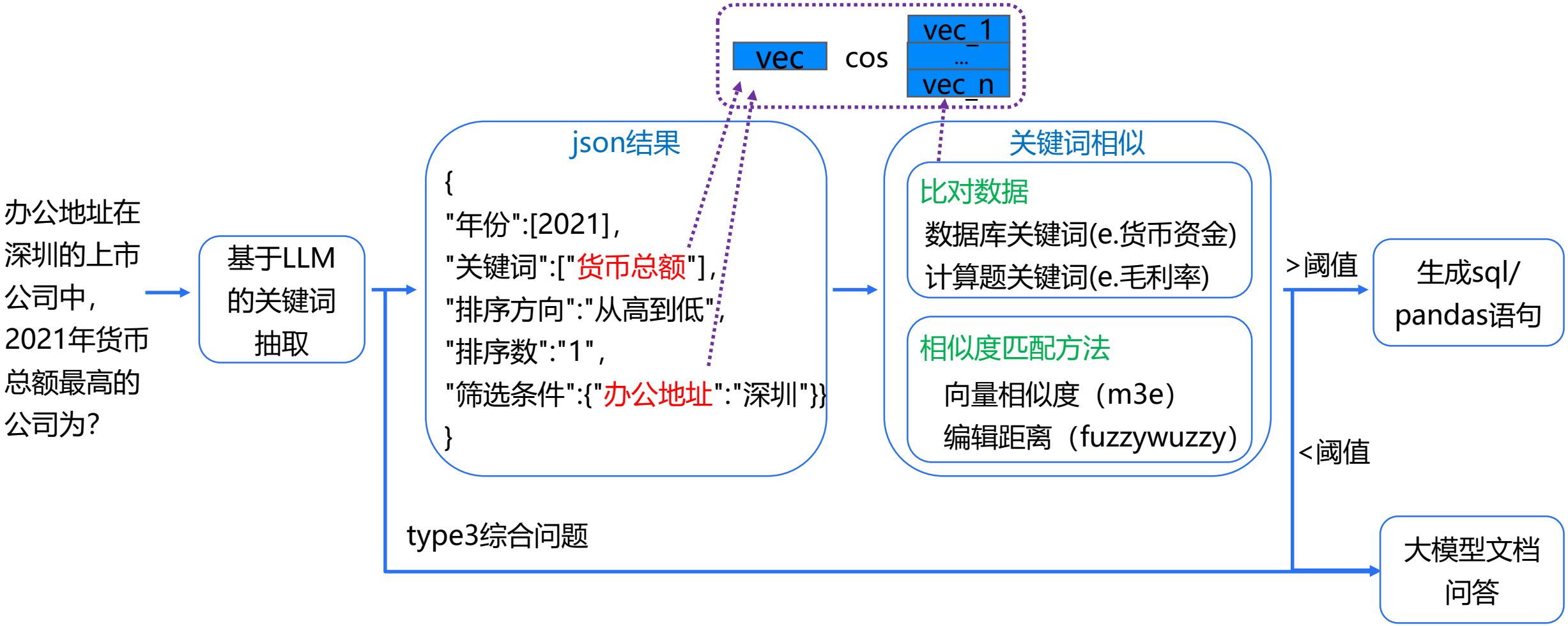
优化做法： {"年份":[2021],"关键词":["货币总额"],"排序方向":"从高到低","排序数":"1","筛选条件":{"注册地点":"深圳"}}

输出结果：

{"年份":[2021],"关键词":["货币总额"],"排序方向":"从高到低","排序数":"1","筛选条件":{"注册地址":"深圳"}}

模拟对话可以帮助LLM理解问题和回答间的关系，更好地学习预期的输出格式

基于LLM的关键词抽取



基于LLM的关键词抽取

效果

```
In [10]: process_question(question_obj)
LLM抽取关键词生成json结果
{"关键词":["职工总人员数"]}
问题
2019年上海大屯能源股份有限公司职工总人员数有多少？
答案:
上海大屯能源股份有限公司在2019年的职工总人员数(职工人数)为13989人。
```

基本数值题

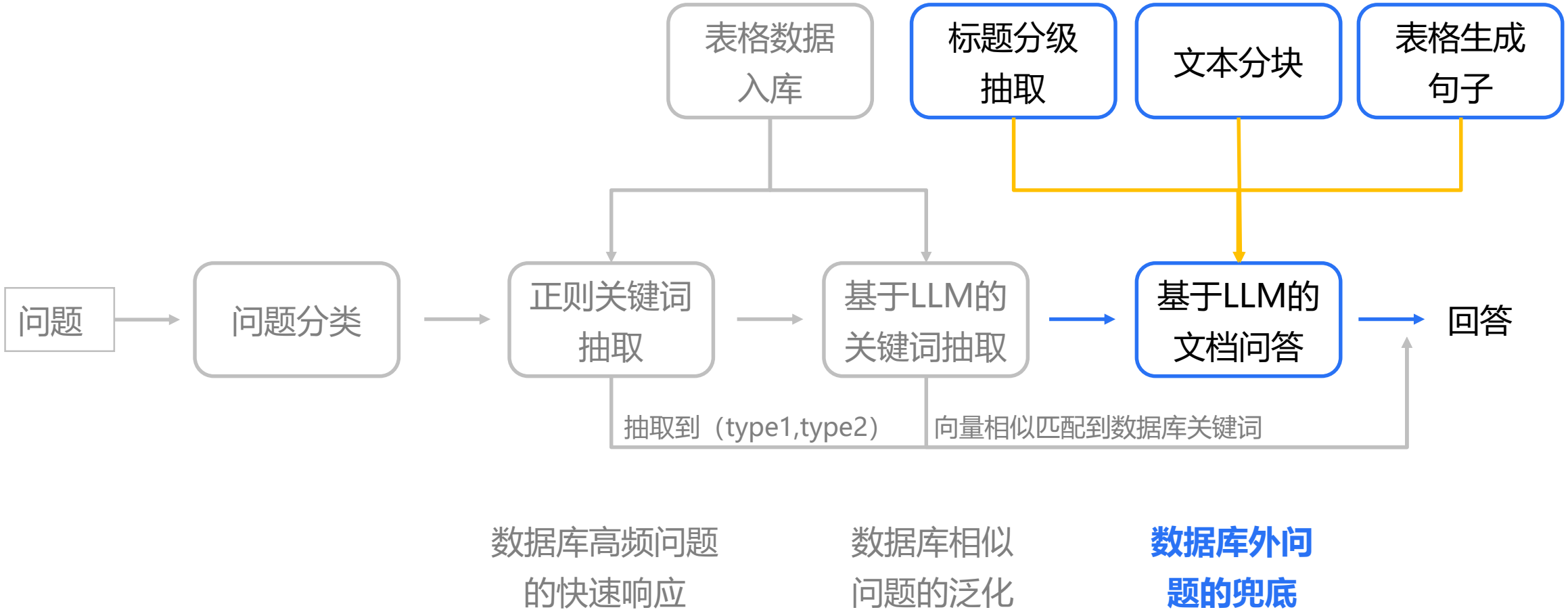
```
In [7]:  process_question(question_obj)
LLM抽取关键词生成json结果
{"年份":[2020],"关键词":["研发人员与职工人数之比"],"排序方向":"从高到低","排序数":"1","筛选条件":{}}
[('研发人员占职工', 77)]
问题
在保留两位小数的情况下，请计算出时代新材2020年的研发人员与职工人数之比
答案:
时代新材2020年研发人员为1194人，2020年职工人数为6244人，根据公式:研发人员占职工=研发人员 / 职工人数，得出研发人员占职工为0.19。
```

计算题

```
In [3]:  process_question(question_obj)
LLM抽取关键词生成json结果
{"年份":[2021],"关键词":["货币总计"],"排序方向":"从高到低","排序数":"3","筛选条件":{"注册地":"深圳"}}
问题
2021年注册地为深圳的货币总计最高的前3家上市公司为?
答案:
2021年注册地为深圳的货币总计最高的前3家上市公司为？在2021年，1、欣旺达电子股份有限公司；2、中集车辆集团股份有限公司；3、深圳市赢合科技股份有限公司；
```

统计题

基于LLM的文档问答

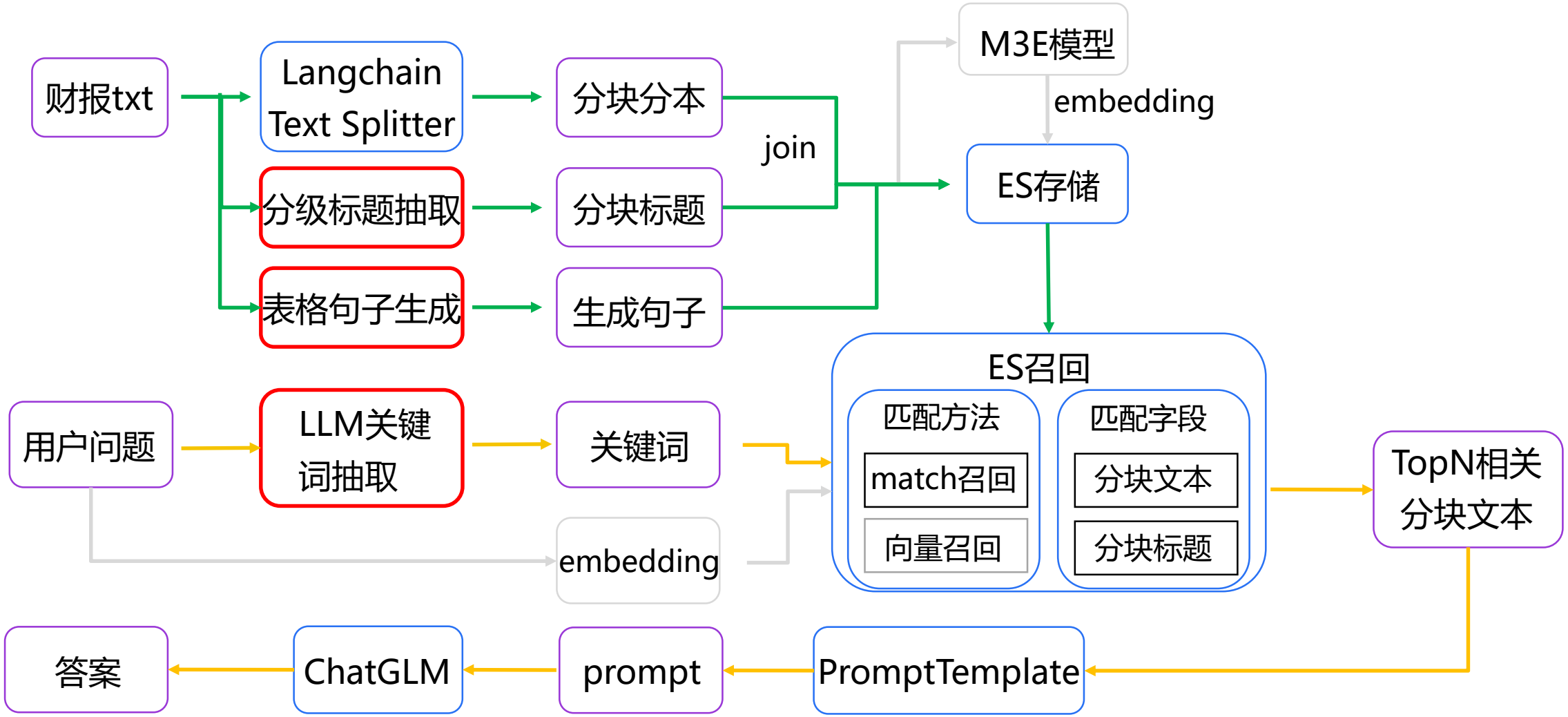


数据库高频问题的快速响应

数据库相似问题的泛化

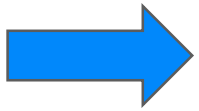
数据库外问题的兜底

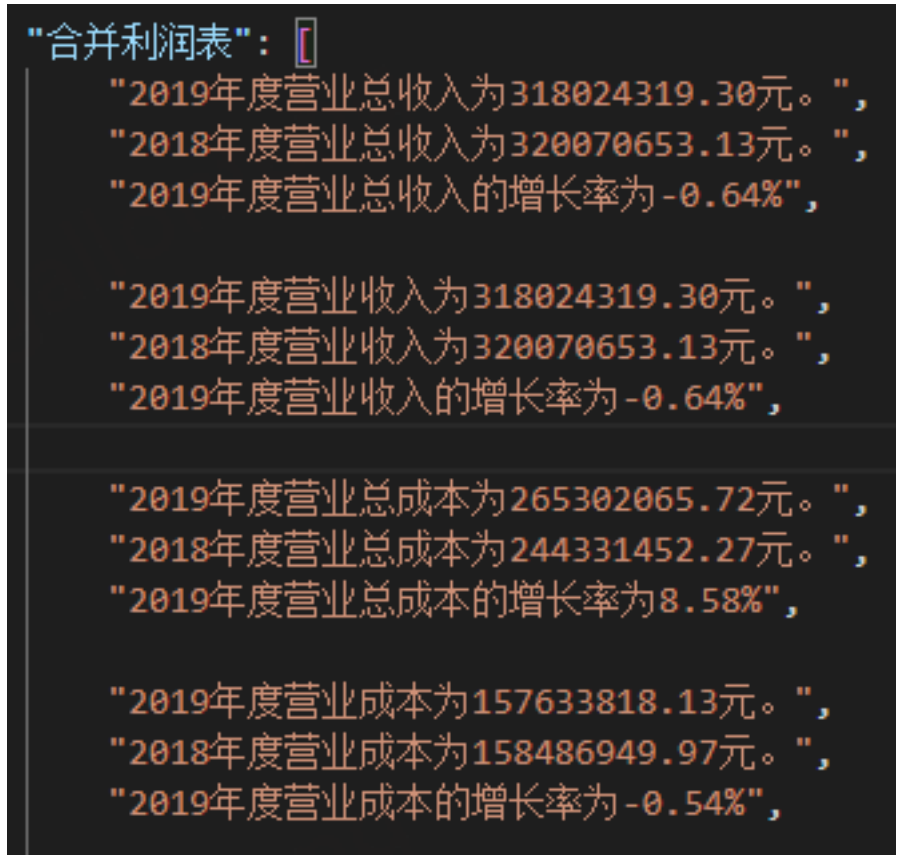
基于LLM的文档问答--框架图



基于LLM的文档问答--表格生成句子

通过找到左、上表头生成句子





优点：

1、***解决表格数据难召回以及LLM理解表格困难的问题；***

2、***可兜底回答数据库没有抽取的表格字段信息***

基于LLM的文档问答--基本数值问题

问题：2021年安硕信息的不能重分类进损益的其他综合收益是多少元?

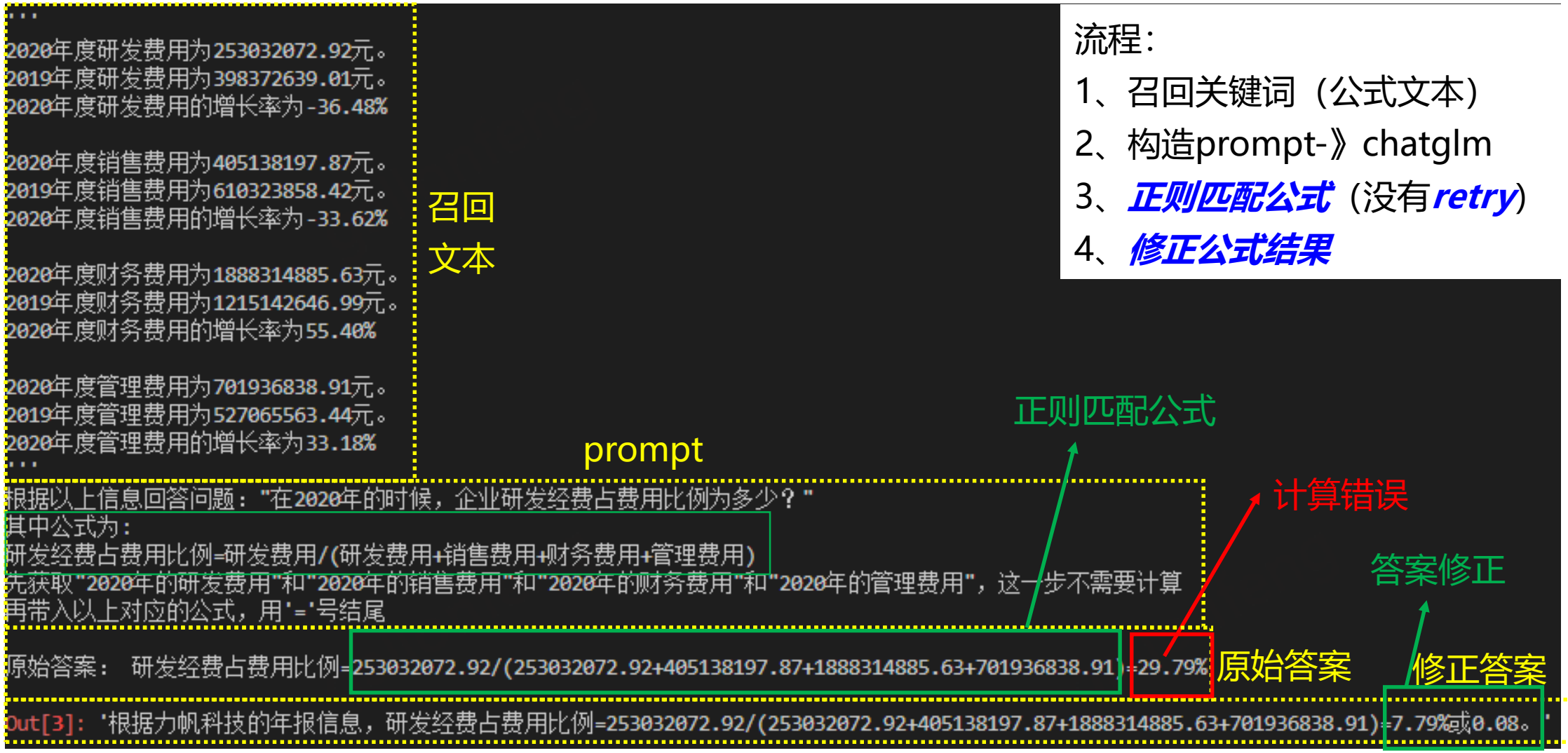
答案：根据安硕信息的年报信息，根据提供的信息，2021年度不能重分类进损益的其他综合收益为454987.07元

```
'''
2021年度不能重分类进损益的其他综合收益为454987.07元。
2020年度不能重分类进损益的其他综合收益为-107727.15元。
2021年度不能重分类进损益的其他综合收益的增长率为-522.35%

财务报告
注册会计师对财务报表审计的责任
就贵公司中实体或业务活动的财务信息获取充分、恰当的审计证据，以对财务报表发表审计意见。我们负责指导、监督和执
母公司利润表
['（二）终止经营净利润（净亏损以“-”号填列）', '', '']
['五、其他综合收益的税后净额', '455222.30', '']
['（一）不能重分类进损益的其他综合收益', '455222.30', '']
['1.重新计量设定受益计划变动额', '', '']
['2.权益法下不能转损益的其他综合收益', '', '']
['3.其他权益工具投资公允价值变动', '455222.30', '']
['4.企业自身信用风险公允价值变动', '', '']
['5.其他', '', '']
['（二）将重分类进损益的其他综合收益', '', '']
['1.权益法下可转损益的其他综合收益', '', '']
['2.其他债权投资公允价值变动', '', '']
['3.金融资产重分类计入其他综合收益的金额', '', '']
['4.其他债权投资信用减值准备', '', '']

财务报告
合并财务报表项目注释
其他综合收益
单位：元
['', '期初余额', '本期发生额本期所得税前发生额', '减：前期计入其他综合收益当期转入损益', '减：前期计入其他综
['项目', '', '', '', '', '', '', '', '']
['一、不能重分类进损益的其他综', '-107727.15', '535242.0', '', '', '80254.94', '454987.0', '', '347259.']
'''
根据以上信息回答问题：'2021的不能重分类进损益的其他综合收益是多少元？'，不需要计算
1148
安硕信息2021的不能重分类进损益的其他综合收益是多少元？
根据安硕信息的年报信息，根据提供的信息，2021年度不能重分类进损益的其他综合收益为454987.07元。
```

召回文本+prompt

答案

基于LLM的文档问答--计算题（给定公式）



流程：

1、召回关键词（公式文本）

2、构造prompt-》chatglm

3、***正则匹配公式***（没有***retry***）

4、***修正公式结果***

基于LLM的文档问答--计算题（不给公式）

召回文本

```
'''
2019年度销售费用为57094406.01元。
2018年度销售费用为53467171.42元。
2019年度销售费用的增长率为6.78%

2019年度管理费用为317524853.89元。
2018年度管理费用为308668321.54元。
2019年度管理费用的增长率为2.87%

2019年度财务费用为225064365.66元。
2018年度财务费用为114401623.38元。
2019年度财务费用的增长率为96.73%
'''
```

问题：通富微电2019年的销售费用、管理费用和财务费用三者互相占比是多少?

不提供公式，也能兜底回答简单的计算问题

prompt

```
先获取"2019年的销售费用"和"2019年的管理费用"和"2019年的财务费用"，这一步不需要计算
再根据以上信息回答问题："在2019年的（销售费用、管理费用和财务费用）三者相互占比是多少？请保留至小数点后两位。"
用'='号结尾
```

答案修正

9.52%或0.10

52.95%或0.53

37.53%或0.38

原始答案

公式正则

```
原始答案：  在2019年的销售费用、管理费用和财务费用三者相互占比是：
```

```
销售费用占比 = 销售费用 / (销售费用 + 管理费用 + 财务费用) = 57094406.01 / (57094406.01 + 317524853.89 + 225064365.66) ≈ 0.4972
管理费用占比 = 管理费用 / (销售费用 + 管理费用 + 财务费用) = 317524853.89 / (57094406.01 + 317524853.89 + 225064365.66) ≈ 0.4784
财务费用占比 = 财务费用 / (销售费用 + 管理费用 + 财务费用) = 225064365.66 / (57094406.01 + 317524853.89 + 225064365.66) ≈ 0.4719
```

基于LLM的文档问答--公司综合问题

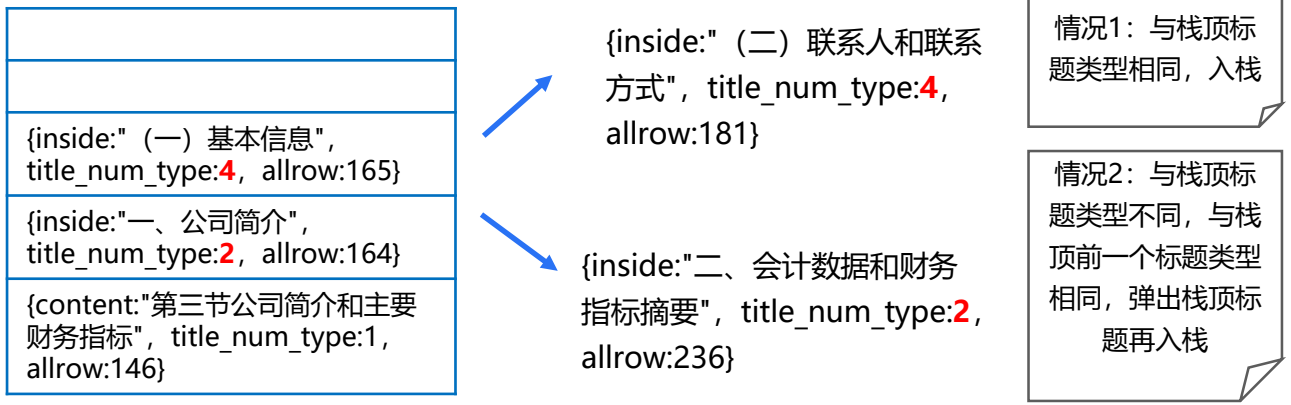
分级标题抽取

1、用正则表达式识别标题行并用序号区分类型

2、用堆栈来储存标题分层递归关系，记录行号

3、根据分层递归的标题层级，对正文进行初步切割

第一节 重要提示、目录和释义 → 标题类型1
一、公司简介 → 标题类型2
1、业务发展情况 → 标题类型3
(一)主要会计数据和财务指标 → 标题类型4
(1)房地产开发业务 → 标题类型5



{
"公司简介和主要财务指标——>公司简介——>基本信息"：":
"中文名称：万科企业股份有限公司（缩写为“万科”）
英文名称：CHINAVANKECO.,LTD.(缩写为“VANKE”）
注册地址：中国深圳市盐田区大梅沙环梅路33号万科中心。。。。。。"
}

基于LLM的文档问答--公司综合问题

文本分块

第三节 管理层讨论与分析

一、报告期内公司所处行业情况

公司需遵守《深圳证券交易所上市公司自律监管指引第 3 号——行业信息披露》中的“软件与信息技术服务业”的披露要求

2022 年 1 月 4 日，央行印发《金融科技发展规划（2022-2025 年）》（下称《规划》），提出将以加强金融数据要素应用为基础，以深化金融供给侧结构性改革为目标，以加快金融机构数字化转型、强化金融科技审慎监管为主线，将数字元素注入金融服务全流程，将数字思维贯穿业务运营全链条，注重金融创新的科技驱动和数据赋能，力争到 2025 年实现整体水平与核心竞争力跨越式提升。

与《金融科技发展规划（2019-2021 年）》相比，本次《规划》更加强调数据应用和金融科技应用两方面内容。数据应用层面来看，《规划》提出要加强数据能力建设、建设绿色高可用数据中心，既要在保障安全和隐私前提下推动数据有序共享与综合应用，充分激活数据要素潜能，提升金融服务质效，也要架设安全泛在的金融网络，布局先进高效的算力体系，进一步夯实金融创新发展的“数字底座”。金融科技应用层面来看，《规划》强调要深化数字技术金融应用，健全安全与效率并重的科技成果应用体制机制，为人民群众提供更加普惠、绿色、人性化的数字金融服务。

银行业是金融领域与信息技术深度融合的重要组成部分。科技从为金融机构业务提供支撑的配角，转为引领业务转型发展的主力。金融科技通过将业务和相关流程进行数据化、自动化和智能化升级，大大增强了金融机构的运营能力和效率。同时，政府也在陆续出台多项政策、规范，鼓励金融机构数字化转型。

公司自成立以来长期服务银行等金融机构，专注信贷风险管理领域，为银行信贷风险业务信息化系统提供一体化解决方案，在行业内有一定的口碑和影响力。随着行业加快数字化转型，公司创新及研发的新方案、新技术、新产品具有更好的市场空间和应用前景。公司的新零售解决方案、分布式微服务架构方案、征信及大数据内容服务体系等将有效帮助银行等金融机构加快数字化转型，为客户增强运营能力、提高运营效率，也为将公司赢得业务规模扩大的发展机会。

截止2021年底，公司已经与4家大型国有银行（共6家），11家股份制银行（共12家），101家城市商业银行（共129家），13家资产规模2000亿以上农村商业银行（共17家），14家民营银行（共19家），10家外资、港资、台资银行，以及7家省级农村信用社联合社等银行金融机构展开合作。另外公司服务了大量的村镇银行、农村金融机构、资产管理公司、保险公司、信托公司、证券公司、消费金融公司、供应链金融公司、融资租赁公司、小额贷款公司等机构。

给分块文本添加标题信息

分块1

LangChain RecursiveCharacterTextSplitter(chunk_size=500)

根据'\n', '\n\n', '。'等符号分割

分块2——（缺少标题信息）

分块3——（缺少标题信息）

询问"行业情况"时会召回不到该文本

基于LLM的文档问答

分块文本添加标题效果

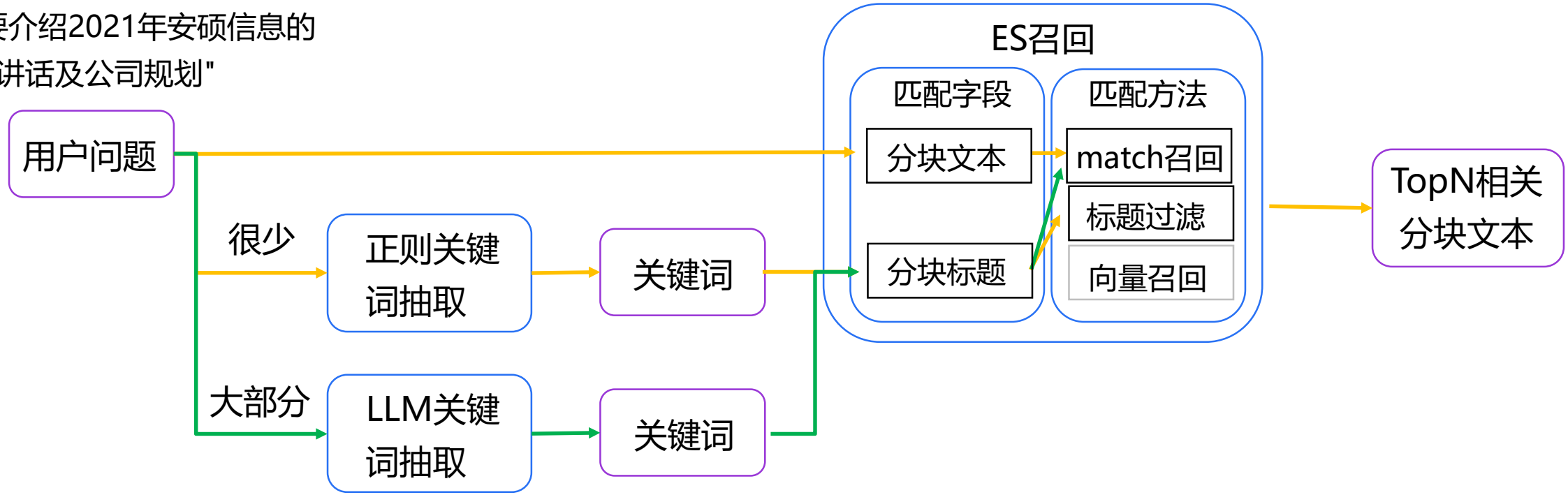
```
'''
管理层讨论与分析
报告期内公司所处行业情况
公司需遵守《深圳证券交易所上市公司自律监管指引第3号——行业信息披露》中的“软件与信息技术服务业”的披露要求2022年1月4日，央行印发《金融科技发展规划（2022-2025年）》（下称《规划》），提出将以加强金融数据要素应用为基础，以深化金融供给侧结构性改革为目标，以加快金融机构数字化转型、强化金融科技审慎监管为主线，将数字元素注入金融服务全流程，将数字思维贯穿业务运营全链条，注重金融创新的科技驱动和数据赋能，力争到2025年实现整体水平与核心竞争力跨越式提升。
与《金融科技发展规划（2019-2021年）》相比，本次《规划》更加强调数据应用和金融科技应用两方面内容。数据应用层面来看，《规划》提出要加强数据能力建设、建设绿色高可用数据中心，既要在保障安全和隐私前提下推动数据有序共享与综合应用，充分激活数据要素潜能，提升金融服务质效，也要架设安全泛在的金融网络，布局先进高效的算力体系，进一步夯实金融创新发展的“数字底座”。金融科技应用层面来看，《规划》强调要深化数字技术金融应用，健全安全与效率并重的科技成果应用体制机制，为人民群众提供更加普惠、绿色、人性化的数字金融服务。

管理层讨论与分析
报告期内公司所处行业情况
截止2021年底，公司已经与4家大型国有银行（共6家），11家股份制银行（共12家），101家城市商业银行（共129家），13家资产规模2000亿以上农村商业银行（共17家），14家民营银行（共19家），10家外资、港资、台资银行，以及7家省级农村信用社联合社等银行金融机构展开合作。另外公司服务了大量的村镇银行、农村金融机构、资产管理公司、保险公司、信托公司、证券公司、消费金融公司、供应链金融公司、融资租赁公司、小额贷款公司等机构。

管理层讨论与分析
报告期内公司所处行业情况
银行业是金融领域与信息技术深度融合的重要组成部分。科技从为金融机构业务提供支撑的配角，转为引领业务转型发展的主力。金融科技通过将业务和相关流程进行数据化、自动化和智能化升级，大大增强了金融机构的运营能力和效率。同时，政府也在陆续出台多项政策、规范，鼓励金融机构数字化转型。
公司自成立以来长期服务银行等金融机构，专注信贷风险管理领域，为银行信贷风险业务信息化系统提供一体化解决方案，在行业内有一定的口碑和影响力。随着行业加快数字化转型，公司创新及研发的新方案、新技术、新产品具有更好的市场空间和应用前景。公司的新零售解决方案、分布式微服务架构方案、征信及大数据内容服务体系等将有效帮助银行等金融机构加快数字化转型，为客户增强运营能力、提高运营效率，也为将公司赢得业务规模扩大的发展机会。
'''
根据以上信息回答问题：'请简要介绍一下2021的公司所处行业情况'

答案:
根据安硕信息的年报信息，2021年,公司所处的行业是金融科技领域,具体来说,是软件与信息技术服务业中的金融科技应用方向。该行业在报告期内受到了政府的大力支持和鼓励,因为政府出台了多项政策,鼓励金融机构数字化转型,并加强金融数据要素应用。公司作为专注于信贷风险管理领域的金融科技企业,在这一行业中具有一定的口碑和影响力。同时,公司的新零售解决方案、分布式微服务架构方案、征信及大数据内容服务体系等将有效帮助银行等金融机构加快数字化转型,为客户增强运营能力、提高运营效率,也为将公司赢得业务规模扩大的发展机会。
```

问题：请简要介绍一下安硕信息2021年的公司所处行业情况

基于LLM的文档问答--公司综合问题

llm抽取关键词的召回增强

"请简要介绍2021年安硕信息的
董事长讲话及公司规划"



{'关键词': ['董事长讲话', '公司规划']}

基于LLM的文档问答--公司综合问题

不抽取关键词 和 抽取关键词 召回的效果对比

```
'''
公司治理
董事、监事和高级管理人员情况
王和忠先生：财务负责人兼董事会秘书，简历详见本节董事会成员介绍。
['', '', '监事', '2016年07月08日', '', '']
['王和忠', '上海安硕首道信息服务有限公司', '', '', '', '否']
['', '', '监事', '2016年06月29日', '', '']
['王和忠', '江苏兀峰信息科技有限公司', '', '', '', '否']
['', '', '监事', '2021年03月02日', '', '']
['王和忠', '安硕致鑫（重庆）信息科技有限公司', '', '', '', '']
['', '', '监事', '2014年11月28日', '', '']
['魏治毅', '北京安硕信息技术有限公司', '', '', '', '否']
['', '', '执行董事', '2020年01月05日', '', '']
['张怀', '上海璋湃企业咨询管理有限公司', '', '', '', '否']
['张怀', '西昌安硕易民互联网金融服务股份有限公司', '监事', '2015年04月10日', '', '否']

管理层讨论与分析
主营业务分析
收入与成本
营业成本构成
行业和产品分类
单位：元
['', '项目', '2021年', '', '2020年', '', '同比增减']
['行业分类', '', '金额', '占营业成本比重', '金额', '占营业成本比重', '']
['信息服务业', '', '499503404.13', '100.00%', '392458263.98', '100.00%', '27.28%']
['信息服务业', '', '499503404.13', '100.00%', '392458263.98', '100.00%', '27.28%']
单位：元
['', '项目', '2021年', '', '2020年', '', '同比增减']
['产品分类', '', '金额', '占营业成本比重', '金额', '占营业成本比重', '']
['信贷管理类系统', '', '370747578.15', '74.22%', '313026134.79', '79.76%', '18.44%']
['风险管理类系统', '', '33955681.46', '6.80%', '17264489.17', '4.40%', '96.68%']

公司治理
董事、监事和高级管理人员情况
王和忠先生：财务负责人兼董事会秘书，简历详见本节董事会成员介绍。
['', '', '执行董事', '2015年11月17日', '']
['高勇', '上海安硕益盛商务咨询有限公司', '', '', '否']
['', '', '执行董事', '2016年05月19日', '']
['高勇', '上海泥鸽信息科技有限公司', '', '', '否']
['', '', '执行董事', '2015年07月30日', '']
['高勇', '上海安硕金融信息服务有限公司', '', '', '否']
['', '', '董事长', '2019年01月18日', '']
['高勇', '上海懿量网络科技有限公司', '', '', '否']
['', '', '董事', '2003年04月29日', '']
['高勇', '上海安硕计算机系统集成有限公司', '', '', '否']
['', '', '董事', '2015年12月04日', '']
['高勇', '贵州安硕金融大数据服务有限公司', '', '', '否']
['', '', '董事', '2007年01月15日', '']
'''
根据以上信息回答问题：'请简要介绍2021年的产品信息'

答案：
根据安硕信息的年报信息，2021年,公司主要生产和销售信贷管理类系统和风险管理类系统,其中信贷管理类系统收入占比较高。具体来说,2021年公司的信贷管理类系统收入为370747578.15元,占营业成本比重为74.22%;风险管理类系统收入为33955681.46元,占营业成本比重为6.80%。
```

召回错误文本，导致回答错误

问题：请简要介绍一下安硕信息2021年公司的产品信息

基于LLM的文档问答--公司综合问题

不抽取关键词 和 抽取关键词 召回的效果对比

```
{'关键词': ['产品信息']}
'''
财务报告
其他信息
贵公司管理层（以下简称管理层）对其他信息负责。其他信息包括贵公司2021年年度报告中涵盖的信息，但不包括财务报表和我们的审计报告。
我们对财务报表发表的审计意见不涵盖其他信息，我们也不对其他信息发表任何形式的鉴证结论。
结合我们对财务报表的审计，我们的责任是阅读其他信息，在此过程中，考虑其他信息是否与财务报表或我们在审计过程中了解到的情况存在重大不一致或者似乎存在重大错报。
基于我们对审计报告日前获取的其他信息已执行的工作，如果我们确定其他信息存在重大错报，我们应当报告该事实。在这方面，我们无任何事项需要报告。
我们阅读2021年年度报告后，如果确定其中存在重大错报，审计准则要求我们与治理层沟通该事项并采取适当措施。

管理层讨论与分析
主营业务及产品
报告期内公司的主要业务仍然是向以银行为主的客户提供信贷风险业务管理咨询、软件开发与服务，产品线主要是银行信贷管理系统、银行风险管理系统、商业智能与数据仓库、非银行金融机构及其他系统。其他系统含监管报送领域、融资租赁
领域、非银行资产管理领域等一系列解决方案。创新业务方面，征信大数据业务发展迅速，客户数量和订单数量快速增多，业务初具规模。
公司通过招投标或协议销售方式获取项目。公司服务模式分为三种，一是根据客户需求在公司产品原型基础上定制化开发项目，二是按照人月计价模式提供开发服务，三是销售标准化产品和服务。
公司主要产品介绍如下：（1）信贷管理系统是公司的核心产品，由众多独立应用产品组成，以数据完备性为基础、以流程管理及信贷风险防范为主线，经过十多年的技术积累沉淀和更新换代，已经成长为一套成熟、先进的金融管理系统，为公
司带来稳定的业绩收入。（2）风险管理系统借鉴国内外先进银行在风险管理领域的成熟经验，融合多种风控手段，为客户及时识别、分析各类风险因素，在信用关系复杂的经济环境下帮助客户有效规避各种潜在风险。（3）数据仓库和商业智

公司治理
董事、监事和高级管理人员情况
王和忠先生：财务负责人兼董事会秘书，简历详见本节董事会成员介绍。
['', '', '监事', '2016年07月08日', '', '']
['王和忠', '上海安硕首道信息服务有限公司', '', '', '', '否']
['', '', '监事', '2016年06月29日', '', '']
['王和忠', '江苏兀峰信息科技有限公司', '', '', '', '否']
['', '', '监事', '2021年03月02日', '', '']
['王和忠', '安硕致鑫（重庆）信息科技有限公司', '', '', '', '']
['', '', '监事', '2014年11月28日', '', '']
['魏治毅', '北京安硕信息技术有限公司', '', '', '', '否']
['', '', '执行董事', '2020年01月05日', '', '']
['张怀', '上海璋湃企业咨询管理有限公司', '', '', '', '否']
['张怀', '西昌安硕易民互联网金融服务股份有限公司', '监事', '2015年04月10日', '', '否']
'''
根据以上信息回答问题：'请简要介绍2021年的产品信息'

答案：
根据安硕信息的年报信息，根据所提供的信息，2021年公司的主要产品包括信贷管理系统、风险管理系统、数据仓库和商业智能与数据仓库。其中，信贷管理系统是公司的核心产品，由众多独立应用产品组成，以数据完备性为基础、以流程管理
及信贷风险防范为主线，经过十多年的技术积累沉淀和更新换代，已经成长为一套成熟、先进的金融管理系统，为公司带来稳定的业绩收入。风险管理系统借鉴国内外先进银行在风险管理领域的成熟经验，融合多种风控手段，为客户及时识别、
分析各类风险因素，在信用关系复杂的经济环境下帮助客户有效规避各种潜在风险。数据仓库和商业智能系统为公司提供了更高效、更精准的数据管理和分析服务。
```

LLM正确抽取到关键词"产品信息"

正确召回文本，回答正确

总结与展望

总结：

1、基于正则分类与抽取关键词的方式，实现数据库字段高频问题的快速回答

2、基于LLM抽取关键词的方式，实现与数据库字段相似问题的泛化

3、基于LLM文档问答的方式，实现数据库字段外问题的兜底

亮点：

1、使用In-Context Learning的方式抽取关键词，无需微调，保留大模型的通用能力

2、通过分块文本增加标题信息，以及LLM关键词的召回增强的方法，显著提升回答公司综合问题的效果

展望：

1、可以做更多向量模型的选型，如 bge，stella等

2、可以在文本召回的时候做更多向量召回的优化

3、探索text2sql来做计算题的能力

4、分类算法的优化


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

Thanks