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

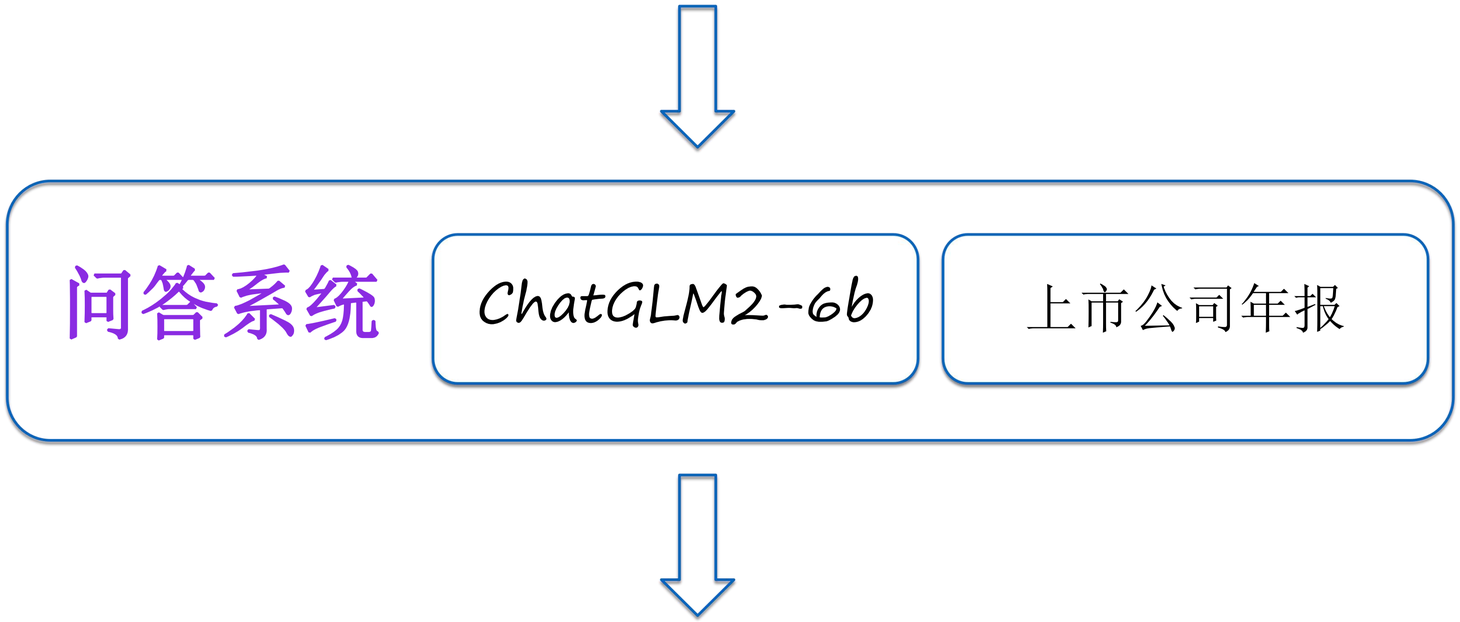
ChatGLM 金融大模型挑战赛

汇报人：刘幼峰

队伍：结婚买房代代韭菜

任务描述

query：2019年中国工商银行财务费用是多少元?



answer：2019年中国工商银行财务费用是12345678.9元。

赛题分析及评测标准

query type	query example
type1	2020 年一品红药业股份有限公司其他非流动金融资产是多少元？
type2	2019 年贵研铂业股份有限公司速动比率为多少？保留两位小数。
type31	根据 2019 年的年报数据，哈空调研发投入的情况，请做简要分析。
type32	什么是其他债权投资？

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

赛题难点

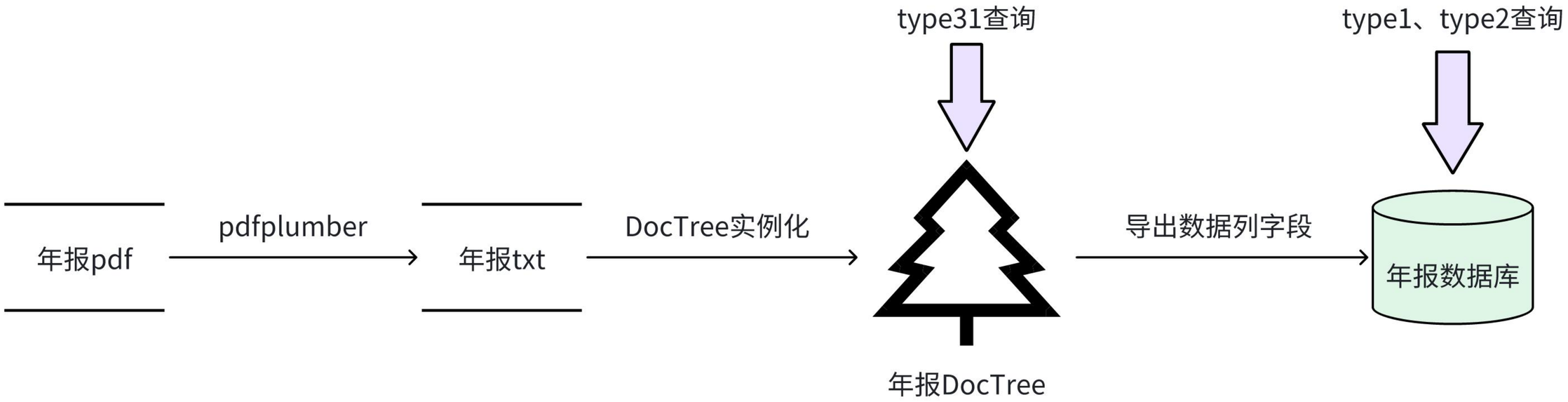
query: 2021年亚普股份的法定代表人与上一年是否相同？

厦门延江新材料股份有限公司2019年的其他非流动资产为多少元？
2020年唐山三孚硅业股份有限公司总资产增长率为多少？保留两位小数。
2021年注册地址在武汉哪三个上市公司的负债总额最高？金额是？
请具体描述一下2019年润禾材料研发投入的情况。

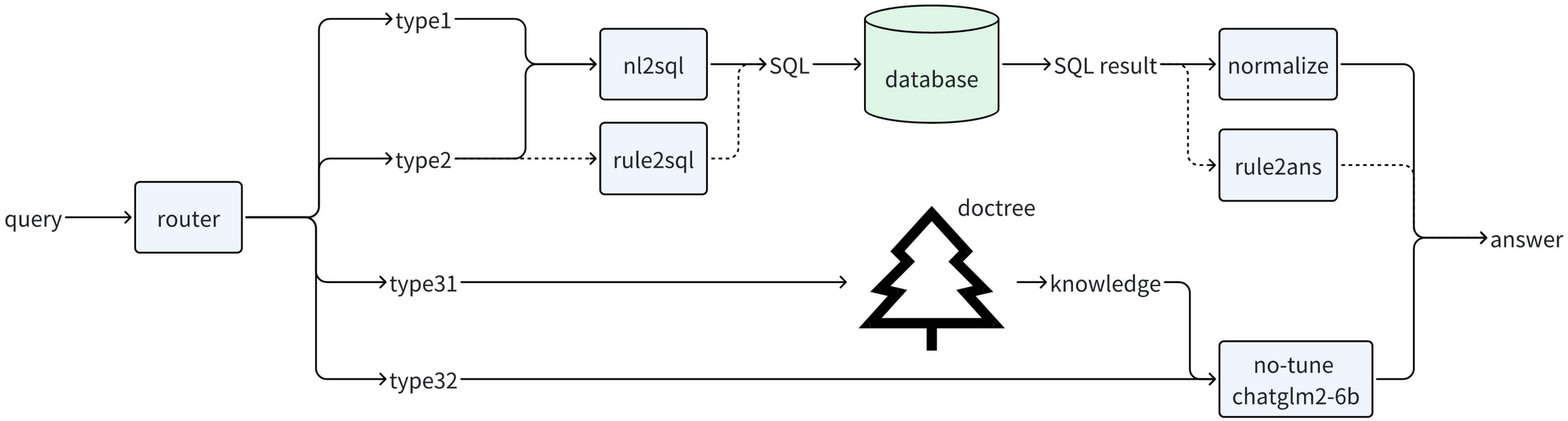
answer: 亚普股份2020年的法定代表人是郝建，2021年的法定代表人是姜林，所以答案是不同。

1. 如何理解问题？(需要哪些数据)
2. 如何找到问题需要的数据？
3. 如何规范地回答问题？

方案架构-数据处理


type31查询
type1、type2查询
年报pdf
pdfplumber
年报txt
DocTree实例化
导出数据列字段
年报数据库
年报DocTree

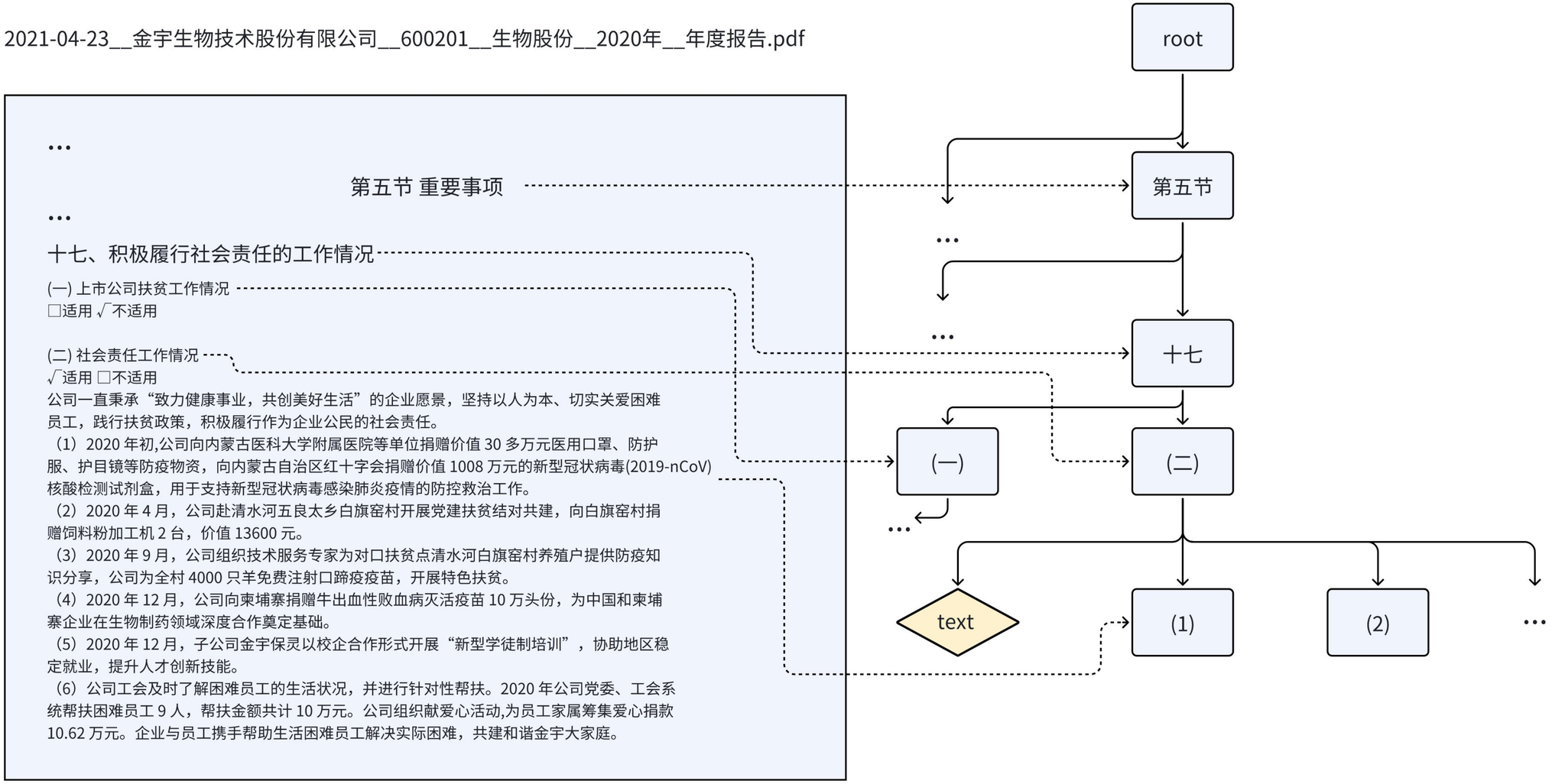
方案架构-问答系统



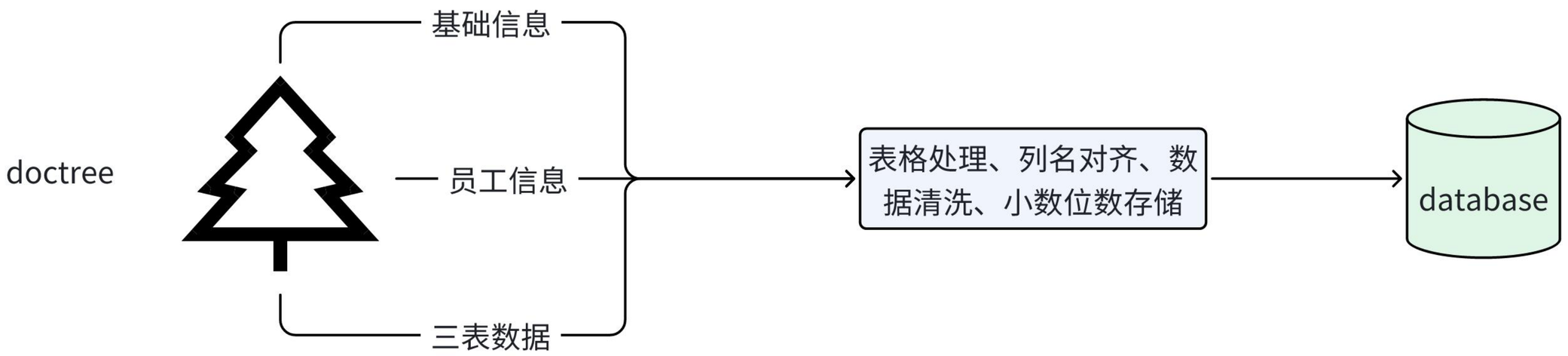
router、nl2sql、normalize均为使用ptuningv2的微调模型

数据结构-DocTree

2021-04-23__金宇生物技术股份有限公司__600201__生物股份__2020年__年度报告.pdf



数据结构-Database



微调模型-router

why router?

不同类型的问题需要的数据不一样，处理流程也不同

1. type1需要具体的数据
2. type2需要具体的数据，还可能需要公式
3. type31需要相关文本段落
4. type32不需要年报相关的文本输入

微调模型-router

how to?

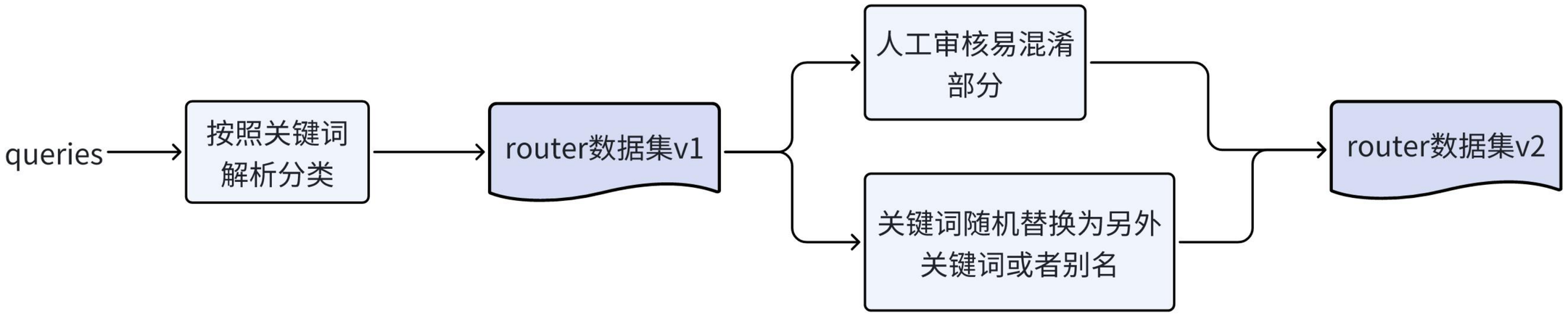
年份　　公司名称　　年报关键词

2020 年一品红药业股份有限公司其他非流动金融资产是多少元？

年份	公司名称	年报关键词	类型分析
√	√	√	type11或者type2。根据年报关键词是否为公式可以区分type1和type2
√	√	×	type31，小概率为type1或者type2（关键词不在词表内）
√	×	√	type12，小概率数据库范围外
×	√	√	type12
√	×	×	type12或者type31
×	×	√	type32
×	√	×	type12，关键词不在词表内
×	×	×	type32

微调模型-router

how to?



效果：B榜2k条，正确率99.9%

微调模型-nl2sql

why nl2sql?

厦门延江新材料股份有限公司2019年的其他非流动资产为多少元？

2020年唐山三孚硅业股份有限公司总资产增长率为多少？保留两位小数。

2021年注册地址在武汉哪三个上市公司的负债总额最高？金额是？

2021年亚普股份的法定代表人与上一年是否相同？

2021年连云港的法定代表人与前年是否相同？

1. 问题形态多变，需要理解问题意图的统一范式
2. 统计类问题SQL查询能很好完成

微调模型-nl2sql

how to?

prompt

你的任务是将问题转化为SQL。

1. SQL语句查询的表名为: big

2. 涉及到的列名有: 注册地址,负债合计,预计负债,股票简称,公司名称,股票代码,年份

【问题】哪家上市公司，在北京注册，2020年总负债最高？金额为？

【SQL】

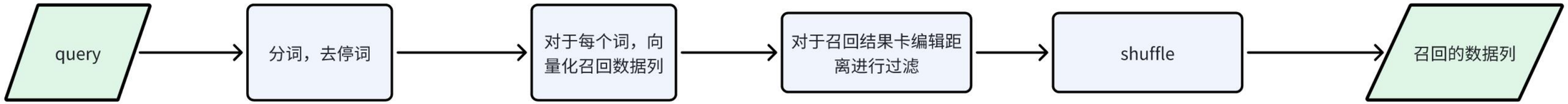
response

```
select 公司名称, 负债合计 from big
where 注册地址 like '%北京%' and 年份 = '2020年'
order by 负债合计 desc limit 1
```

微调模型-nl2sql

columns how to?

1. 分词：jieba
2. 向量模型：GanymedeNil_text2vec-large-chinese



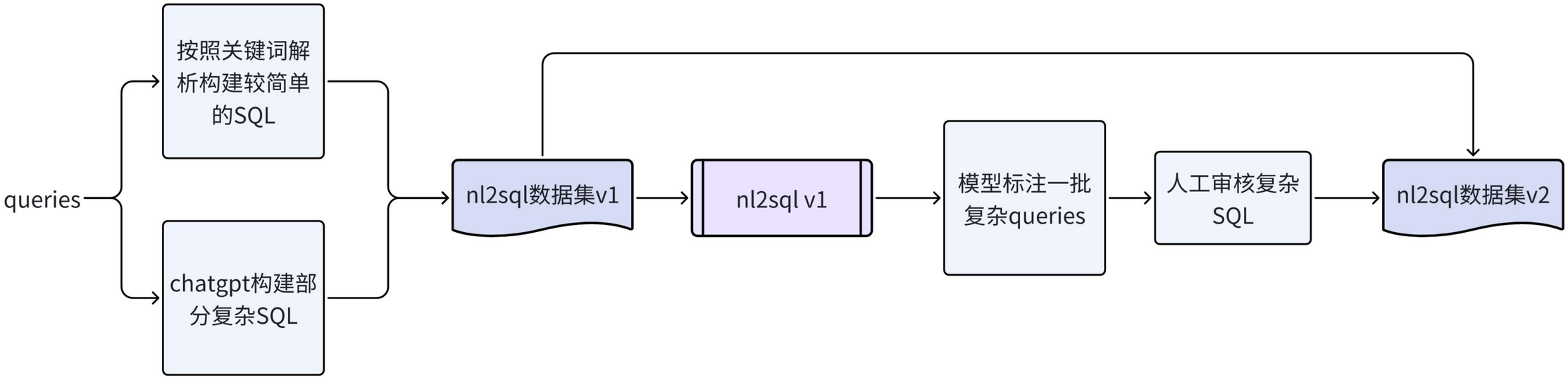
[注册地址]　[负债合计,预计负债]

哪家 上市公司 ， 在 北京 注册 ， ~~2020~~ 年 总负债 最高 ~~?~~ 金额 为 ~~?~~

[ ]　[ ]　[ ]　[ ]

微调模型-nl2sql

how to?



微调模型-nl2sql

tricks

1. 通过解析提取公司名高亮来简化任务

你的任务是将问题转化为SQL。

1. SQL语句查询的表名为: big

2. 涉及到的列名有: 年份,其他非流动资产,其他流动资产,公司名称,股票代码,股票简称

【问题】公司名称为厦门延江新材料股份有限公司的公司2019年的其他非流动资产为多少元?

【SQL】

2. 通过简化部分问题SQL来降低任务难度

query: 开滦股份2019-2020年这两年的法定代表人是否都相同?

sql:

```
select 年份, 法定代表人
from big
where 年份 in ('2019年', '2020年') and 股票简称 in ('开滦股份')
```

query: 请提供太平鸟2020年的无形资产增长率~~并保留2位小数~~

sql:

```
select 年份, 无形资产
from big
where 年份 in ('2020年', '2019年') and 股票简称 in ('太平鸟')
```

效果：抽检type11、type2无问题，type12存在少数错误，主要是对注册地址实体的误判

微调模型-normalize

why normalize?

1. 用户体验，人机接口

- 规范化的回答可以让用户确认自己所有的意图都被理解并执行
- 数据库查询结果不可读，需要以自然语言解释

Q: 2019年中国工商银行财务费用是多少元?

A: 12345678.90元。

Q: 中国工商银行2021年总负债是多少?

A: 中国工商银行2021年负债合计是98765432.10元。

2. 关键词、相似度得分

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

微调模型-normalize

how to?

prompt

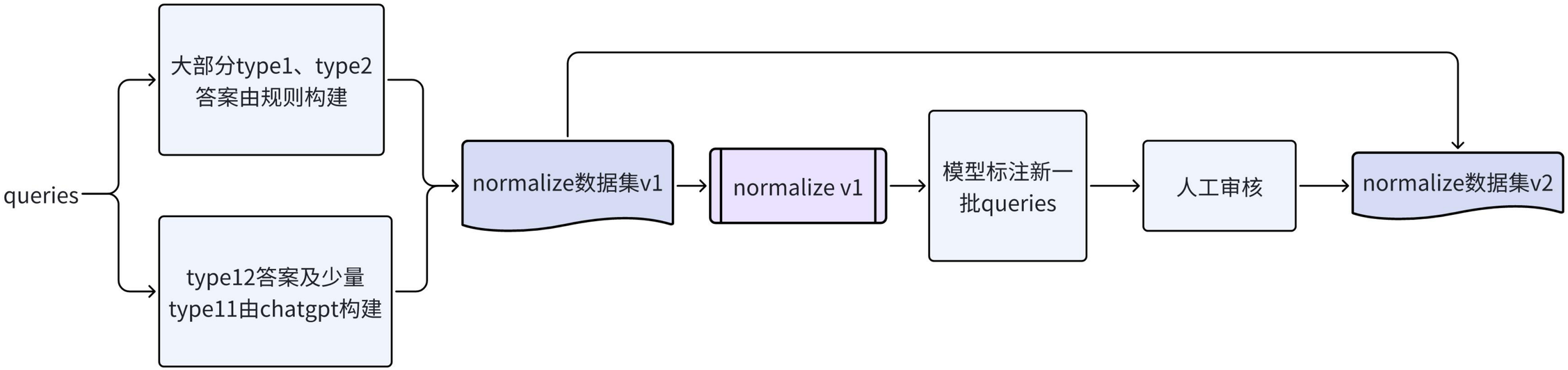
根据查询结果回答问题。
【查询结果】{'年份': ['2019年', '2020年'], '法定代表人': ['苏科舜', '刘宝珠']}
【问题】开滦股份2019-2020年这两年的法定代表人是否都相同？
【回答】

response

开滦股份2019年的法定代表人是苏科舜，2020年的法定代表人是刘宝珠，所以答案是不相同，不同。

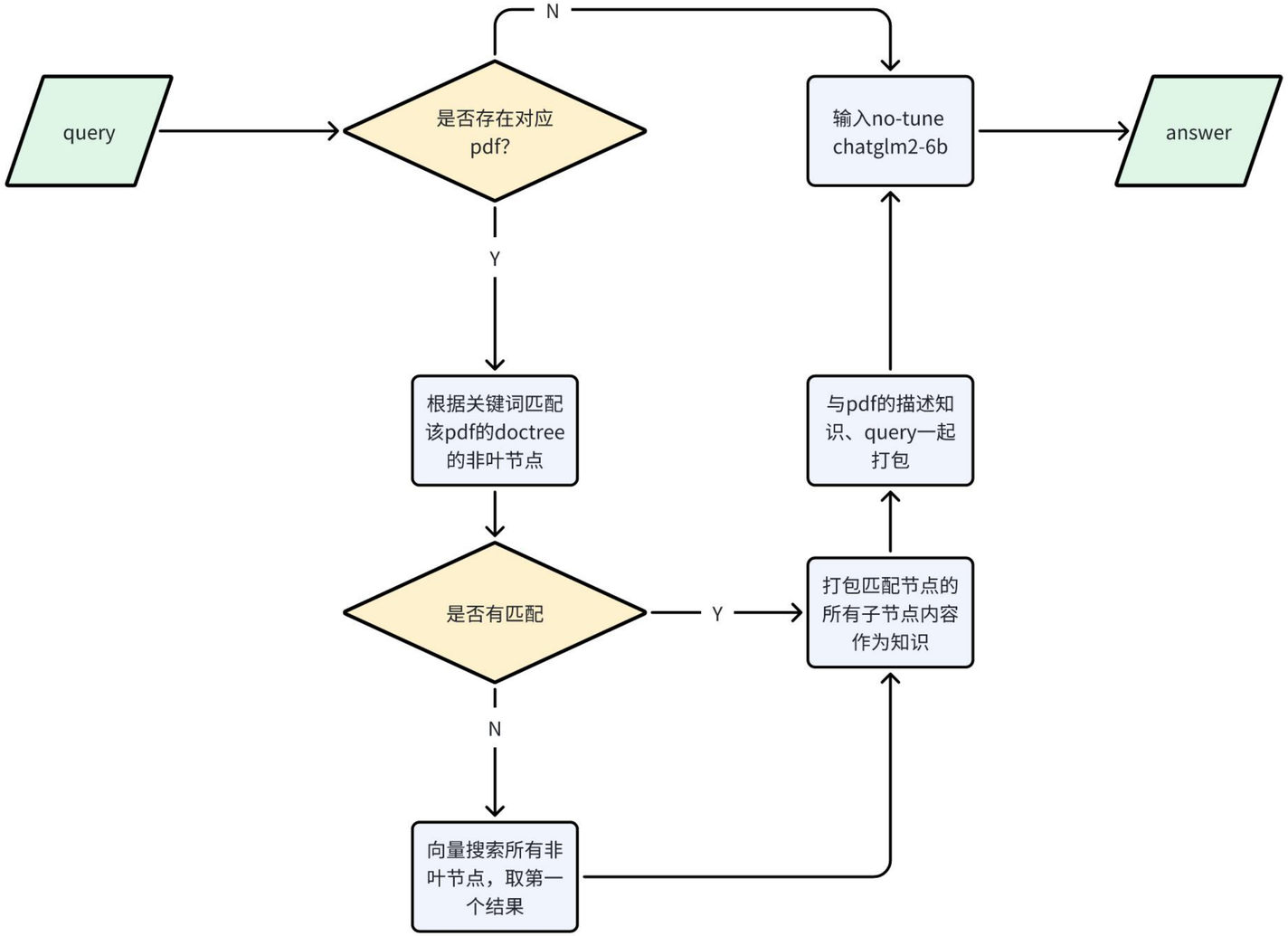
微调模型-normalize

how to?



效果：约85%无问题，部分答案丢关键词。经检查是数据集质量问题

知识搜索-type31解题



推理效率

- 测试环境
  - 1张3090
  - Huggingface transformers python调用
- 测试数据
  - B榜中200条query，150条type1、type2；50条type3
- 方法
  - 测三次取平均值。

问题类型	router	nl2sql	normalize	no-tune	total
type1、type2	0.17s/it (0.18s/it、0.16s/it、0.18s/it)	1.79s/it (1.81s/it、1.78s/it、1.78s/it)	2.59s/it (2.61s/it、2.61s/it、2.56s/it)	\	4.55s/it
type3		\		8.67s/it (8.64s/it、8.68s/it、8.70s/it)	8.84s/it

比赛心得

1. 大模型对微调数据质量要求非常高，重要性在数据量之上。
2. 6b的幻觉问题比较重，下游微调可以有效缓解。
3. 大模型学习能力极强，能够高效学习各种seq2seq任务。
4. QA系统搭建经验：精准问答走数据库nl2sql引擎，非精准问答走向量搜索。

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

汇报人：刘幼峰

2023.9.23

数据流程示例-type11

stage	prompt	response
router	厦门延江新材料股份有限公司2019年的其他非流动资产为多少元?	type11
nl2sql	你的任务是将问题转化为SQL。 1. SQL语句查询的表名为: big 2. 涉及到的列名有: 年份,其他非流动资产,其他流动资产,公司名称,股票代码,股票简称 【问题】公司名称为厦门延江新材料股份有限公司的公司2019年的其他非流动资产为多少元? 【SQL】	select 其他非流动资产 from big where 年份 in ('2019年') and 公司名称 in ('厦门延江新材料股份有限公司')
normalize	根据查询结果回答问题。 【查询结果】{'其他非流动资产': [80080633.59]} 【问题】厦门延江新材料股份有限公司2019年的其他非流动资产为多少元? 【回答】	厦门延江新材料股份有限公司2019年的其他非流动资产为80080633.59元。

数据流程示例-type11-2

stage	prompt	response
router	2021年京投发展的法定代表人与前两年是否相同？	type11
nl2sql	你的任务是将问题转化为SQL。 1. SQL语句查询的表名为: big 2. 涉及到的列名有: 合同资产,公司名称,法定代表人,股票简称,股票代码,年份 【问题】2021年股票简称为京投发展的公司的法定代表人与前两年是否相同？ 【SQL】	select 年份, 法定代表人 from big where 年份 in ('2019年', '2020年', '2021年') and 股票简称 in ('京投发展')
normalize	根据查询结果回答问题。 【查询结果】{'年份': ['2019年', '2020年', '2021年'], '法定代表人': ['魏怡', '魏怡', '魏怡']} 【问题】2021年京投发展的法定代表人与前两年是否相同？ 【回答】	京投发展2019年的法定代表人是魏怡，2020年的法定代表人是魏怡，2021年的法定代表人是魏怡，所以答案是相同。

数据流程示例-type12

stage	prompt	response
router	2019年哪4家在上海注册的上市公司，营业外收入最高？金额是？	type12
nl2sql	你的任务是将问题转化为SQL。 1. SQL语句查询的表名为: big 2. 涉及到的列名有: 年份,股票代码,营业收入,营业外收入,公司名称,注册地址,股票简称,营业总收入 【问题】2019年哪4家在上海注册的上市公司，营业外收入最高？金额是？ 【SQL】	select 公司名称, 营业外收入 from big where 年份 = '2019年' and 注册地址 like '%上海%' order by 营业外收入 desc limit 4
normalize	根据查询结果回答问题。 【查询结果】{'公司名称': ['上海汽车集团股份有限公司', '华域汽车系统股份有限公司', '上海建工集团股份有限公司', '上海电力股份有限公司'], '营业外收入': [766952147.81, 416322468.4, 318139572.37, 107571059.0]} 【问题】2019年哪4家在上海注册的上市公司，营业外收入最高？金额是？ 【回答】	2019年营业外收入最高的四家在上海注册的上市公司分别是：上海汽车集团股份有限公司，金额为766952147.81元；华域汽车系统股份有限公司，金额为416322468.4元；上海建工集团股份有限公司，金额为318139572.37元；上海电力股份有限公司，金额为107571059.0元。

数据流程示例-type21

stage	prompt	response
router	在2019年的时候，柳药集团速动比率为多少？	type2
nl2sql	你的任务是将问题转化为SQL。 1. SQL语句查询的表名为: big 2. 涉及到的列名有: 公司名称,股票代码,流动负债合计,年份,流动资产合计,存货,股票简称 3. 涉及到的公式有: 速动比率=(流动资产合计-存货)/流动负债合计 【问题】在2019年的时候，股票简称为柳药集团的公司速动比率为多少？ 【SQL】	select 流动资产合计, 流动负债合计, 存货, 1.0 * (流动资产合计-存货)/流动负债合计 as 速动比率 from big where 年份 in ('2019年') and 股票简称 in ('柳药集团')
normalize	根据查询结果回答问题。 【查询结果】{'流动资产合计': [10094286331.65], '流动负债合计': [6872726085.58], '存货': [1508922641.63], '速动比率': ['124.92%或1.25']} 【问题】在2019年的时候，柳药集团速动比率为多少？ 【回答】	柳药集团2019年流动负债合计为6872726085.58元，流动资产合计为10094286331.65元，存货为1508922641.63元，根据公式，速动比率=(流动资产合计-存货)/流动负债合计，得出结果柳药集团2019年速动比率为124.92%或1.25。

数据流程示例-type22

stage	prompt	response
router	荣晟环保2020年的流动负债增长率保留两位小数是多少？	type2
nl2sql	你的任务是将问题转化为SQL。 1. SQL语句查询的表名为: big 2. 涉及到的列名有: 股票代码,其他流动负债,股票简称,年份,公司名称,流动负债合计 【问题】股票简称为荣晟环保的公司2020年的流动负债增长率保留两位小数是多少？ 【SQL】	select 年份, 流动负债合计 from big where 年份 in ('2020年', '2019年') and 股票简称 in ('荣晟环保')
normalize	根据查询结果回答问题。 【查询结果】{'2019年流动负债合计': 183831654.18, '2020年流动负债合计': 206000840.67, '流动负债合计增长率': '12.06%或0.12'} 【问题】荣晟环保2020年的流动负债增长率保留两位小数是多少？ 【回答】	荣晟环保2019年流动负债为183831654.18元，2020年流动负债为206000840.67元，根据公式，流动负债增长率=(流动负债合计-上年流动负债合计)/上年流动负债合计，得出结果荣晟环保2020年流动负债增长率为12.06%或0.12。