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

SMP 2023
ChatGLM 金融大模型挑战赛

演讲人 于子涵

队伍名称 南哪都队

团队构成

于子涵

南京大学自然语言处理研究组，研二在读

郭俊杰

南京大学自然语言处理研究组，研二在读

毛云麟

南京大学自然语言处理研究组，研三在读

- 指导老师：　　吴震　　戴新宇

南京大学自然语言处理研究组

- 20世纪80年代开始，从事NLP领域研究工作近40年
- 7位指导老师，近80人的博士及硕士研究生科研队伍
- 国家七五科技攻关重大成果奖、教育部科技进步二等奖等
- 在人工智能及自然语言处理领域国际顶级期刊和会议上发表论文逾120篇
- 中国人工智能学会优秀博士学位论文奖、中国中文信息学会优秀博士学位论文等
- 国际国内竞赛第一名：中文分词评测（SIGHAN 05, NLPCC 12）、命名实体国际评测（SIGHAN 06）、 全国统计机器翻译评测（CWMT13）、 2018 CCF BDCI大赛情感分析赛题、WMT2022-QE评测（MQM质量标记，英德方向）

主页　　B站　　公众号

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

汇报大纲

赛题理解

架构总览

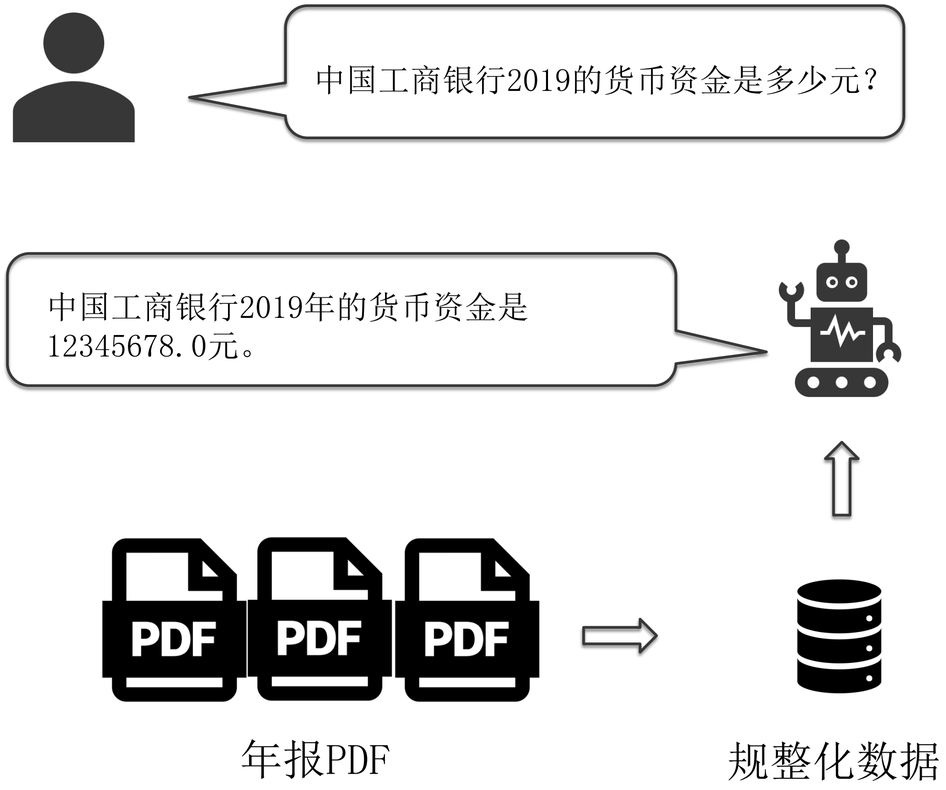
具体方法

项目总结

赛题任务

根据提供的年报数据，围绕ChatGLM2-6B打造一个问答系统，能够回答用户三类不同难度层次的问题。

- 数据基础查询
- 数据基础统计查询
- 开放类问题



评测指标：

$$\begin{cases} \max\limits_{\text{similar}}\ (sentence_1, sentence_2, sentence_3) & \text{无基础信息及关键词} \\ 0.25 + 0.25 + \max\limits_{\text{similar}}\ (sentence_1, sentence_2, sentence_3) * 0.5, & \text{基础信息正确, 关键词正确} \\ 0.25 + 0 + \max\limits_{\text{similar}}\ (sentence_1, sentence_2, sentence_3) * 0.5, & \text{基础信息正确, 关键词错误} \\ 0, & \text{基础信息错误} \end{cases}$$

基础信息：查询结果，如12345678.9元。

关键词：问题中关键信息，如2019年、货币资金等。

相似度：要求回答的句式结构和参考答案尽可能接近

年报数据分析

我们所使用的全部数据均基于官方提供的11587份PDF格式的年报数据，除此以外不包含任何外部数据

年报本身结构多样，内部表格数据的规整程度也各不相同


```
"合并资产负债表": [
    ["项目", "2019年12月31日","2018年12月31日"],
    ["货币资金", "5,259,263,917.80","3,719,593,866.02"],
    ["客户资金存款","3,826,829,042.27","2,738,957,073.31"],
    ["结算备付金","1,478,369,471.73","980,498,996.49"],
    ["客户备付金","1,197,489,555.67","793,873,093.49"],
    ["贵金属","",""],
    ["拆出资金","",""],
    ["融出资金","2,346,617,995.81","1,614,425,014.40"],
```


规整程度较高的表格示例

不同表对于“硕士及以上”属性的差异化表示：


```
["硕士及以上","25"],
["博士","14"],
["硕士","392"],
["研究生","421"],
```


包含合并单元格情况的复杂表：


```
"['股票简称', '银信科技', '股票代码', '300231']",
"['公司的中文名称', '北京银信长远科技股份有限公司', '', '']",
"['公司的中文简称', '银信科技', '', '']",
"['公司的外文名称（如有）', 'Beijing Trust&Far Technology CO.,LTD', '', '']",
"['公司的外文名称缩写（如有）', 'TRUST&FAR TECH.', '', '']",
"['公司的法定代表人', '詹立雄', '', '']",
"['注册地址', '北京市海淀区苏州街29号维亚大厦12层071室', '', '']",
"['注册地址的邮政编码', '100080', '', '']",
"['办公地址', '北京市朝阳区安定路35号北京安华发展大厦8层', '', '']",
"['办公地址的邮政编码', '100029', '', '']",
```


一些规整程度较低的表格示例

赛题挑战

1、测试集**问题种类多样**，包含计算类、查询类、分析类等，**单个框架难以处理**

训练统一的**分类器**，高效准确地完成**问题分类**、**信息抽取**和**回答模板生成**，大幅提高模型的**泛化性**，简化模型流程、节省了大量人工标注成本

2、**数据格式复杂**，难以直接从pdf中抽取出**规范化数据**

将pdf年报转化为json格式的**结构化数据**，　　针对不同数据特征发挥**规则抽取**、**模型抽取**、**文档检索**等多种数据处理策略的优势，减轻不同情况下的数据噪音影响，增强数据的利用效率

3、没有微调数据集，**人工标注数据成本太高**

设计一套**高质量**、**低成本**的数据标注框架

4、大模型**生成不可控**，难以保证输出结果的规范性

通过问题生成问题**回答模板**，保证模型生成格式的**规范性和可控性**

汇报大纲

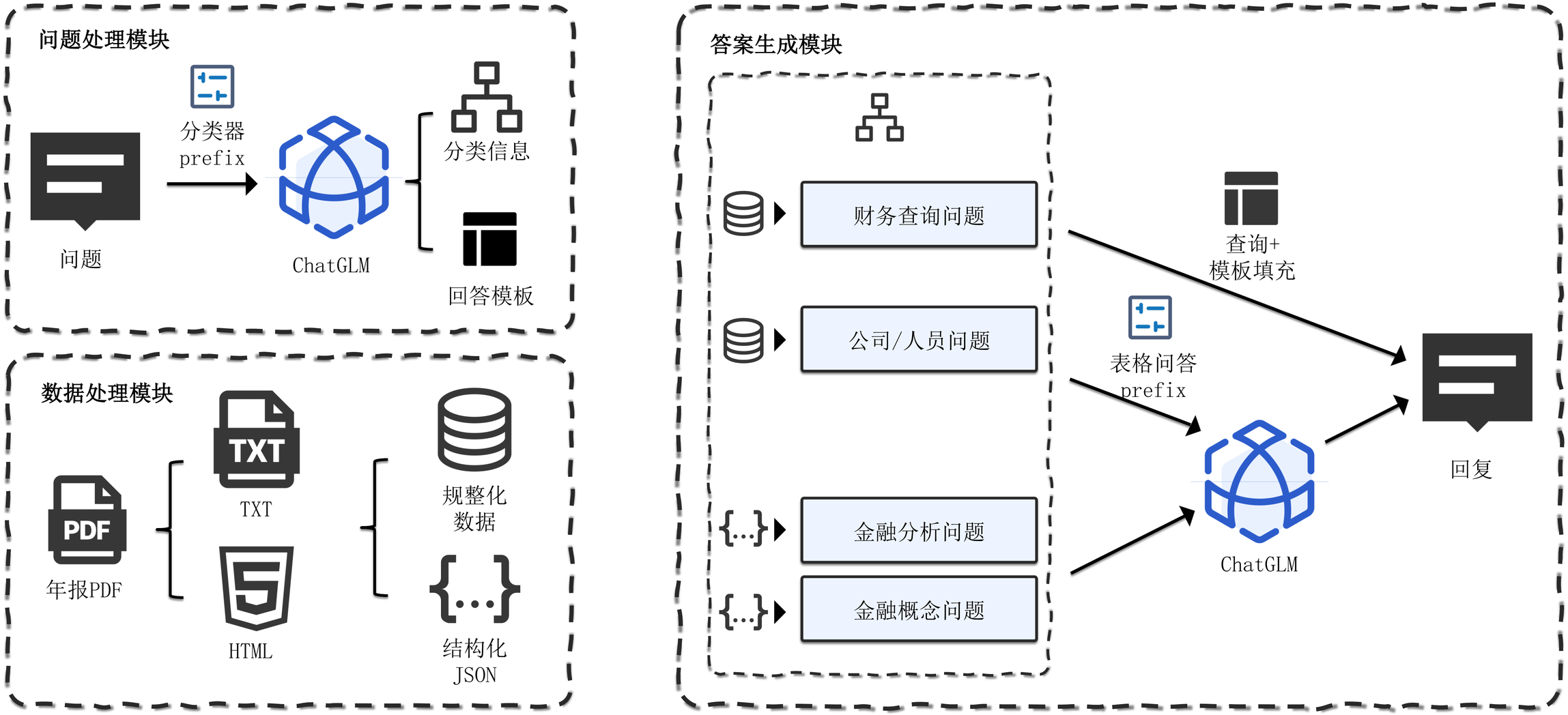
赛题理解

架构总览

具体方法

项目总结

架构总览


问题处理模块
分类器
prefix
问题
ChatGLM
分类信息
回答模板
数据处理模块
PDF
年报PDF
TXT
TXT
HTML
规整化
数据
结构化
JSON
答案生成模块
财务查询问题
公司/人员问题
金融分析问题
金融概念问题
查询+
模板填充
表格问答
prefix
ChatGLM
回复

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

汇报大纲

赛题理解

架构总览

具体方法

项目总结

数据处理模块

年报文档结构化处理

```
211, "type": "text", "inside": "第三节公司业务概要"}
212, "type": "text", "inside": "一、报告期内公司从事的主要业务"}
213, "type": "text", "inside": "（一）主要业务"}
214, "type": "text", "inside": "公司主要致力于高压及超高压电缆连接件、GIL及相关产品的研发和生产，并以上述产品为基础
215, "type": "text", "inside": "（二）主要产品及其用途"}
216, "type": "text", "inside": "公司的主要产品为电缆连接件系列及GIL系列，并为客户提供地下智能输电系统整体解决方案、
217, "type": "text", "inside": "（三）经营模式"}
218, "type": "text", "inside": "报告期内，一方面，公司重要下游客户，国家电网、南方电网及五大发电集团等电缆连接件需求
219, "type": "text", "inside": "公司的主要经营模式如下："}
220, "type": "text", "inside": "1、盈利模式"}
221, "type": "text", "inside": "公司的主要盈利模式为通过销售电缆连接件系列产品、GIL系列产品及相关配套产品，承接电力
222, "type": "text", "inside": "2、生产模式"}
223, "type": "text", "inside": "公司向客户销售电缆连接件系列产品、GIL系列产品，提供地下智能输电系统整体解决方案服务
224, "type": "页脚", "inside": "11"}
225, "type": "text", "inside": ""}
226, "type": "页眉", "inside": "江苏安靠智能输电工程科技股份有限公司2019年年度报告全文"}
227, "type": "text", "inside": "3、销售模式"}
228, "type": "text", "inside": "目前公司产品主要销往国内，采用直销模式，即由公司直接将产品销售给客户，不通过经销商销
229, "type": "text", "inside": "4、研发模式"}
230, "type": "text", "inside": "公司主要有两种研发模式，一是根据对市场趋势的判断而进行的自主研发，二是根据不同客户的
231, "type": "text", "inside": "（四）主要的业绩驱动因素"}
232, "type": "text", "inside": "报告期内，基于电缆连接件和GIL关键技术的核心竞争优势及城市电力架空线迁改与入地工程业
233, "type": "text", "inside": "报告期内，公司新业务即GIL输电业务取得有效进展。一方面，公司签署并完成了《无锡荣巷街
234, "type": "text", "inside": "万元。上述两个架空线迁改与入地工程的一次性顺利完工，意味着公司架空线迁改与入地业务
235, "type": "text", "inside": "年第一次35-220千伏设备协议库存采购招标”项目，总金额4033万元。"}
```

```
"第三节公司业务概要": [
    {
        "一、报告期内公司从事的主要业务": [
            {
                "（一）主要业务": [
                    "公司主要致力于高压及超高压电缆连接件、GIL及相关产品的研发和生产，并以上述产品为基础，为客户提供地下智能输电系统
                ],
                "（二）主要产品及其用途": [
                    "公司的主要产品为电缆连接件系列及GIL系列，并为客户提供地下智能输电系统整体解决方案、城市电力架空线迁改与入地、电力
                ],
                "（三）经营模式": [
                    "报告期内，一方面，公司重要下游客户，国家电网、南方电网及五大发电集团等电缆连接件需求客户主要执行招标采购制度，未
                    "公司的主要经营模式如下：",
                    {
                        "1、盈利模式": [
                            "公司的主要盈利模式为通过销售电缆连接件系列产品、GIL系列产品及相关配套产品，承接电力工程承包业务、城市电
                        ],
                        "2、生产模式": [
                            "公司向客户销售电缆连接件系列产品、GIL系列产品，提供地下智能输电系统整体解决方案服务。公司主要实行“以销定
                        ],
                        "3、销售模式": [
                            "目前公司产品主要销往国内，采用直销模式，即由公司直接将产品销售给客户，不通过经销商销售产品。公司下游需求
                        ],
                        "4、研发模式": [
                            "公司主要有两种研发模式，一是根据对市场趋势的判断而进行的自主研发，二是根据不同客户的具体需要进行的定制研
                        ]
                    }
                ],
                "（四）主要的业绩驱动因素": [
                    "报告期内，基于电缆连接件和GIL关键技术的核心竞争优势及城市电力架空线迁改与入地工程业绩，公司的行业地位和品牌市场
```

TXT文档

噪声多、不易于检索

JSON结构化文档

噪声少、层次分明、各层级标题易于检索

数据处理模块

表格数据抽取

将11587份PDF转换为HTML以及TXT格式，集成从两种格式的文件中提取的表格

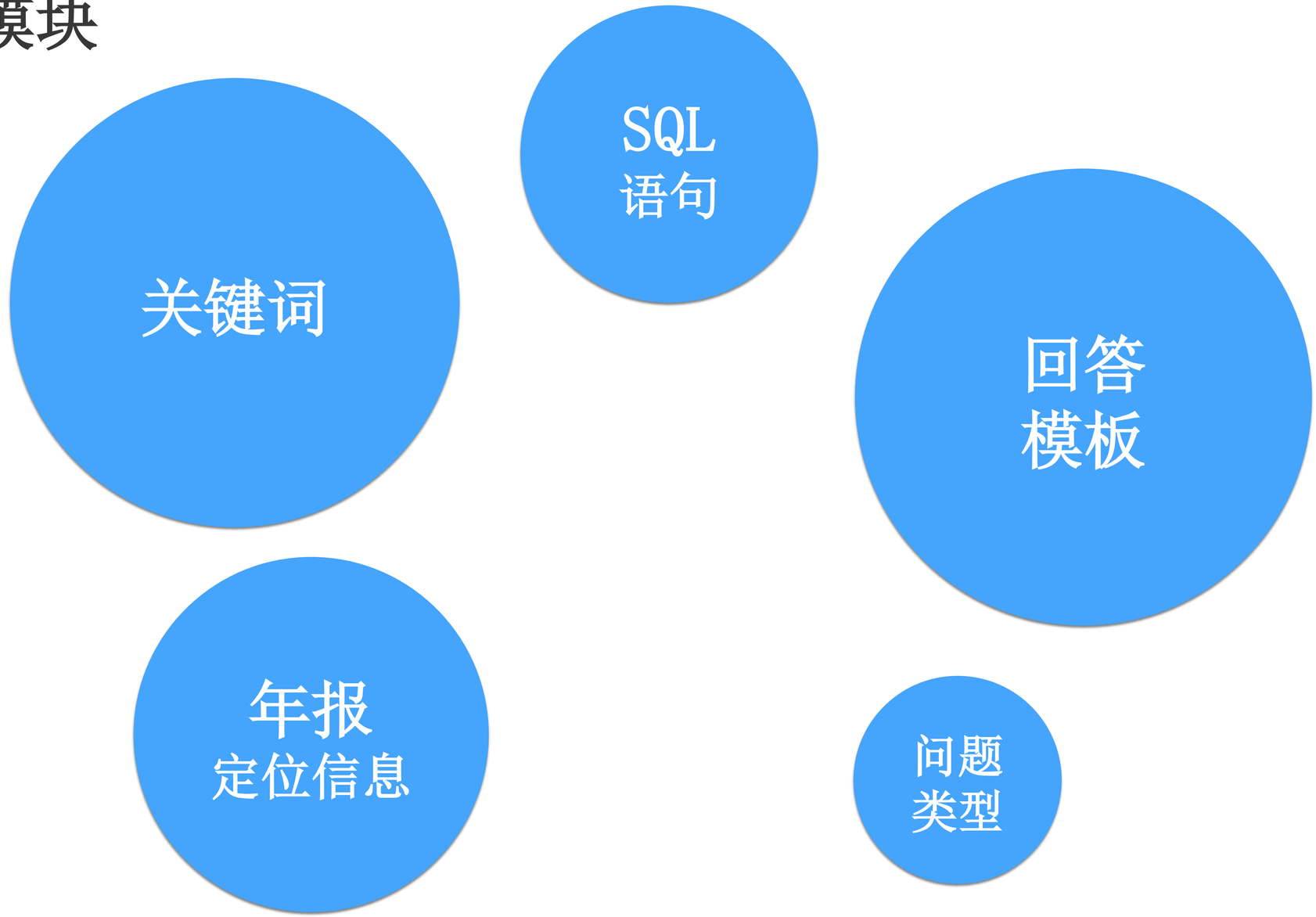
来源	资产负债表	现金流量表	利润表	公司信息相关	员工数量、专业构成及教育程度	研发投入
HTML	9334	10299	10223	96	10380	6832
TXT	11123	11000	11108	10528	11307	8557
合并	11190 （96.57%）	11180 （96.49%）	11200 （96.66%）	10529 （90.87%）	11364 （98.08%）	9380 （80.95%）

表格提取结果的统计

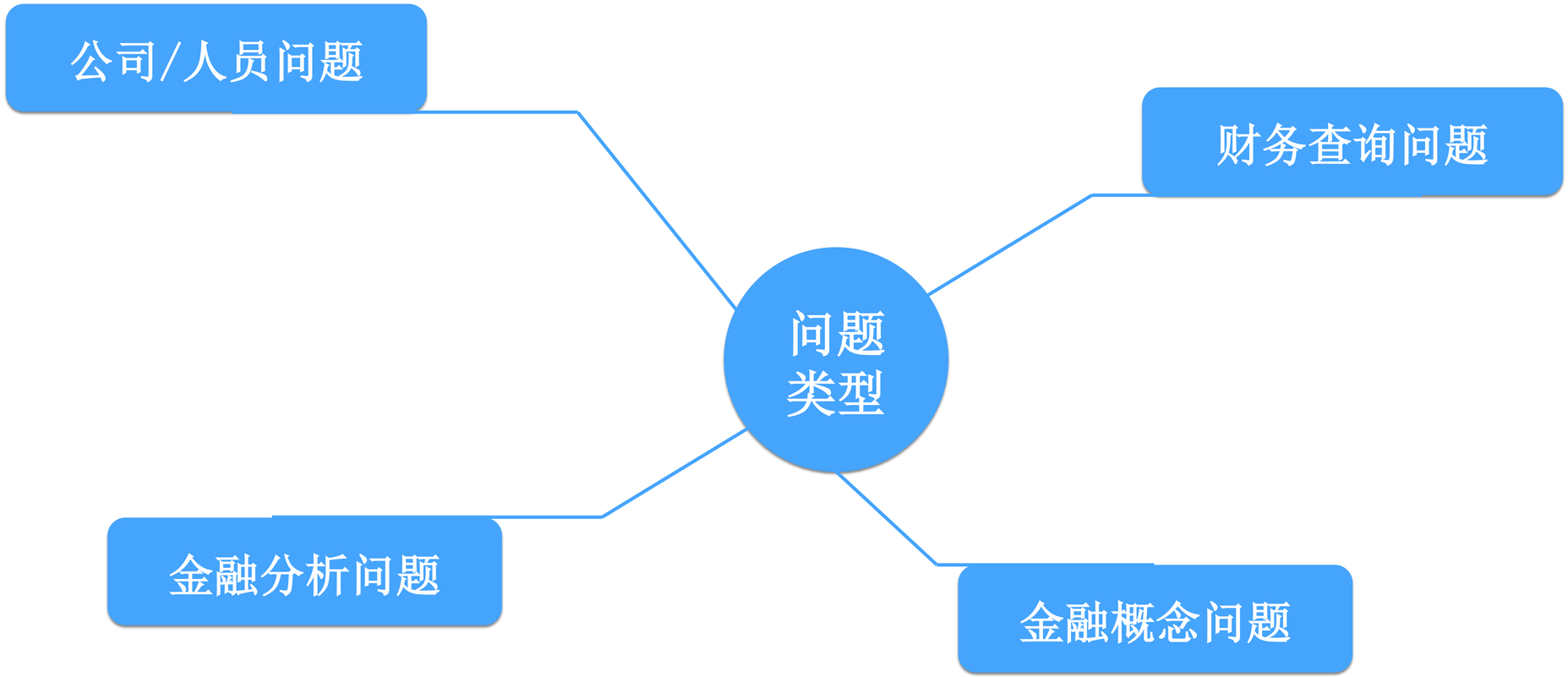
规整化表格——利用规则抽取数据

非规整化表格——利用模型抽取数据

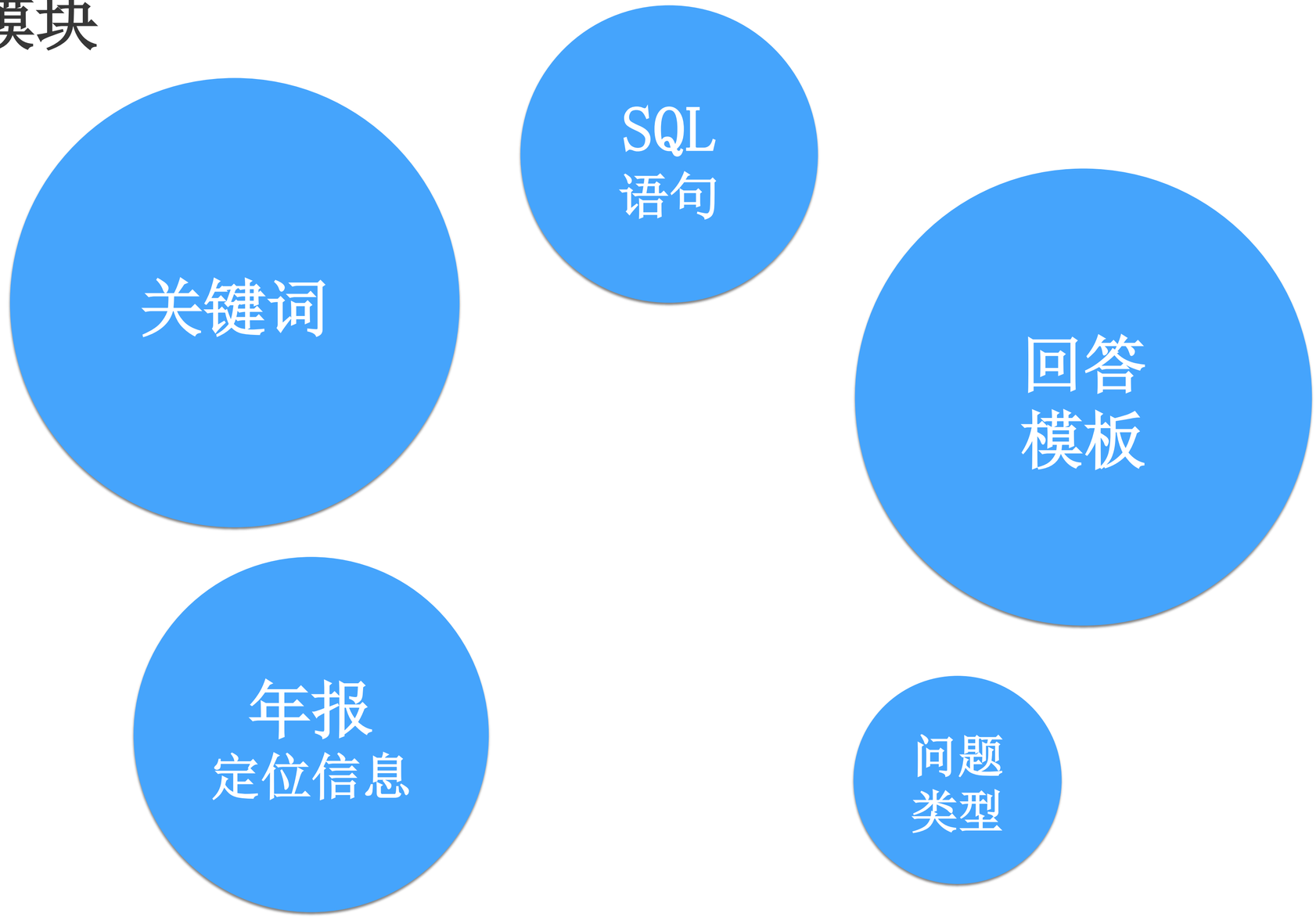
问题处理模块



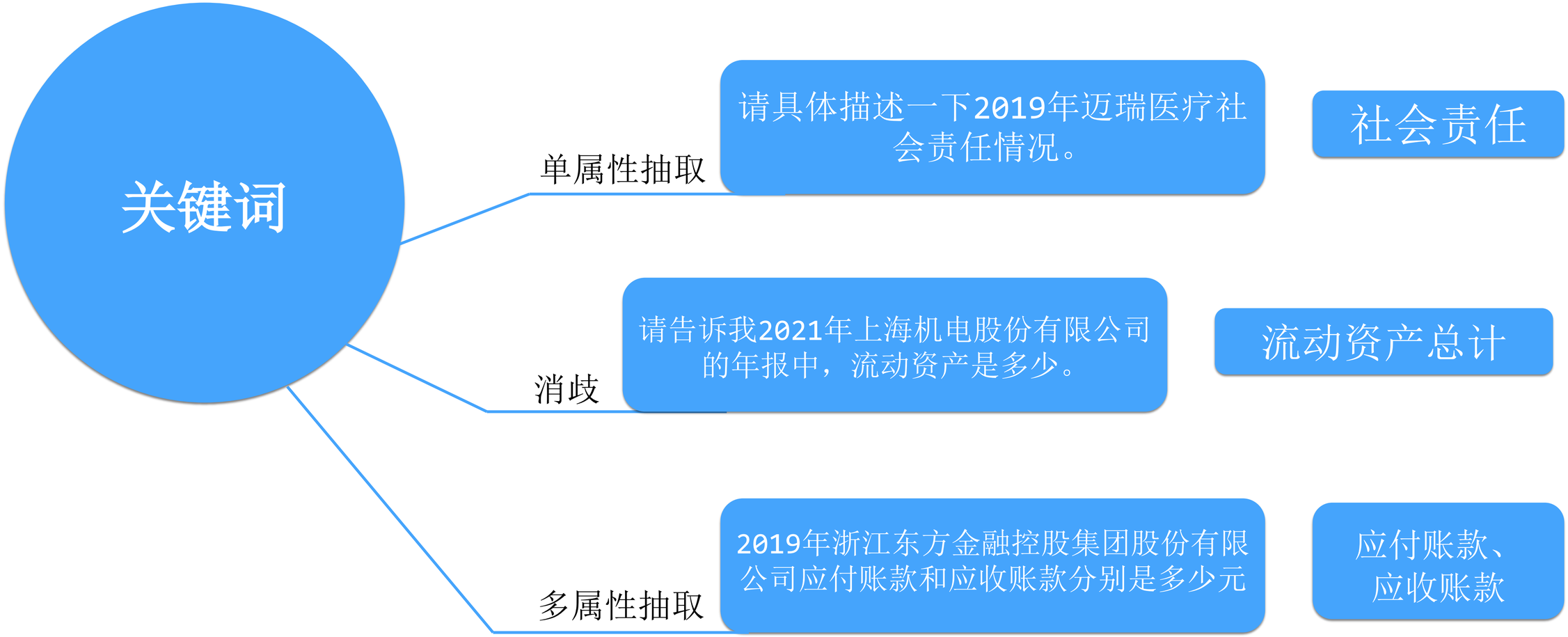
问题处理模块



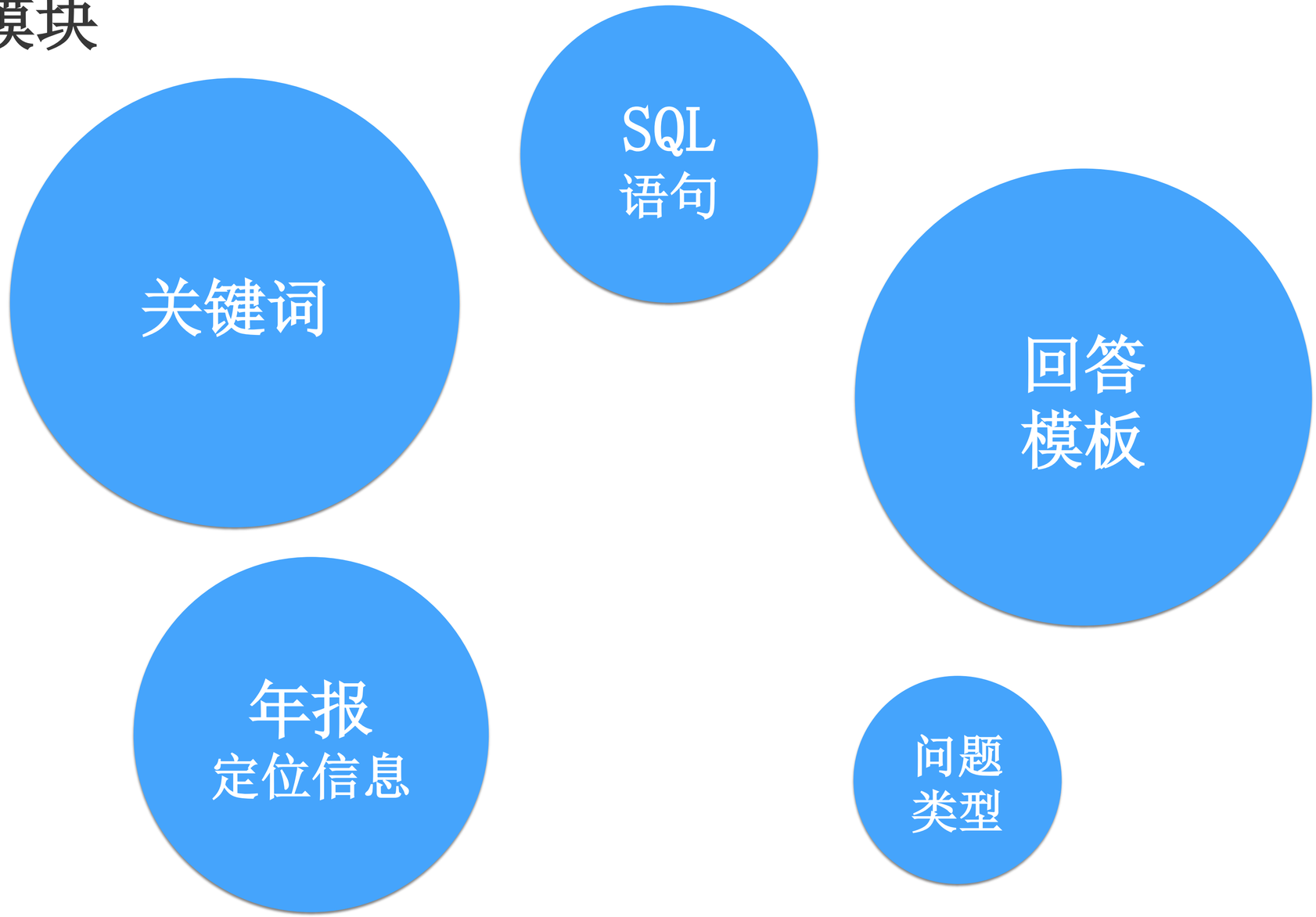
问题处理模块



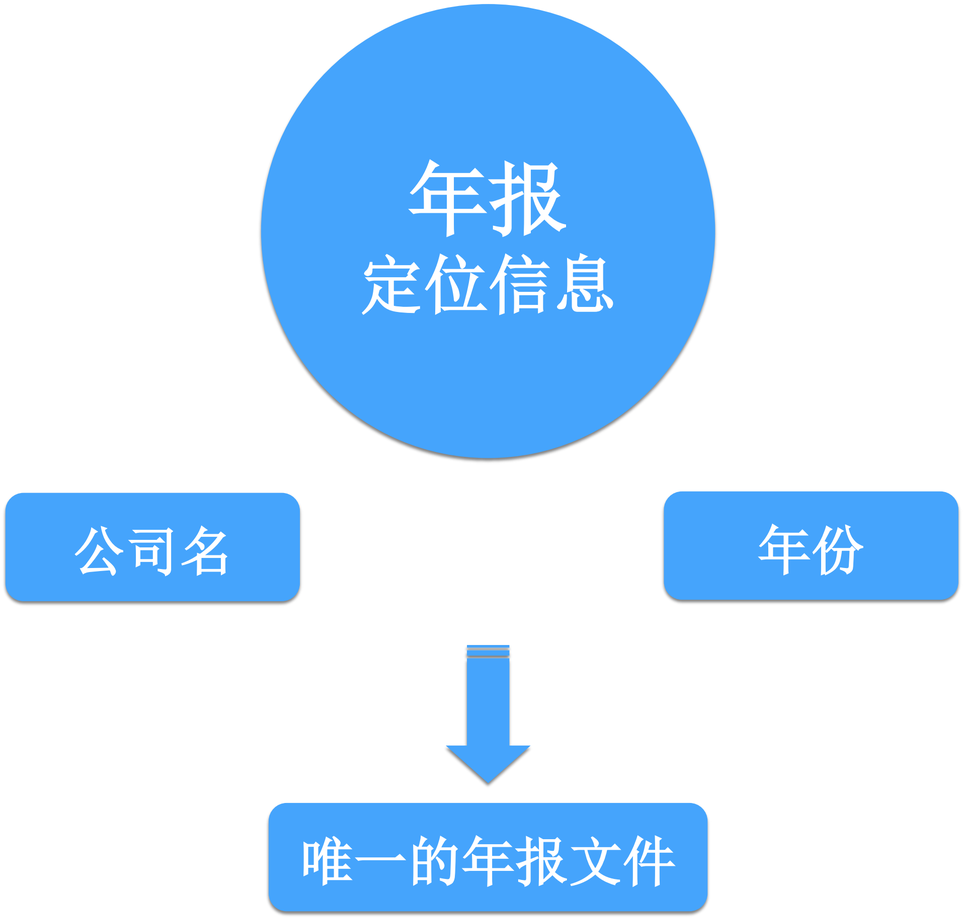
问题处理模块


关键词
单属性抽取
请具体描述一下2019年迈瑞医疗社会责任情况。
社会责任
消歧
请告诉我2021年上海机电股份有限公司的年报中，流动资产是多少。
流动资产总计
多属性抽取
2019年浙江东方金融控股集团股份有限公司应付账款和应收账款分别是多少元
应付账款、应收账款

问题处理模块



问题处理模块



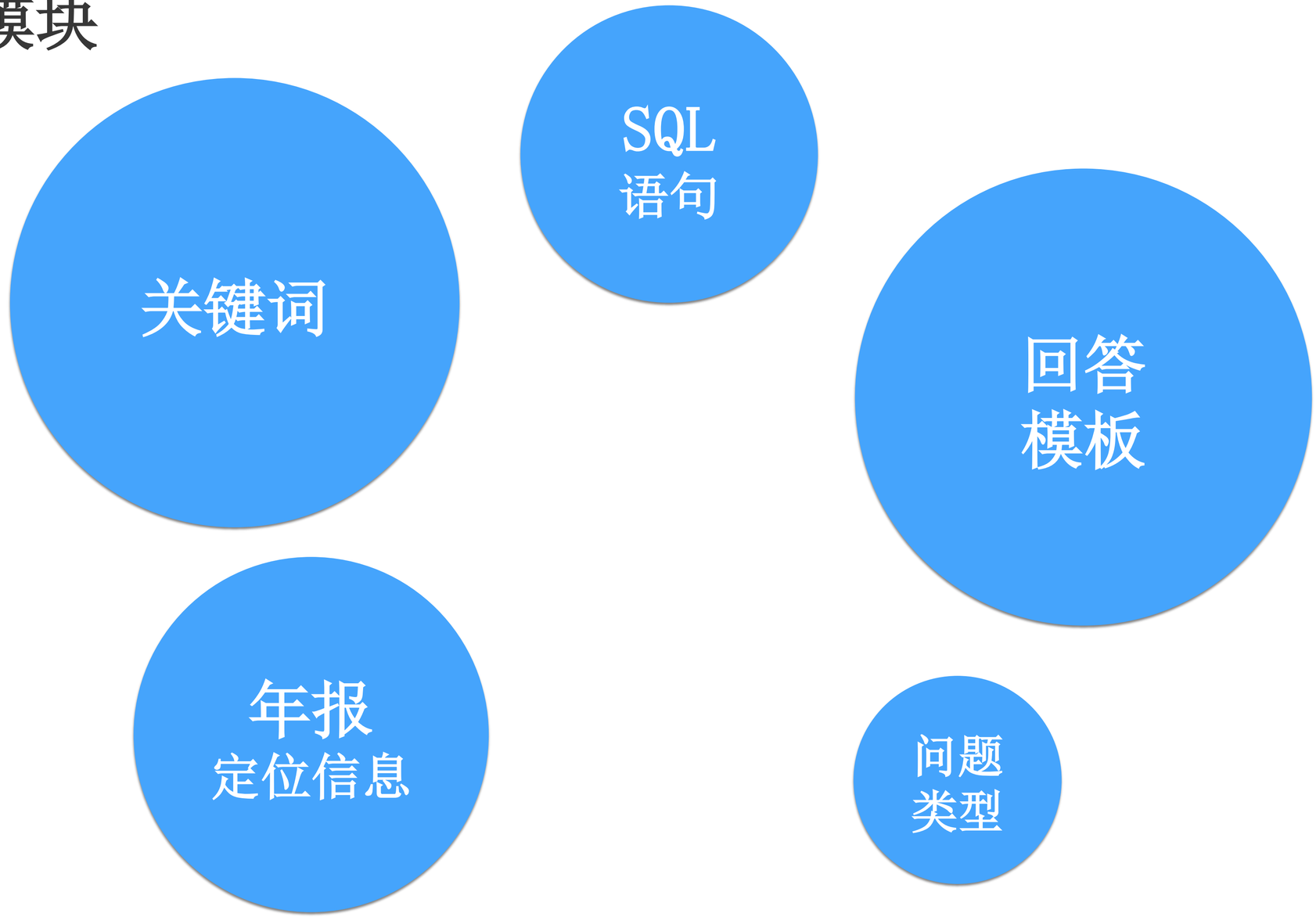
输入问题：请帮我查一下，在2020年兴业皮革科技股份有限公司的货币资金额是多少，结果保留两位小数.

公司名：兴业皮革科技股份有限公司

年份：2020

年报文件：2021-04-20__兴业皮革科技股份有限公司__002674__兴业科技__2020年__年度报告.pdf

问题处理模块



问题处理模块

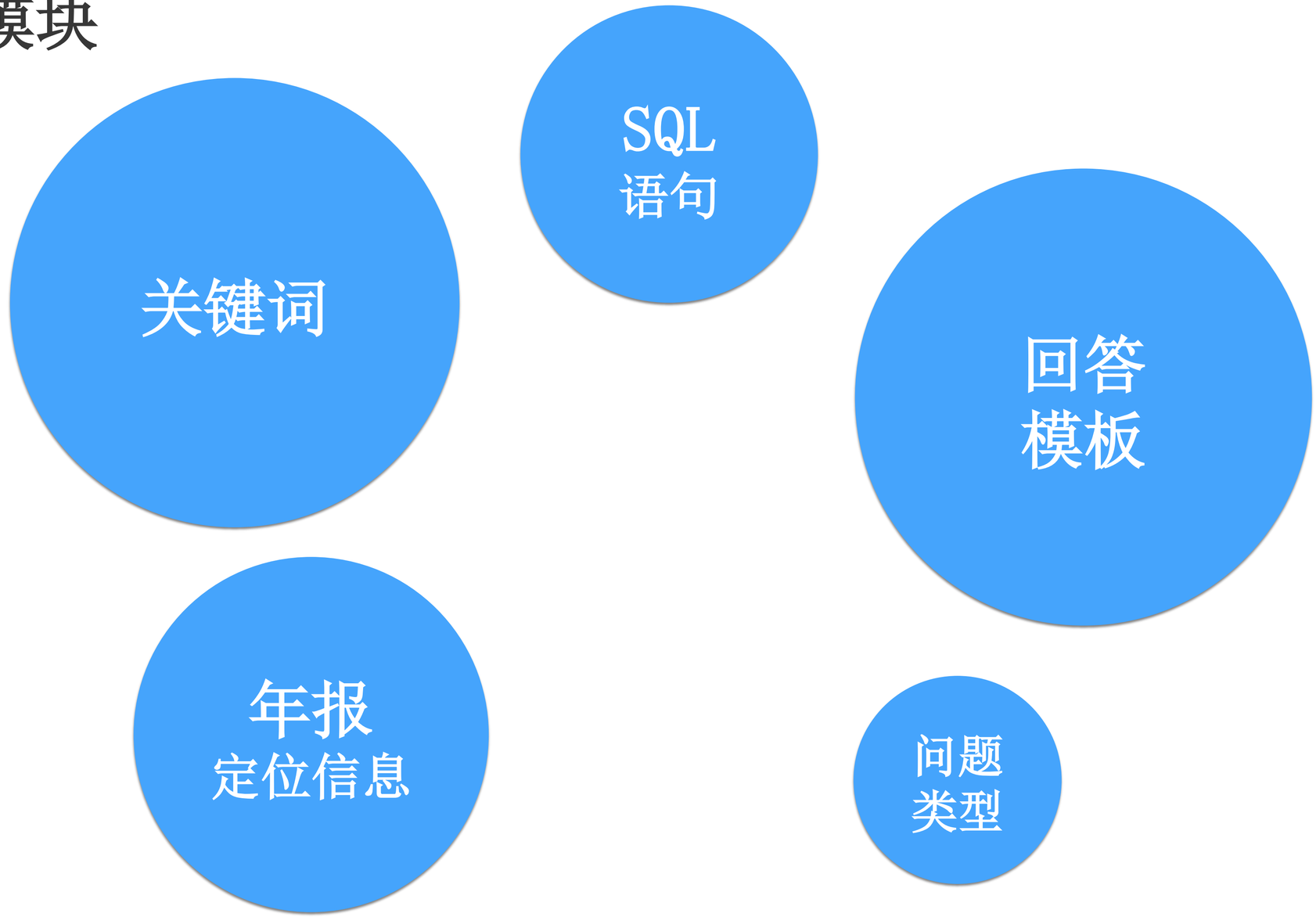
SQL
语句

基于问题输出准确的SQL查询代码

输入问题：哪家上市公司2019年货币总额最低

SQL语句：SELECT 公司的中文名称 FROM finance WHERE 年份='2019' ORDER BY 货币资金 ASC LIMIT 1

问题处理模块



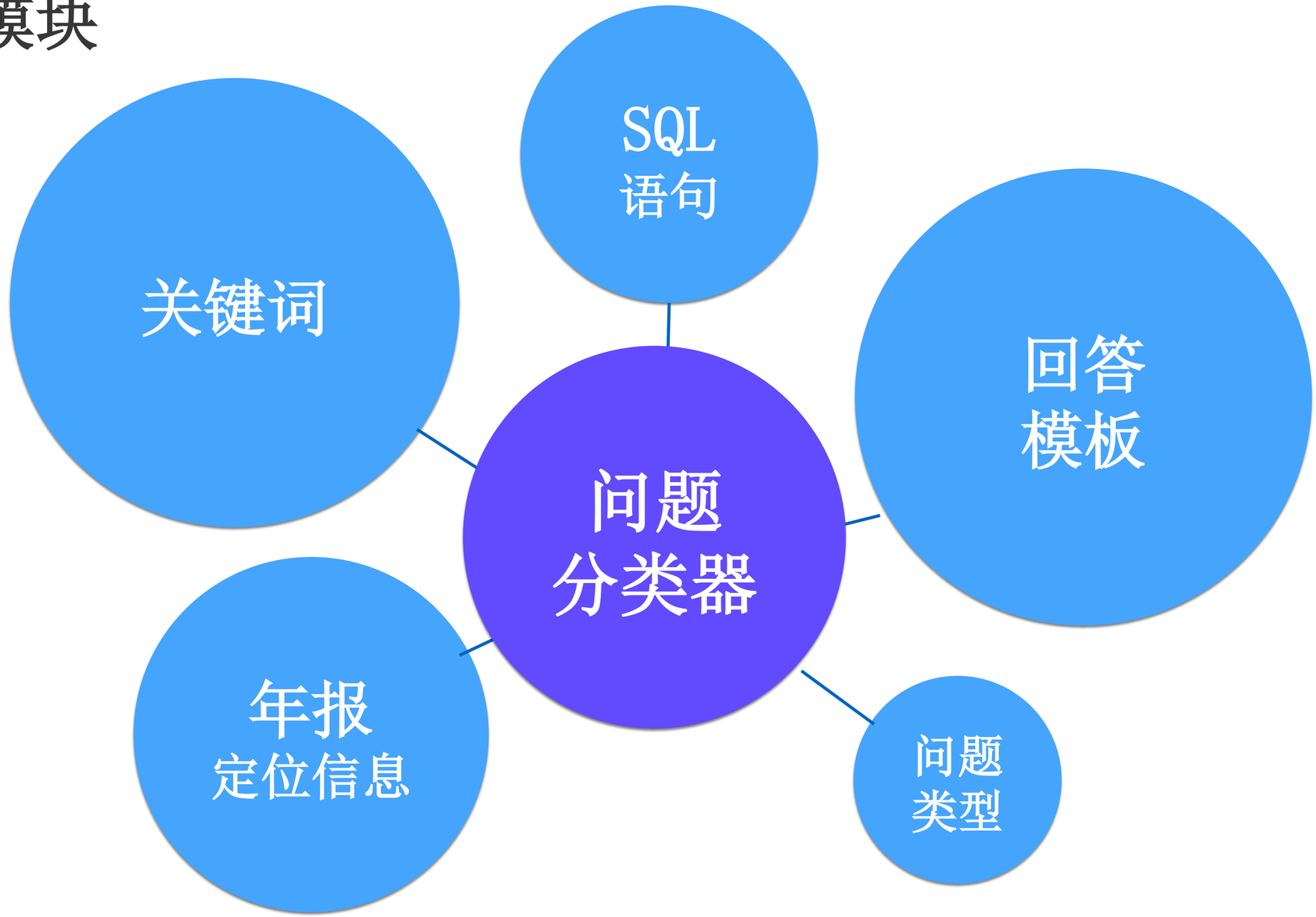
问题处理模块

输入问题：请帮我查一下，在2020年兴业皮革科技股份有限公司的货币资金额是多少，结果保留两位小数.

回答模板：2020年兴业皮革科技股份有限公司的货币资金额是{:.2f}。

	Type-1 Score	Type-2 Score
有回答模板	**88.4992**	**79.4448**
无回答模板	79.6078	76.3637

问题处理模块



问题处理模块

输入：

“请帮我查一下，在2020年兴业皮革科技股份有限公司的货币资金额是多少，结果保留两位小数.”

输出：

```
{
    "类型": "财务问题",
    "关键词": ["货币资金"],
    "公司名称": "兴业皮革科技股份有限公司",
    “年份”: [“2020"],
    "回答模板": "2020年兴业皮革科技股份有限公司的货币资金额是{:.2f}。"
}
```

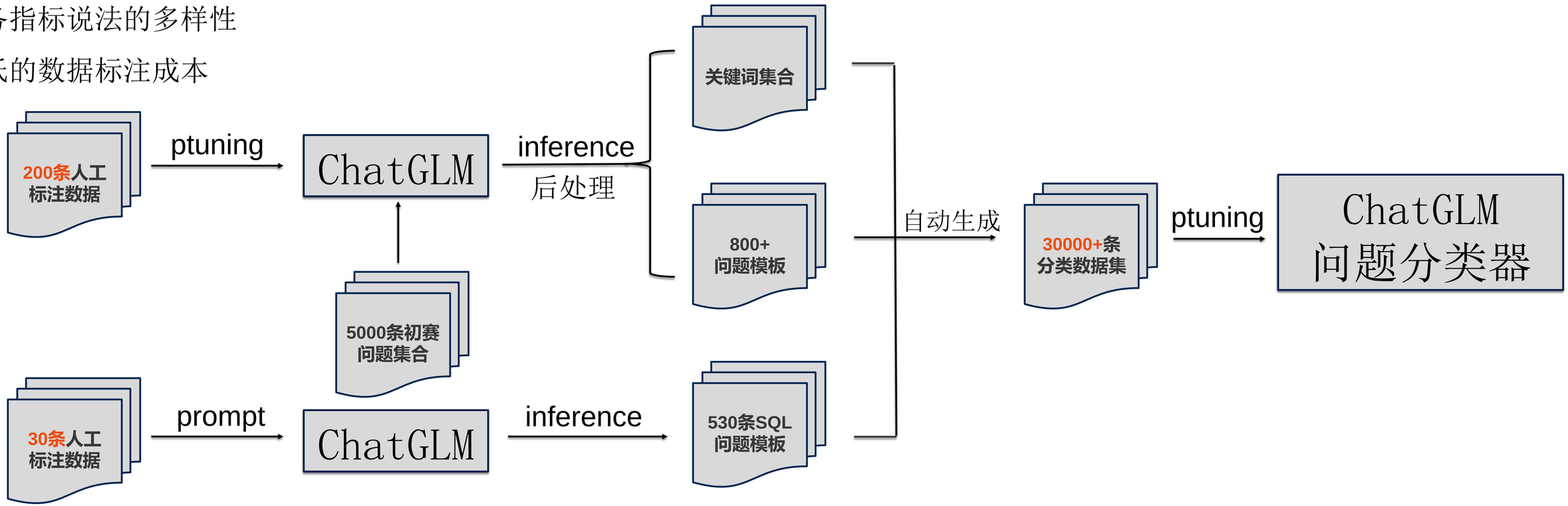
分类器输入输出示例

问题处理模块

我们构建了一个规模为30000+的数据集来训练分类器

构造数据集的几个关键问题：

- 问题问法的多样性
- 财务指标说法的多样性
- 较低的数据标注成本



训练的分类器在我们的测试集上达到了99.2%的抽取准确率

答案生成模块

财务问题/查询问题

这类问题需要查询的数据相对规整，可以直接解析为dataframe

- 简单查询——直接在dataframe中进行检索即可。
- 计算问题——按属性对应到计算所需的财务指标，外部计算得到结果。
- SQL查询——使用duckdb对dataframe执行SQL。

} 工具学习

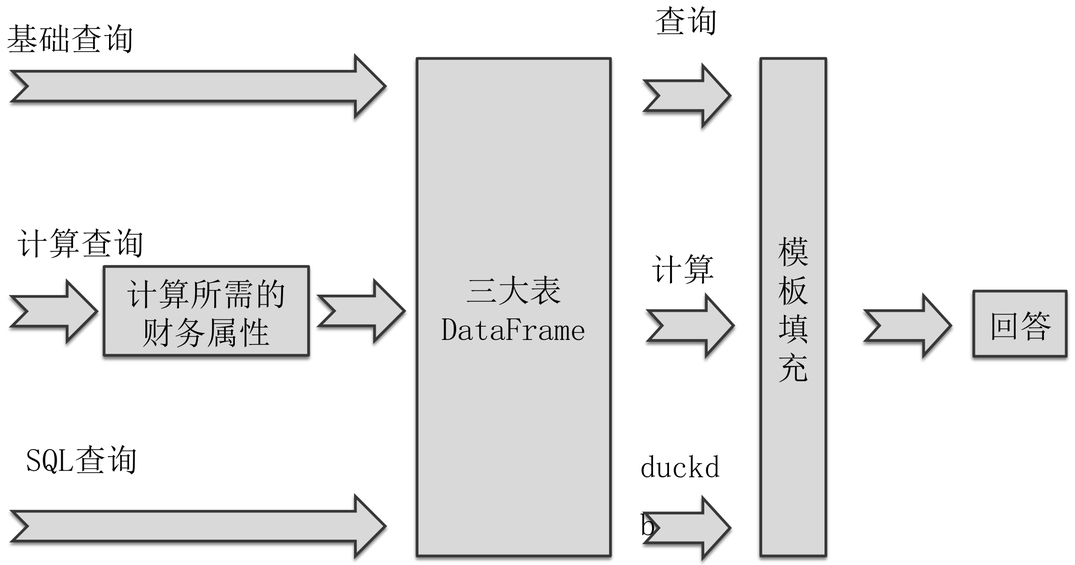
最后，填充对应的模板，输出最后的结果，整个流程如下所示。

输入：

“请帮我查一下，在2020年兴业皮革科技股份有限公司的货币资金额是多少，结果保留两位小数.”

输出：

```
{
    "类型": "财务问题",
    "关键词": ["货币资金"],
    "公司名称": "兴业皮革科技股份有限公司",
    “年份”: [“2020"],
    "回答模板": "2020年兴业皮革科技股份有限公司的货币资金额是{:.2f}。"
}
```



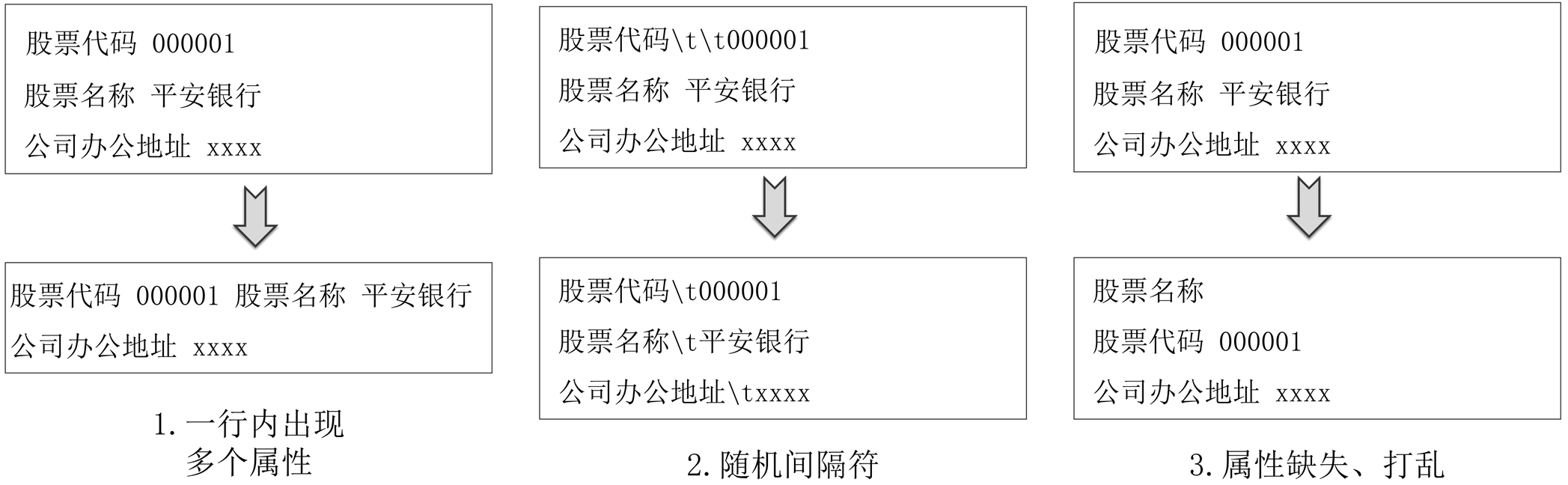
答案生成模块

公司问题/人员问题

相关表格不规整，难以直接抽取

我们解析了一部分相对规整的公司和人员表格，生成了约20000条数据，对模型进行ptuning

因为这部分的表格大多数不够规整，因而需要对输入的数据进行增强，增强方式主要有以下三类：



答案生成模块

财报分析问题

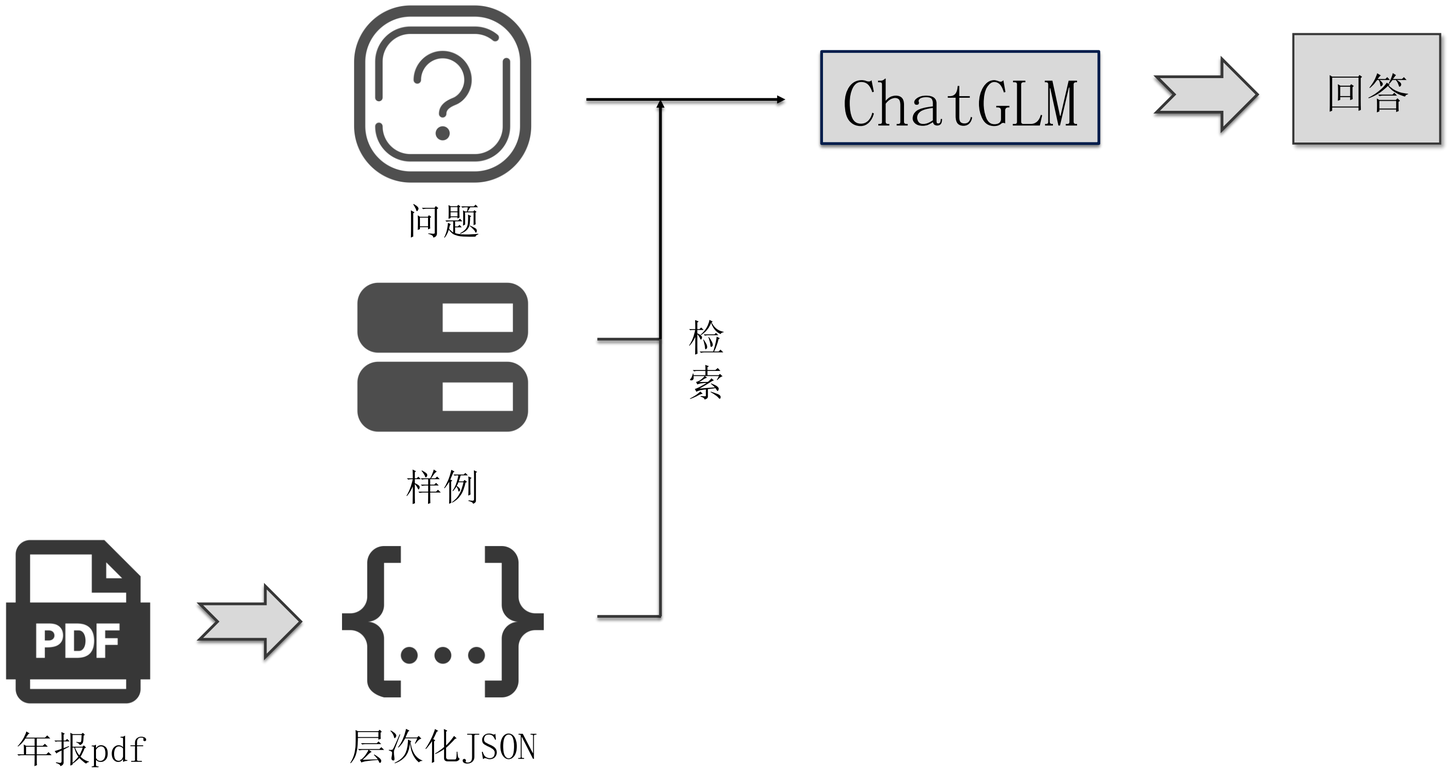
这类问题相关信息遍布整个财报，需要基于整个年报检索

利用已有的pdf解析生成结构化JSON，并将标题向量化作为文档索引用于检索，保障了文档的层次完整性和问题相关性

为了保证生成内容的效果，我们针对已知的财务分析类问题每类问题标注了1~2个样例作为In-context learning的样例加入prompt中

层次化检索和ICL的优化在这部分题目上

分别提升了6和3的最终分数。

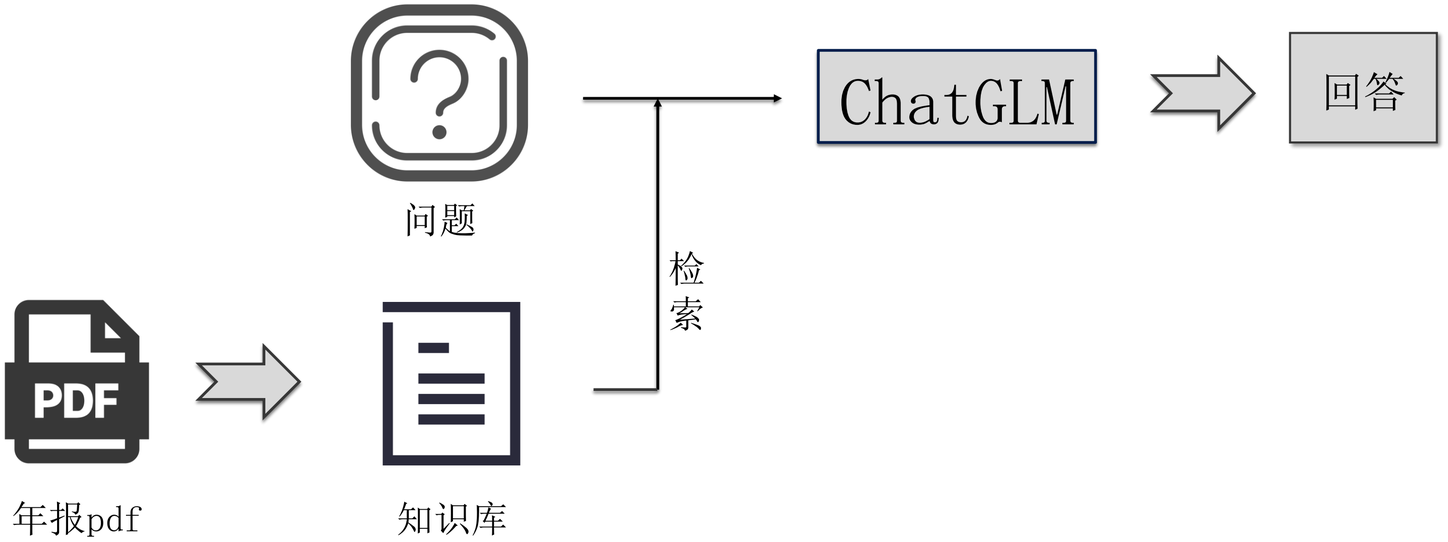


答案生成模块

金融概念问题

这类问题需要一个大的金融领域知识库

我们从年报的财务报表编制板块提取并清洗了相关的知识数据，用于问答时检索。



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

汇报大纲

赛题理解

架构总览

具体方法

项目总结

框架优势

- 统一的问题分类器：提升模型泛化性，简化后续流程
- 差异化的数据处理方式：降低噪声影响，增强数据利用效率
- 差异化的模型利用方式：利用ChatGLM打通金融问答的不同环节，最大限度发挥模型理解、分析和推理能力
- 高效、低成本的数据标注框架：数据和人工成本低，数据质量高
- 规范化的问题回答模板：保证了模型生成的可控性

经验总结

- 金融领域年报数据处理方式和噪声去除方式对后续任务效果有影响。
- 大模型在语言理解类问题上表现较好，但在计算和逻辑类问题上有限。
- 在训练数据充足的情况下，大模型可以完成关键信息的格式化抽取。
- 背景文档检索质量与模型输出效果相关，结构化的文档有助于提升检索效率。
- 对于关键词确定的问题，简单的BM25检索通常效果良好。
- 情景学习有助于规范大模型的输出。


智谱·AI

marsoft
安硕信息

北京交通大學
BEIJING JIAOTONG UNIVERSITY

ModelScope
魔 搭 社 区

阿里云

Thanks