



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

安硕硕眼探企

# 项目分享说明

赛道	算法赛道/模型开发赛道
赛题	SMP 2023 ChatGLM金融大模型挑战赛
项目进展程度	新项目

## 1、赛题来源

## 2、方案介绍

- 整体流程
- 数据处理流程
- 微调流程
- 问答流程

## 3、总结

## 4、开源计划-FinGLM

# 赛题来源

场景模拟

财务方向

角色：投资者（A），本模型（B）

A：您好，我对Z公司的财务数据有一些问题，可以向您咨询吗？

B：当然，我很乐意回答您的问题。

A：非常感谢。我首先想了解，Z公司的流动比率和速动比率如何，能否满足短期内的偿债压力？

B：根据该公司的年报里的财务数据，公司的流动比率和速动比率都在行业的平均水平之上，应该可以满足短期偿债的需要。不过，具体的偿债压力还会取决于未来的财务状况和市场环境。

A：那么，关于公司的存货和应收账款管理，Z公司的存货周转率和应收账款周转率表现如何？

B：报告显示，公司的存货周转率和应收账款周转率都在行业平均水平。这表明公司的存货管理和应收账款管理相对有效，但具体效果还需视未来的经营状况和市场环境而定。

A：听起来不错。再问一个关于公司投资的问题，Z公司在固定资产和无形资产方面的投入，是否得到了有效的利用？

B：根据报告，公司的固定资产和无形资产的投入与公司的收益以及行业的平均水平相比较，表现得相当不错。然而，这并不代表这些投资在未来一定能继续产生良好的回报，因为它们受到市场环境和经营策略的影响。

A：对于Z公司的长期投资，尤其是参股和联营公司投资，你是否能提供这些投资的价值和回报的详细信息？

B：很遗憾，我无法提供公司投资价值 and 回报的详细信息，因为这类信息通常包含了敏感的商业信息。然而，我可以确认的是，Z公司在年报中已经对这些投资进行了充分的披露，并得出了合理的价值评估。

A：明白了。那么，Z公司的资本结构如何？是否有足够的财务灵活性来应对可能的市场变化？

B：据我从年报中看到，Z公司的资本结构相对均衡，负债和权益的比例适中。这应该能为公司提供一定的财务灵活性，以应对可能的市场变化。然而，这需要结合公司的具体经营情况和市场环境来考虑。

A：Z公司的杠杆比率如何？是否过高，对公司的风险承受能力造成威胁？

B：在我审计的期间，Z公司的杠杆比率并未超过行业的平均水平，我不认为这会对公司的风险承受能力构成威胁。当然，这需要视未来的市场环境和公司的经营情况而定。

A：Z公司的现金及现金等价物比例是否充足，来应对可能的现金流压力？

B：根据年报，Z公司的现金及现金等价物比例相对健康，应该可以应对大部分短期内的现金流压力。但这也需要考虑到公司未来的投资计划和可能的财务状况变化。

A：Z公司的长期负债和短期负债的比例如何？是否反映了公司健康的财务管理？

B：Z公司的长期负债和短期负债的比例相对均衡，这通常表明公司的财务管理在一定程度上是健康的。当然，对于公司的健康财务管理，我们还需要考虑更多其他的财务指标。

A：Z公司的应付账款周转率如何？是否有延迟支付供应商的情况？

B：据我从年报中了解，Z公司的应付账款周转率基本稳定，没有明显的延迟支付供应商的情况。这个情况也反映了公司的健康的供应链管理。

A：Z公司的权益收益率（净利润/权益）是否满足投资者的期望？

B：Z公司的权益收益率相比行业平均水平有所超越，但是是否满足投资者的期望则取决于各位投资者自身的期望和投资目标。

# 赛题来源

## 重要提示或概述

## 公司简介和主要财务指标

- 公司简介
  - 公司信息
  - 联系人和联系方式
  - 信息披露及备置地点
  - 注册变更情况
  - 其他有关资料
- 主要会计数据和财务指标
- 境内外会计准则下会计数据差异
- 分季度主要财务指标
- 非经常性损益项目及金额

哪些公司的董事长发生了变更？  
哪些公司的董秘发生了变更？  
哪些公司的办公地址发生了跨省分变更？

请简要分析公司的主要会计数据和财务指标及产生变化的原因。（结合主要财务指标及下面的备注）

## 公司业务概要

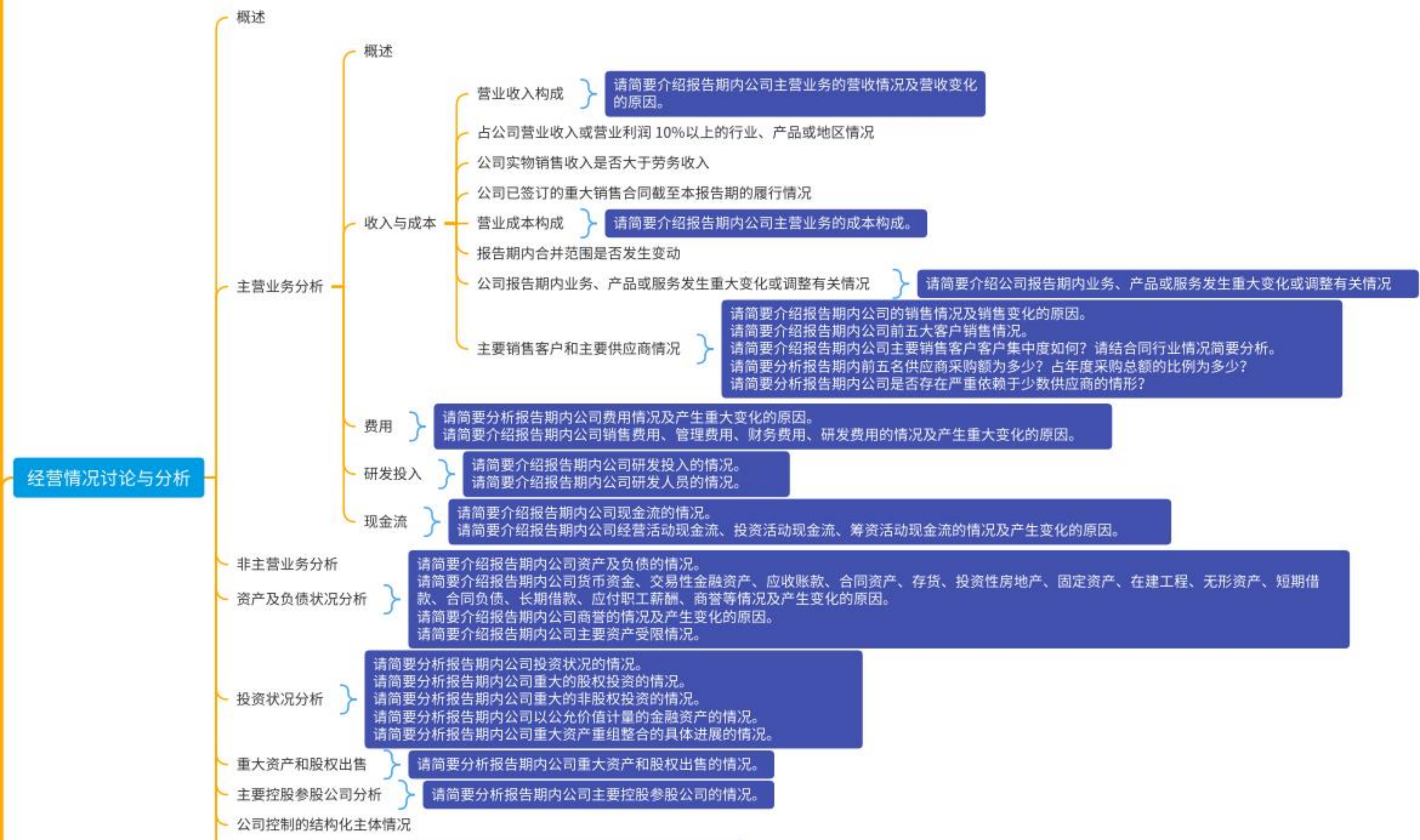
- 报告期内公司从事的主要业务
- 主要资产重大变化情况
- 核心竞争力分析

请简要介绍公司的主要业务及产品。  
请简要介绍报告期内公司的经营模式。  
请简要介绍报告期内公司所处行业的情况。  
请简要介绍公司的行业地位。

请简要介绍报告期内公司的核心竞争力。  
\*请简要介绍报告期内公司的竞争策略是低成本还是差异化，如何战胜竞争对手并建立壁垒保持长期的竞争优势。  
\*请简要介绍报告期内公司的经营范围是多元化还是专业化。  
\*请简要介绍报告期内公司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？



# 赛题来源





# 赛题来源

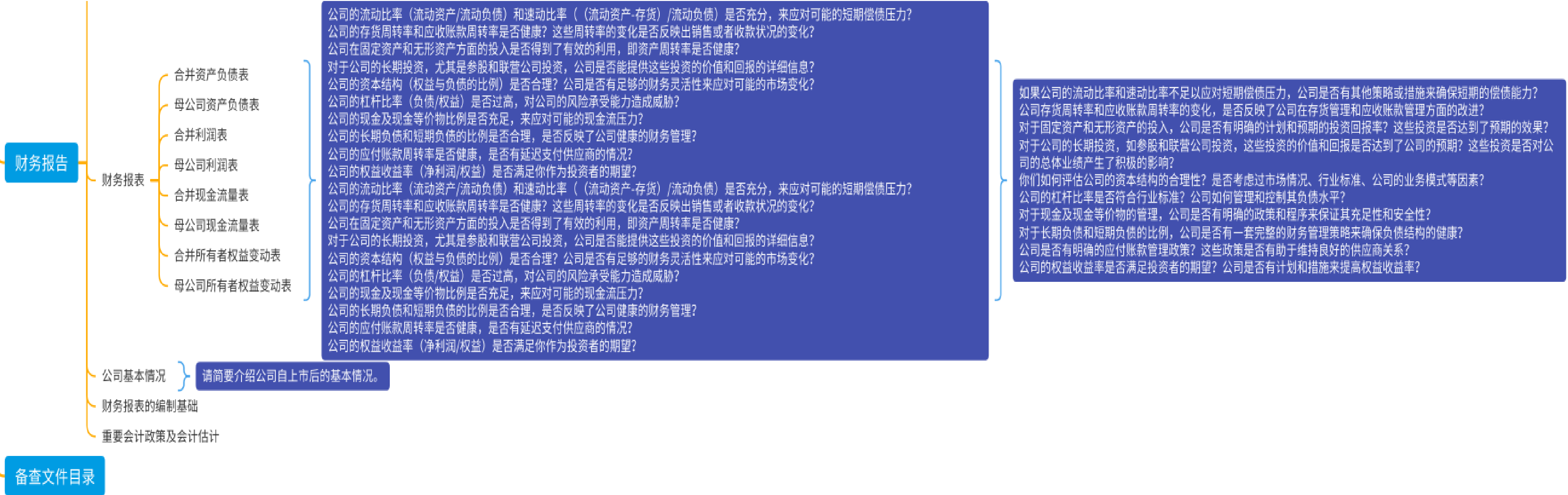


# 赛题来源





# 赛题来源



# 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

$$\begin{cases} \max_{\text{similar}} (\text{sentence1}, \text{sentence2}, \text{sentence3}), \\ 0.25 + 0.25 + \max_{\text{similar}} (\text{sentence1}, \text{sentence2}, \text{sentence3}) * 0.5, \\ 0.25 + 0 + \max_{\text{similar}} (\text{sentence1}, \text{sentence2}, \text{sentence3}) * 0.5, \\ 0, \end{cases}$$

无基础信息及关键词

基础信息正确，关键词正确

基础信息正确，关键词错误

基础信息错误

# 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？

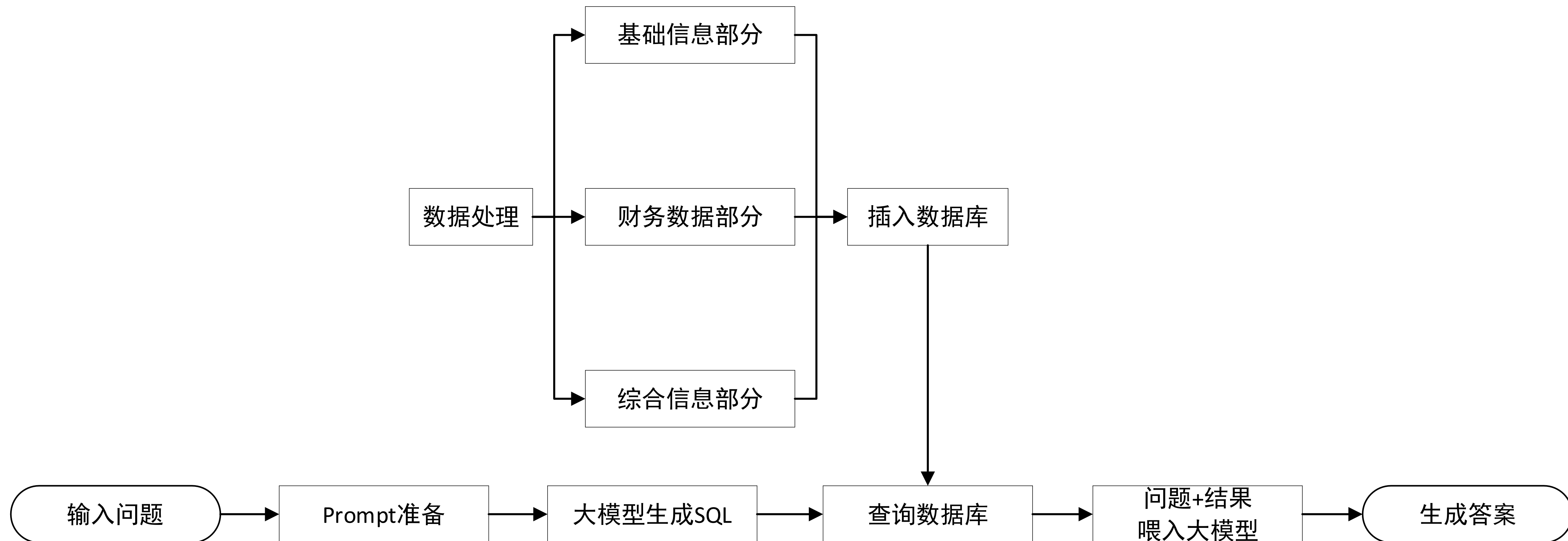
- **关键词：** 财务费用、2019
- **prompt：** "财务费用"： "12345678.9元"
- **答案：** 12345678.9元
- **示例答案：** ["2019年中国工商银行财务费用是12345678.9元。", "2019年工商银行财务费用是12345678.9元。", "中国工商银行2019年的财务费用是12345678.9元。" ]

➤ **评测计算示例：**

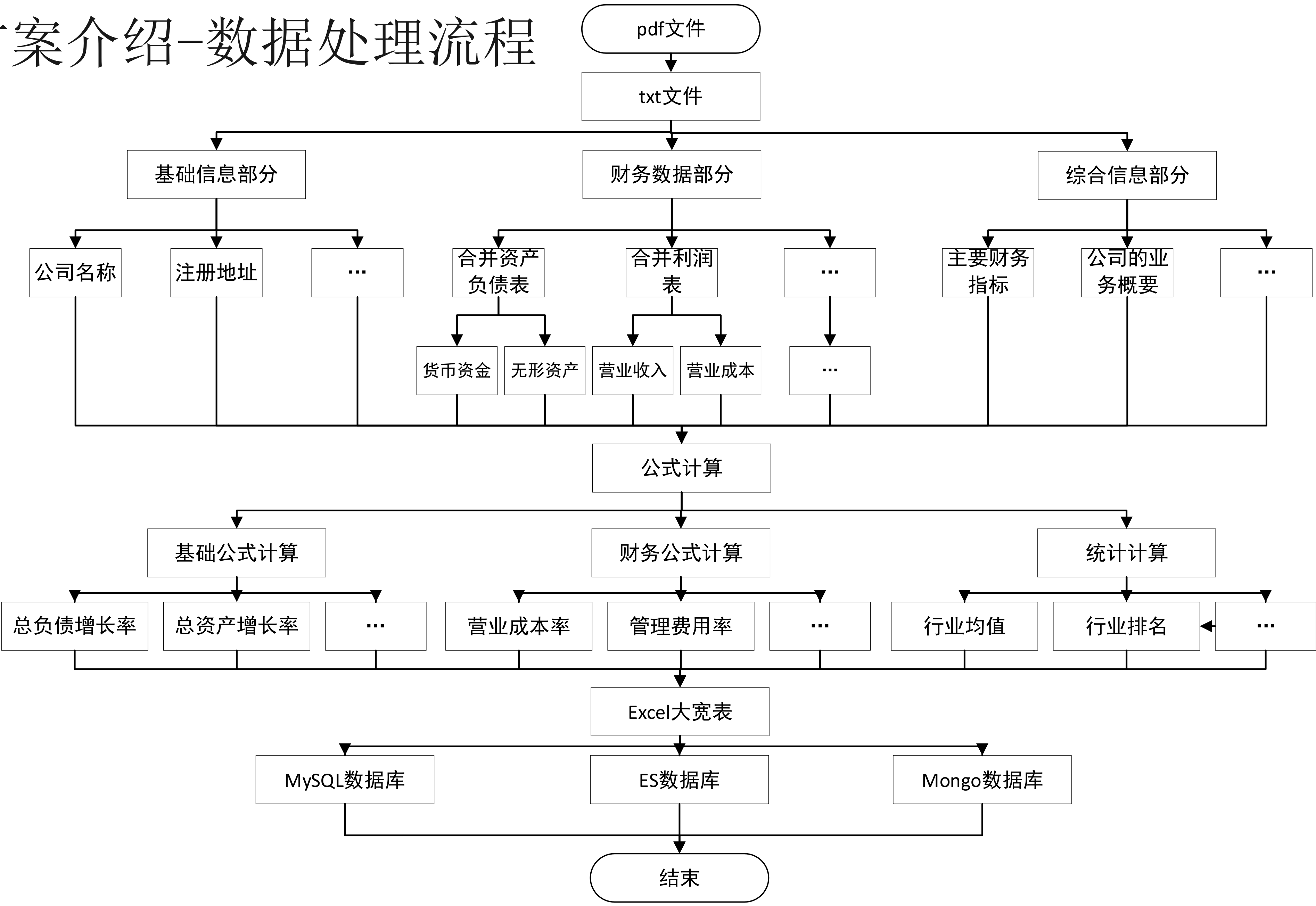
- **答案一：** 工商银行2019年财务费用是12345678.9元。
- **most similar sentences：**
  - ✓ 2019年工商银行财务费用是12345678.9元。 (Score: 0.9915)
  - ✓ 中国工商银行2019年的财务费用是12345678.9元。 (Score: 0.9820)
  - ✓ 2019年中国工商银行财务费用是12345678.9元。 (Score: 0.9720)
- **评分：**  $0.25+0.25+0.9915*0.5=0.9958$ 分。
- **评分解释：** prom\_answer正确、包含所有key\_word、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



# 方案介绍-整体流程



# 方案介绍-数据处理流程



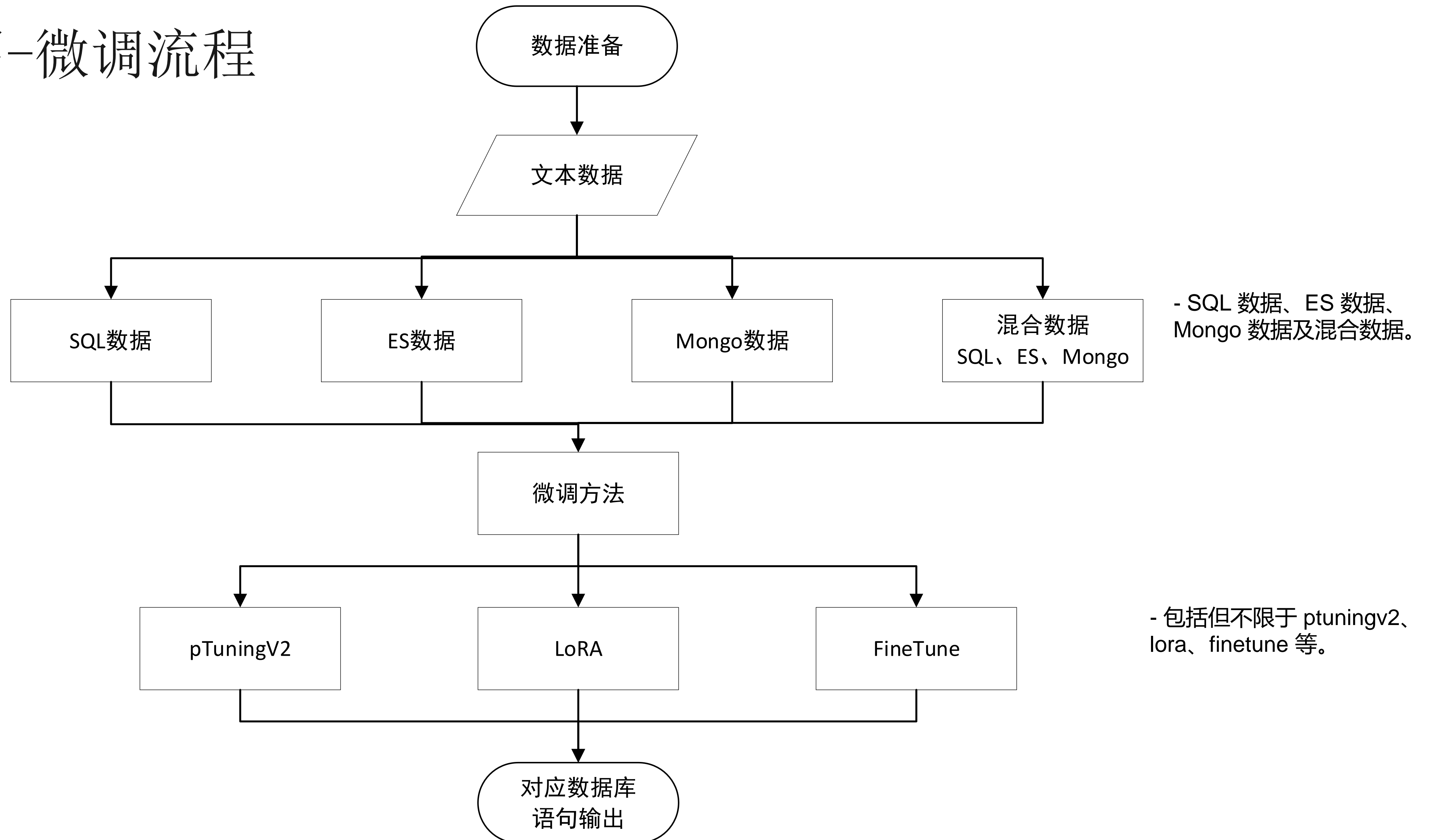
- 转换 PDF 文件为 TXT 格式。
- 保留表格格式并合并相关单元格。

- 基础信息：如公司名称、简称、英文全称等。
- 财务数据：如合并资产负债表、合并利润表等。
- 综合信息：如主要财务指标、公司业务概要等。

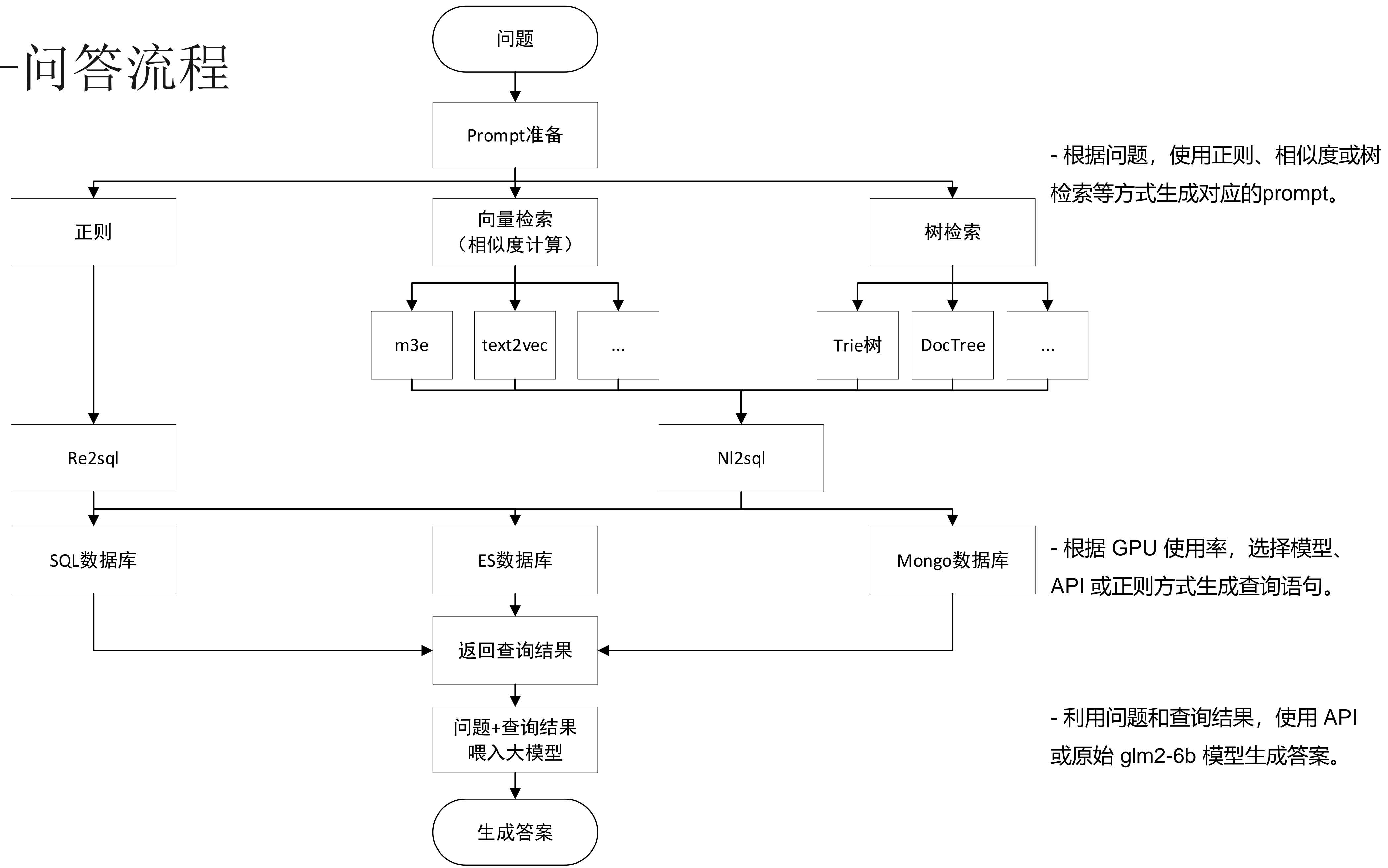
- 计算各种基础公式：如营业成本率、投资收益率等。
- 计算增长率。
- 计算行业均值及排名。

- 存入 SQL、Mongo 和 ES 数据库中。
- 涉及到建表及存表的过程。

# 方案介绍-微调流程



# 方案介绍-问答流程





## 答案示例

```
{"id": 0, "question": "2021年其他流动资产第12高的是哪家上市公司?", "answer":  
"2021年其他流动资产第12高的公司是苏美达股份有限公司。"}  
{"id": 1, "question": "注册地址在重庆的上市公司中, 2021年营业收入大于5亿的有多少家?", "answer":  
"2021年注册在重庆, 营业收入大于5亿的公司一共有4家。"}  
{"id": 2, "question": "广东华特气体股份有限公司2021年的职工总人数为?", "answer":  
"2021年广东华特气体股份有限公司职工总人数是1044人。"}  
{"id": 3, "question": "在保留两位小数的情况下, 请计算出金钼股份2019年的流动负债比率", "answer":  
"2019金钼股份流动负债比率是61.10%。其中流动负债是1068418275.97元; 总负债是1748627619.69元;"}  
{"id": 4, "question": "2019年负债总金额最高的上市公司为?", "answer":  
"2019年负债合计最高的是上海汽车集团股份有限公司。"}  
{"id": 5, "question": "2019年总资产最高的前五家上市公司是哪些家?", "answer": "2019年资产总计最高前五家  
是上海汽车集团股份有限公司、中远海运控股股份有限公司、国投电力控股股份有限公司、华域汽车系统股份有限公司、广州  
汽车集团股份有限公司。"}  
{"id": 6, "question": "2020年营业收入最高的3家并且曾经在宁波注册的上市公司是? 金额是?", "answer": "注册  
在宁波, 2020年营业收入最高的3家是宁波均胜电子股份有限公司营业收入47889837616.15元; 宁波建工股份有限公司营业  
收入19796854240.57元; 宁波继峰汽车零部件股份有限公司营业收入15732749552.37元。"}  
{"id": 7, "question": "注册地址在苏州的上市公司中, 2020年利润总额大于5亿的有多少家?", "answer":  
"2020年注册在苏州, 利润总额大于5亿的公司一共有2家。"}  
{"id": 8, "question": "浙江运达风电股份有限公司在2019年的时候应收款项融资是多少元?", "answer":  
"2019年浙江运达风电股份有限公司应收款项融资是51086824.07元。"}  
{"id": 9, "question": "神驰机电股份有限公司2020年的注册地址为?", "answer":  
"2020年神驰机电股份有限公司注册地址是重庆市北碚区童家溪镇同兴北路200号。"}  
{"id": 10, "question": "2019年山东惠发食品股份有限公司营业外支出和营业外收入分别是多少元?", "answer":  
"2019年山东惠发食品股份有限公司营业外收入是1018122.97元; 营业外支出是2513885.46元。"}  
{"id": 11, "question": "福建广生堂药业股份有限公司2020年年报中提及的财务费用增长率具体是什么?",  
"answer": "2020福建广生堂药业股份有限公司财务费用增长率是34.33%。其中, 财务费用是7766850.48元; 上年财务费  
用是5781839.51元。"}  
{"id": 12, "question": "华灿光电股份有限公司2021年的法定代表人与上年相比相同吗?", "answer":  
"不相同, 华灿光电股份有限公司2020年法定代表人是俞信华, 2021年法定代表人是郭瑾。"}  

```

# 总结

## ## 1. 数据准备流程

### ### 主要步骤：

- **\*\*PDF 转 TXT\*\***：
  - 转换 PDF 文件为 TXT 格式。
  - 保留表格格式并合并相关单元格。
- **\*\*数据切分\*\***：
  - 基础信息：如公司名称、简称、英文全称等。
  - 财务数据：如合并资产负债表、合并利润表等。
  - 综合信息：如主要财务指标、公司业务概要等。
- **\*\*数据处理\*\***：
  - 计算各种基础公式：如营业成本率、投资收益率等。
  - 计算增长率。
  - 计算行业均值及排名。
- **\*\*存入数据库\*\***：
  - 存入 SQL、Mongo 和 ES 数据库中。
  - 涉及到建表及存表的过程。

## ## 2. 模型微调流程

- **\*\*数据分类\*\***：
  - SQL 数据、ES 数据、Mongo 数据及混合数据。
- **\*\*选择微调策略\*\***：
  - 包括但不限于 ptuningv2、lora、finetune 等。
- **\*\*进行微调\*\***：
  - 使用选定的策略对模型进行微调。

### ### 优点：

- 为特定数据和任务提供针对性优化。
- 支持多种微调策略，满足不同的需求。

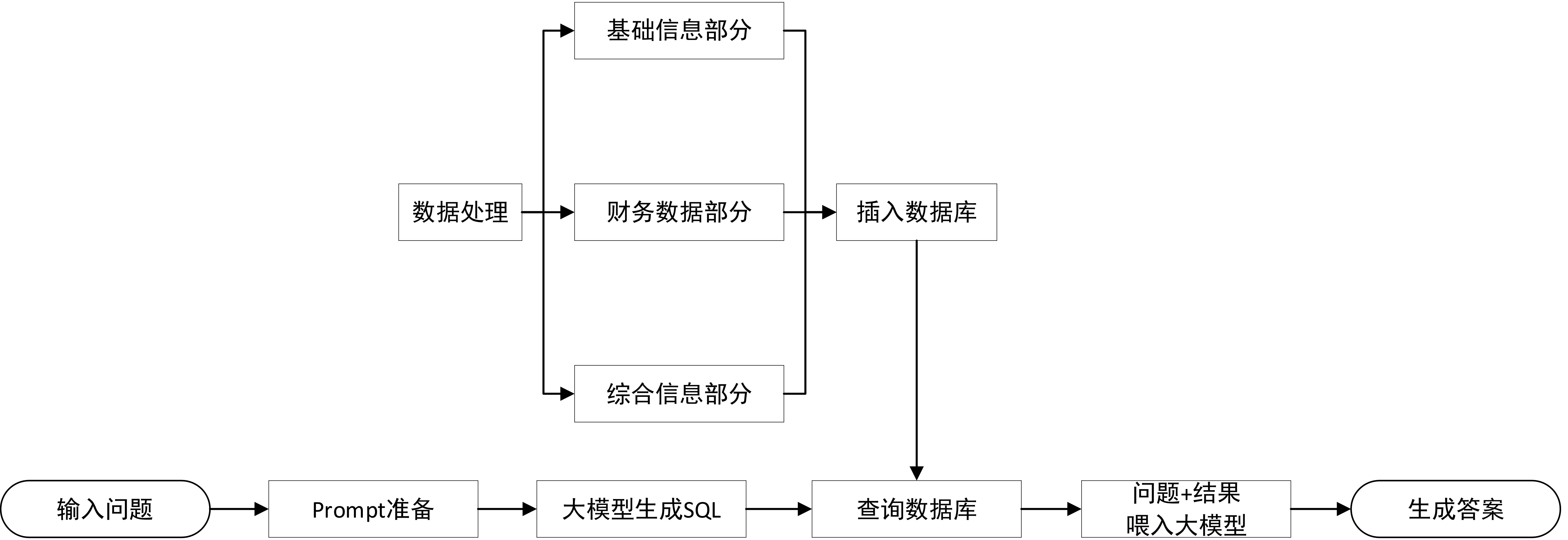
### ### 缺点：

- 微调过程可能耗时。
- 对于不同的数据可能需要重新微调。

## ## 3. 问答流程

- **\*\*输入问题\*\***。
- **\*\*Prompt 准备\*\***：根据问题，使用正则、相似度或树检索等方式生成对应的 prompt。
- **\*\*生成查询语句\*\***：根据 GPU 使用率，选择模型、API 或正则方式生成查询语句。
- **\*\*查询数据库\*\***：返回查询结果。
- **\*\*答案生成\*\***：利用问题和查询结果，使用 API 或原始 glm2-6b 模型生成答案。

# 总结



# 开源计划-FinGLM

## 1. 赛事转型

比赛将在天池平台转变为持续性学习赛，持续鼓励更多人参与和贡献。

## 2. 代码整合

- 所有参与开源的比赛选手的代码将被加入到FinGLM项目中。

- 与九支已经加入的团队(馒头科技、南哪都队、结婚买房代代韭菜、nsddd、Chatglm反卷总局、龙盈战队、小打小闹、东北大土豆、安硕硕眼探企)合作，共同审查代码并确保代码的质量。

## 3. 学习教程的准备

基于选手的代码，我们将整合并为大学生制备以下学习教程：

- 数据预处理教程
- 数据库使用教程
- GLM的使用教程
- Prompt编写教程
- 模型微调数据准备
- 模型微调技巧和步骤
- 全流程落地

## 4. 证书认证

参与比赛并为开源项目做出贡献的学生将获得官方证书，证明其在此领域的知识和贡献。



# 开源计划-FinGLM

## 5. 学生项目评价

我们会维护一个名单，该名单中列出了所有参与学生的项目完成情况。

## 6. 开源社区支持

- 参与开源项目的人员将参与答疑、教程设计，帮助新的参与者更好地了解项目和代码。
- 定期优化和更新项目代码，确保项目的持续健康发展。

## 7. 技术人员认证

为了感谢所有为项目做出贡献的技术人员，我们将实名公布他们的姓名和贡献。

## 8. 社区建设与激励

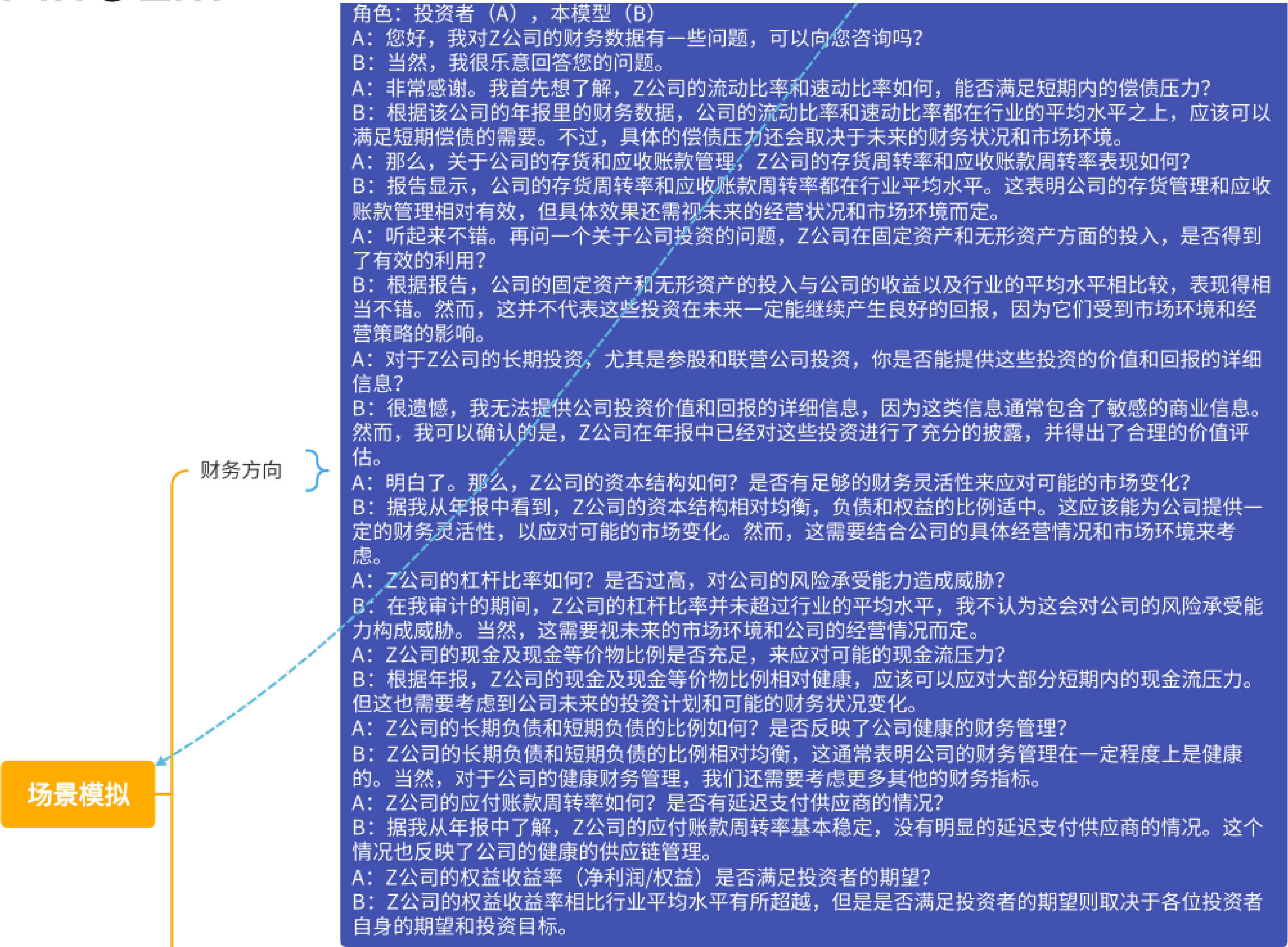
- 建立项目群，便于成员之间的沟通和合作。
- 定期发布任务，鼓励成员参与。
- 对于积极参与的成员，我们将给予费用支持。
- 定期评选并给予“特别贡献奖励”以激励成员。

## 9. 发布与同步

主要的开源代码库将位于魔搭社区。

GitHub 仓库将作为镜像存在，与魔搭社区上的代码库同步更新。

# 开源计划-FinGLM





智谱·AI

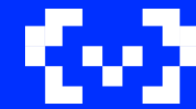


marsoft

| 安硕信息



北京交通大学  
BEIJING JIAOTONG UNIVERSITY



ModelScope  
魔搭社区



阿里云

Thanks\_