2024金融行业·大模型挑战赛

# 让大模型像人一样思考

队伍：ShallowRest_浅寻止步；

答辩人：郭学威

# 目录
# CONTENTS

01

PART ONE

# 参赛的背景与挑战

单击此处输入你的正文，文字是您思想的提炼，为了最终演示发布的良好效果，请尽量言简意赅的阐述观点；

第一单元

DEFENSE

# 参赛的背景（可能去掉）


## 新质生产力

网上找内容

### 市场需求

放市场数据

### XXXXX

中行单击此处输入你的正文，文字是您思想的提炼，为了最终演示发布的良好效果，请尽量言简意赅的阐述观点

### 愿景

跟人一样思考，xxxx

# 落地过程的挑战

## 极高准确率诉求

在特定垂直行业场景（如金融、党政等）中，对结果准确率的要求远超通用场景。在重要汇报时，即使正确率是99.99%，也是不能接受的。

概率事件 → 确定事件

## 算力资源瓶颈

为适配垂直领域知识体系，通常需微调通用大模型。但项目初期（如POC阶段）难提供足量资源，成为落地的关键障碍之一。

低资源 → 优效果

## 智能决策机制缺失

现实应用中存在大量“快与深”之间的权衡需求，对简单问题应快速响应，复杂问题应自主判断深度推理自纠正，实现“零交互纠错”。

02

PART TWO

# 总体方案

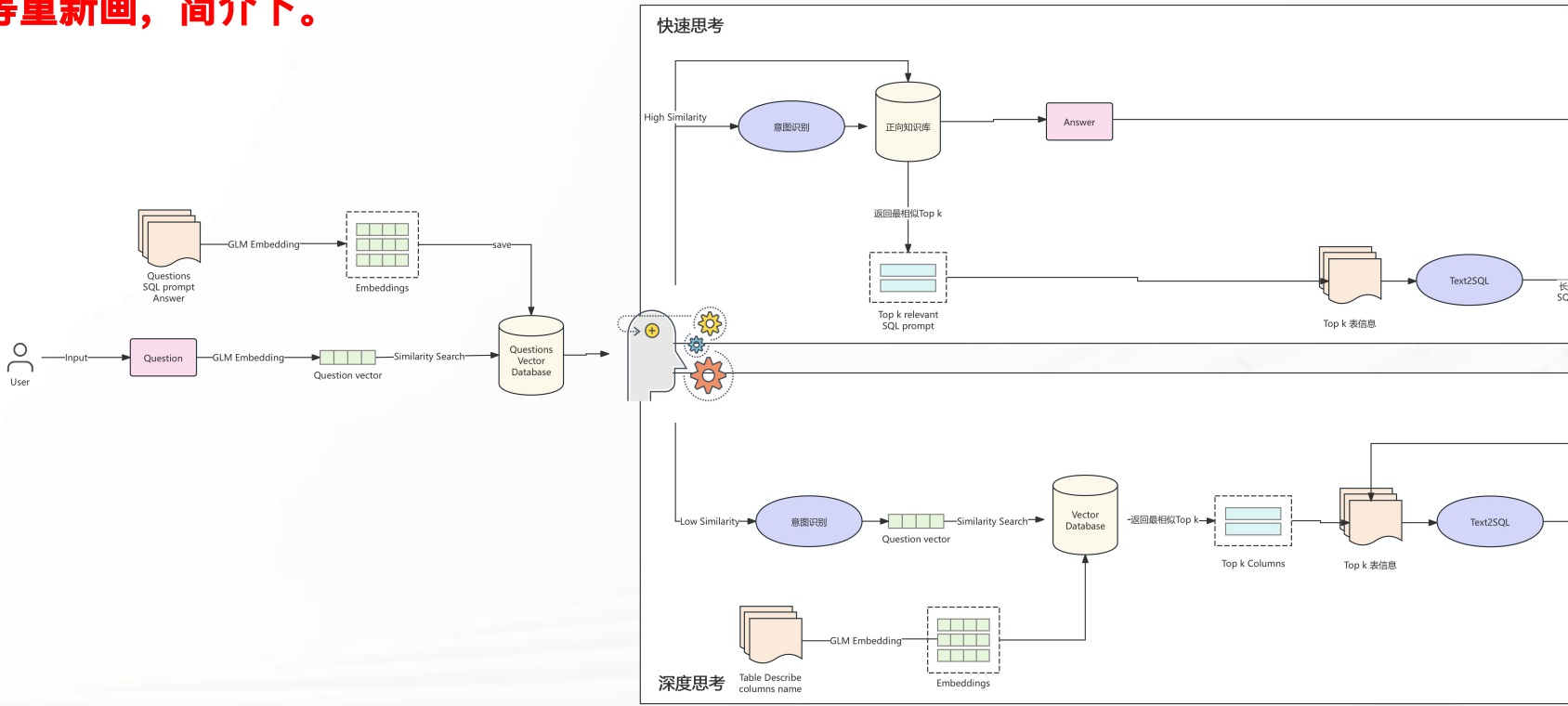
单击此处输入你的正文，文字是您思想的提炼，为了最终演示发布的良好效果，请尽量言简意赅的阐述观点；

第二单元

DEFENSE

# 架构图–像人一样思考

**图丑，等重新画，简介下。**

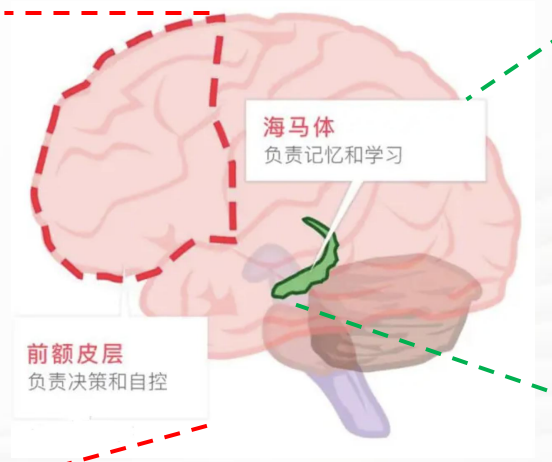


# 创新点-自适应思维机制

**自适应思维机制：快思考与深思考的智能协同**

我们提出一种自适应思维机制，根据自然语言查询与知识库的语义匹配度智能选择不同处理路径。

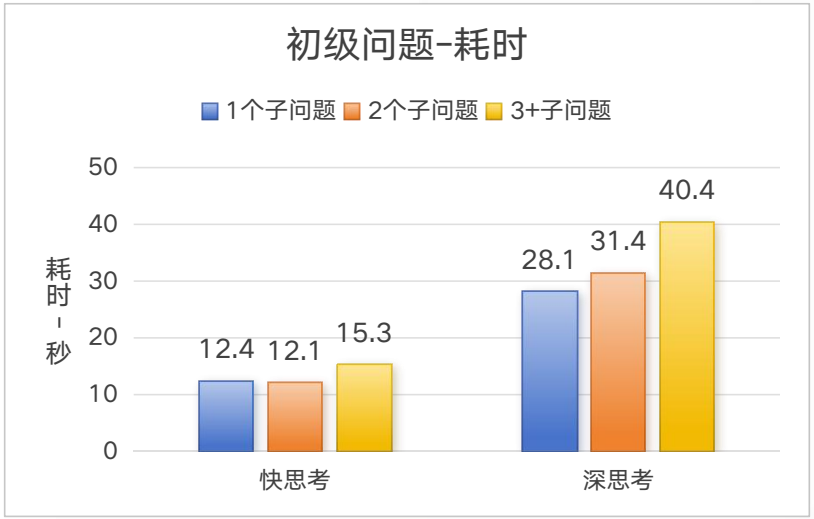
深思考模式：当匹配度较低或问题较复杂时，系统采用多轮链式推理，通过深度搜索模式，将复杂查询拆解成若干子查询进行自纠正分解推理，逐步生成最终答案。

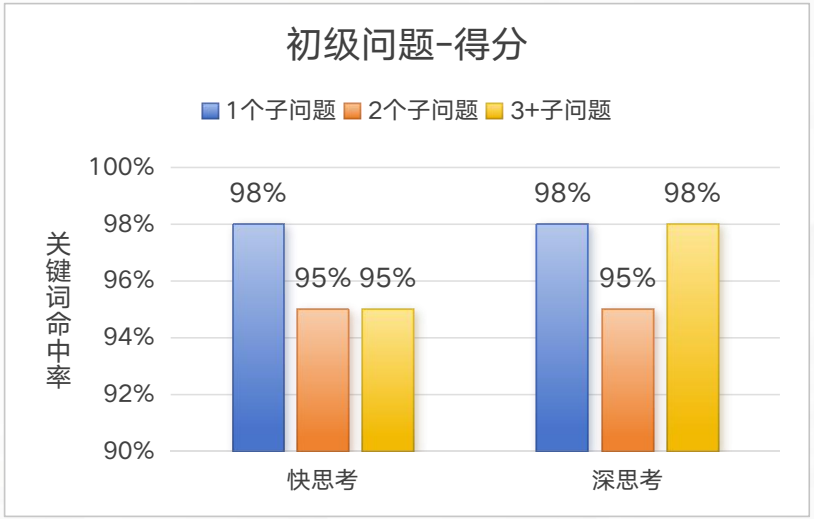



快思考模式：当查询与知识库匹配度高时，系统依托动态SQL promt，生成一步到位的 SQL，响应迅速且精准。

# 创新点–自适应思维机制

解答初级（简单）问题的时，快思考比深思考更具效能优势（**超7倍速度**）。

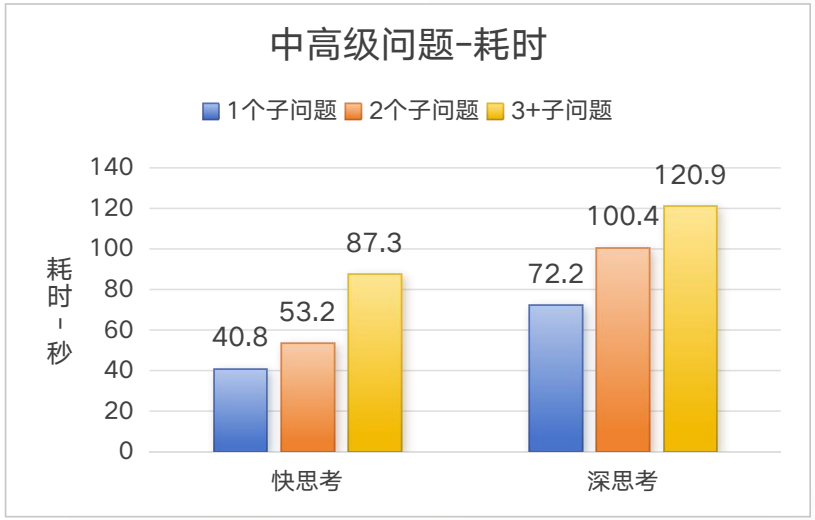


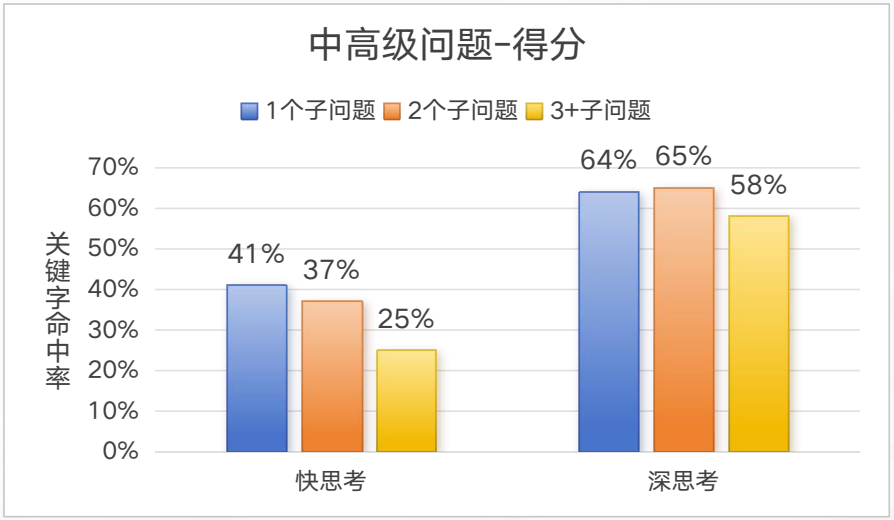




注：单进程顺序执行，由于样本数量少且大模型速度和策略存在一定波动，数据仅为当前条件下的准确率，存在不稳定性。

# 创新点-自适应思维机制

解答中高级（中难）问题的时，深思考比快思考更具性能优势（**近一倍效果提升**）。

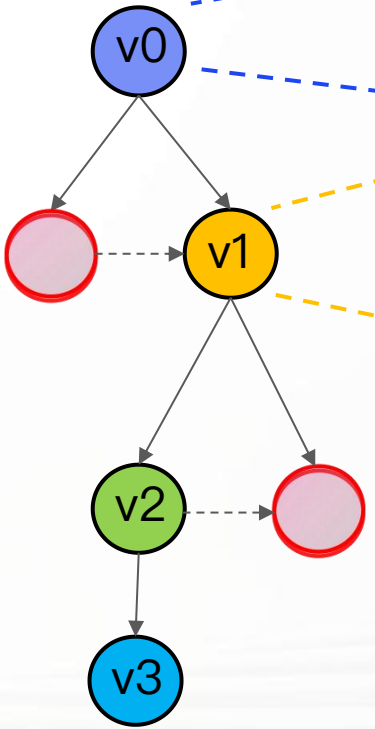



注：
1. 子问题：1组问答中的1问有多少个子问题。
2. 单进程顺序执行，由于样本数量少且大模型速度和策略存在一定波动，数据仅为当前条件下的准确率，存在不稳定性。

# 创新点–自纠正分解决策树

能够自我选择和校验，从而选到更优的答案。




问题：“600872的全称、A股简称、法人、法律顾问、会计师事务所及董秘是？”
文本：通用prompt、相似SQL prompt、库表信息、多步拆解提示。

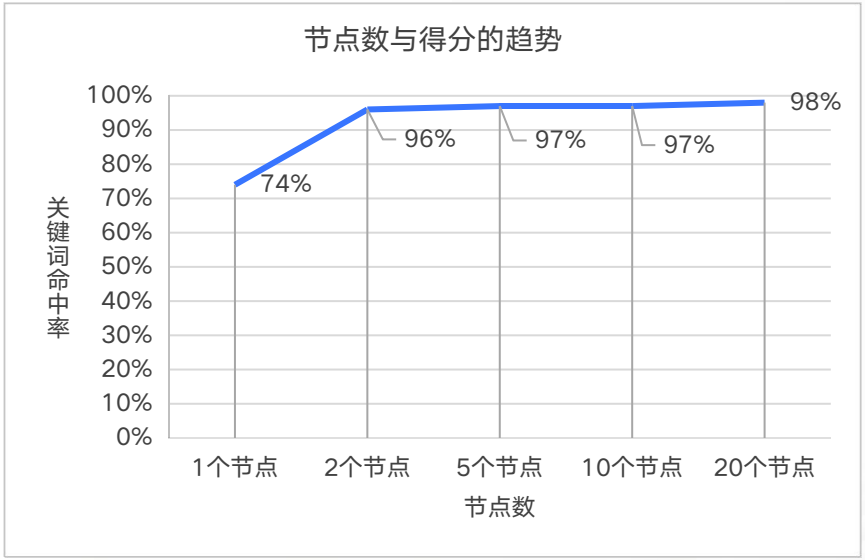
后剪枝：推断选择更优的。

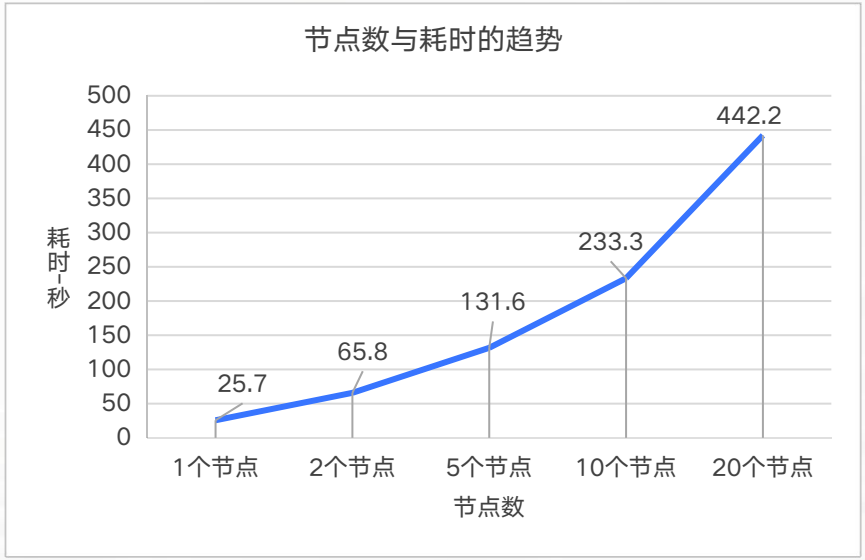
**选最优（相同则选前者）**
node1生成：“SELECT EngName AShareAbbr.....”
文本：“这是node1的结果，根据上文生成更优的结果”
选择：“node2是更优回答‘SELECT ChiName AShareAbbr.....’”
结果：“中炬高新.....”

**根据返回结果规划下一步**
文本：“.....上一轮的sql结果，这是原问题，如果没回答完成，规划下一步”
node1生成：“SELECT .....”
选择：“node1已是更优回答‘SELECT .....’”
结果：“中炬高新.....”

# 创新点–消融实验

当节点数为2以上后，关键词命中率不再有明显上涨，但耗时依然呈现正相关。




注：上述数据为34个正确答案样本的线下测试结果。

# 创新点-消融实验

微调9B模型
我们的方案+9B
我们的方案+4-plus

要对比吗?
不如我们的方案模型。
对比Base，84.2。

注：上述数据为34个正确答案样本的线下测试结果。

YOUR LOGO
稻壳学院

# 03 性能和效能指标

PART THREE

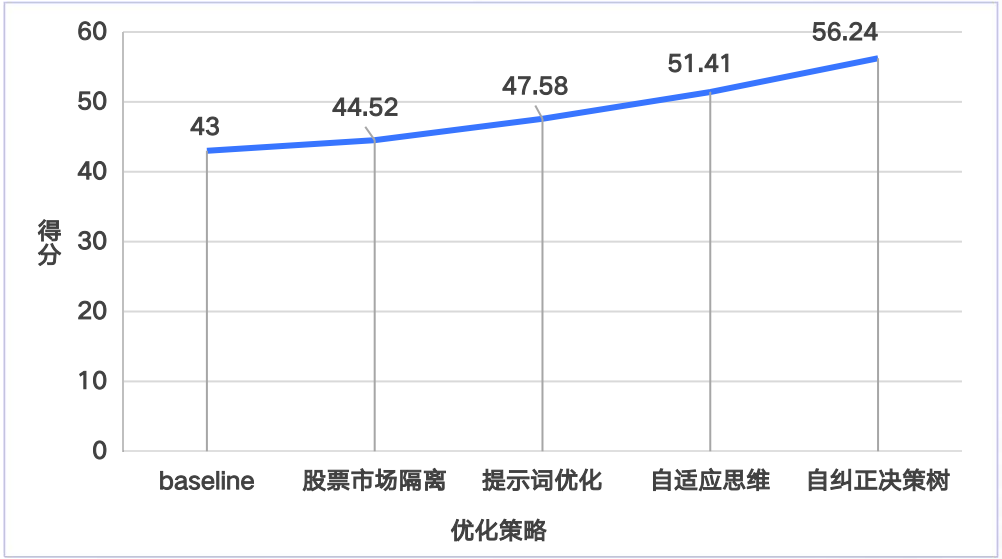
单击此处输入你的正文，文字是您思想的提炼，为了最终演示发布的良好效果，请尽量言简意赅的阐述观点；

第三单元

DEFENSE

# 性能指标

以B榜为主的分数涨分路径：




- **baseline：**多轮交互、date格式优化和表召回等。
- **股票市场隔离：**分割A股、港股和美股，避免大模型使用错误股票市场表。
- **提示词优化：**强化代词替代、多轮交互优化等。
- **深度检索：**当问题与正向知识库相似度过低，则启动“库表描述”、“列中文名”的Embeding匹配等。
- **自纠正决策树：**参考第11页。

注：
1. 感谢baseline分享者“公交车的轮子转啊转”。
2. Embedding：GLM的Embedding-3文本向量模型。

# 性能指标–复现情况

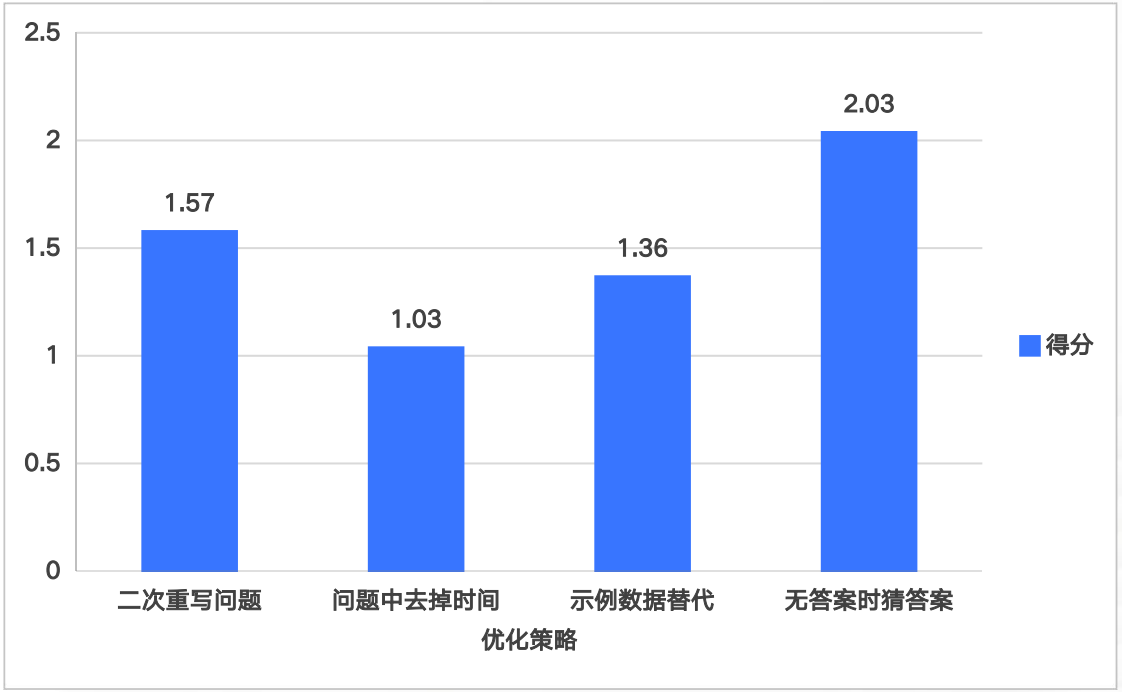
前十队伍里，唯一在官方复现中涨分的：

| 团队名称 | 复赛得分1 (A*0.3+B*0.7) | 复赛得分2 (A*0.3+B复现*0.7) | 总涨跌值 | b榜涨跌值 |
|---|---|---|---|---|
| xxxx | 65.53 | 63.97 | -1.56 | -2.24 |
| xxxx | 64.89 | 63.77 | -1.12 | -1.59 |
| xxxx | 64.05 | 62.62 | -1.43 | -2.04 |
| xxxx | 63.72 | 62.35 | -1.37 | -1.96 |
| xxxx | 63.01 | 61.72 | -1.29 | -1.84 |
| 我们 | 58.68 | 58.79 | 0.11 | 0.16 |
| xxxx | 56.79 | 56.58 | -0.21 | -0.3 |
| xxxx | 57.63 | 55.24 | -2.39 | -3.42 |
| xxxx | 57.07 | 54.73 | -2.34 | -3.35 |
| xxxx | 55.05 | 54.25 | -0.8 | -1.14 |

注：
1. 总涨跌值 = 复赛得分2 – 复赛得分1。
2. b榜涨跌值 = b榜官方复现得分 – b榜得分

# 性能指标–未使用但有效的

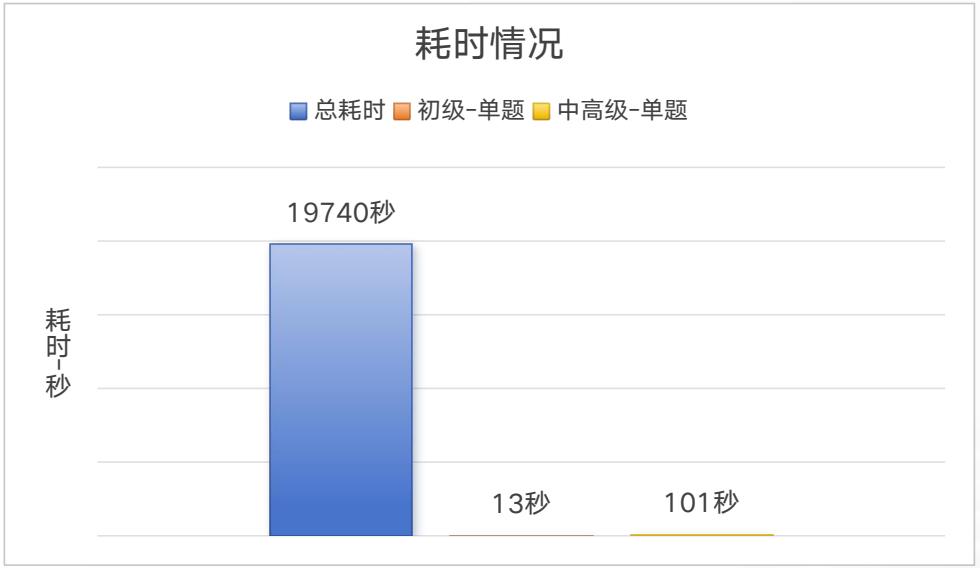
较高得分但不落地性差的（**约5.99分**）：




- **二次重写问题：**针对第一轮的重写问题，进行二次重写优化。
- **问题去掉时间：**解决可能不需要时间的问题。
- **示例数据替代：**解决大模型编写的SQL使用了示例数据而非实际数据问题。
- **无答案时猜答案**：解决问题过难时间无答案的情况。

# 效能指标

简单问题**平均13秒**即可完成结果输出。
复杂问题平均101秒完成结果输出。




待美化，还比较空旷

注：
1. 总耗时：100组问答完成所需的时间。
2. 单题：1组问答中的1问。

YOUR LOGO
稻壳学院

# 04 业务价值和应用前景

PART FOUR

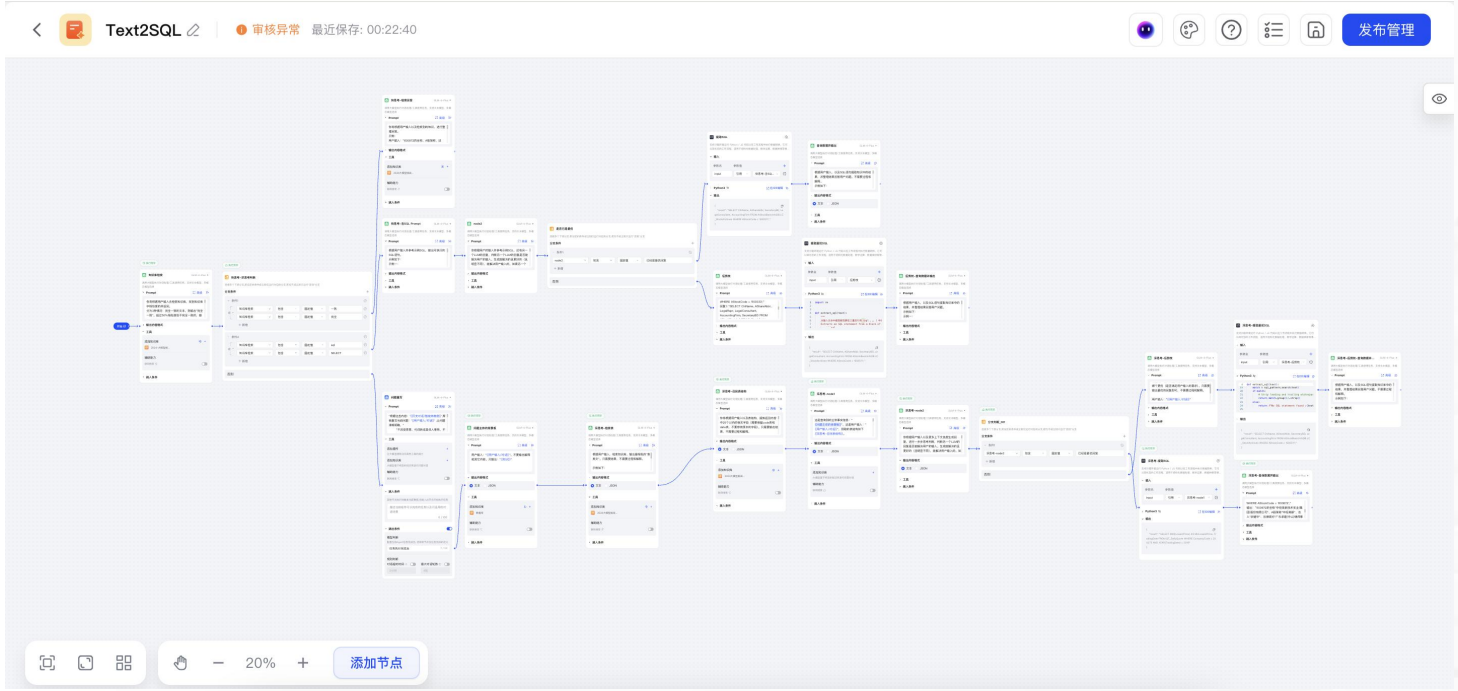
单击此处输入你的正文，文字是您思想的提炼，为了最终演示发布的良好效果，请尽量言简意赅的阐述观点；

第四单元

DEFENSE

# 应用落地–基于BigModel的智能体

待变成视频



# 业务价值和应用前景

单击此处输入你的正文，文字是您思想的提炼

指业务价值

1）某个细分场景（金融类）：实际落地情况：南山经济分析和数据汇聚计算。

2）我们方案的可解决哪些痛点；

2.1 部分答案为确定事件。解决关键汇报人担忧的错误问题。

2.2 落地速度加快。资源简化，不再需要大量算力做微调，提高用户的SQL编写速度，加速数智化转型。

2.3 自适应思维机制。极大降低资源浪费，沉淀知识库，让模型越用越聪明。

指该技术或方案在未来是否具有广泛的扩展性和成长空间。

能不能推广到更多场景？明亚、小悦智法

未来有没有发展趋势？未来一定是更加智能的，不需要我们去选择到底是不是深度思考。

是否能引领行业方向？政务金融领域的100%准确率。lazy prompting

未来在ai加持下的大模型形态：更像人类最后超越人类。

让我们一起迈向AGI

# 感谢大家观看