

2024金融行业·大模型挑战赛

浅尝止步-让大模型像人一样思考

演讲人：郭学威

时间：2025年4月19日



# 目录

**01 大模型落地过程中的挑战**

**02 总体方案**

**03 性能和效能指标**

**04 业务价值和应用前景**



01

# 大模型落地过程中的挑战



# 落地过程中的挑战

## 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。

## 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。

## 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”。

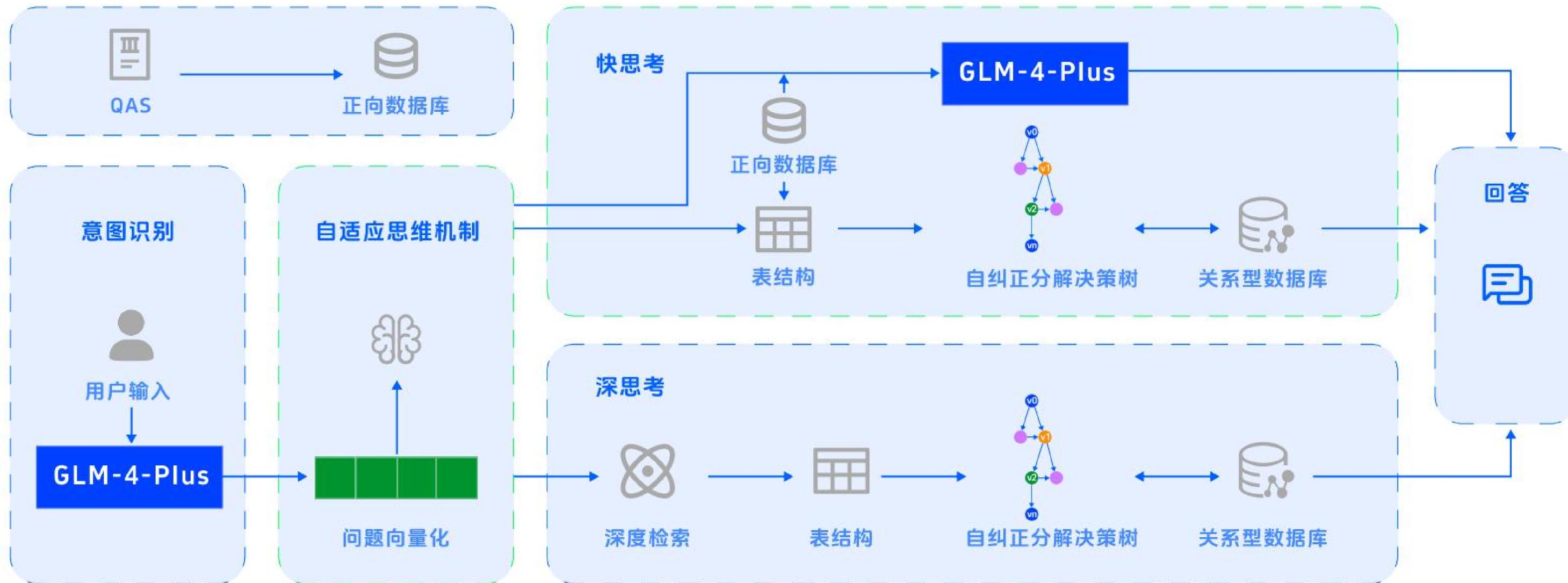


# 02

## 总体方案



# 02 总体框架



以“智谱大模型”为核心，当用户输入问题后，由自适应思维机制进行决策调度，简单的问题则快速思考并回答。复杂则通过深度检索，并配合自纠正分解决策树来整理多步的查询结果进行回答。



# 02 实机演示-基于智谱的智能体

The screenshot shows a web browser window with the URL `open.bigmodel.cn/console/appcenter_v2/flow?app_id=1911648317289496576&detail_id=1912157182724538368`. The page title is "Text2Result" and it indicates it has been published. The interface is divided into three main sections:

- Flowchart (Left):** A complex workflow diagram with multiple nodes connected by arrows. Each node contains configuration details for different AI models, including prompts and parameters.
- Preview/Debug (Middle-Right):** A section titled "预览调试" (Preview/Debug) showing the agent's output. It features a large orange icon with a document and the text "Text2Result". Below this, a welcome message reads: "欢迎来到Text2Result! 这是由“浅尝止步”开发的智能体。你可以通过自然语言获取数据库中的结果!" (Welcome to Text2Result! This is an AI agent developed by "Shan Chang Zhi Bu". You can get results from the database through natural language!). There are three example prompts in rounded rectangles:
  - 600511的全称、A股简称、法人、法律顾问、会计师事务所及董秘是?
  - 厦门钨业在2019年全年的最低收盘价是多少, 出现在哪一天, XXXX年XX月XX日?
  - 发送消息, 点击 Shift + Enter 实现换行输入At the bottom of this section is a search icon and a blue play button.
- Logs (Right):** A section titled "日志" (Logs) which currently displays "暂无数据" (No data).

At the bottom left of the interface, there are navigation icons (back, forward, home) and a zoom control set to 20%, along with a button labeled "添加节点" (Add Node).

# 02 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。

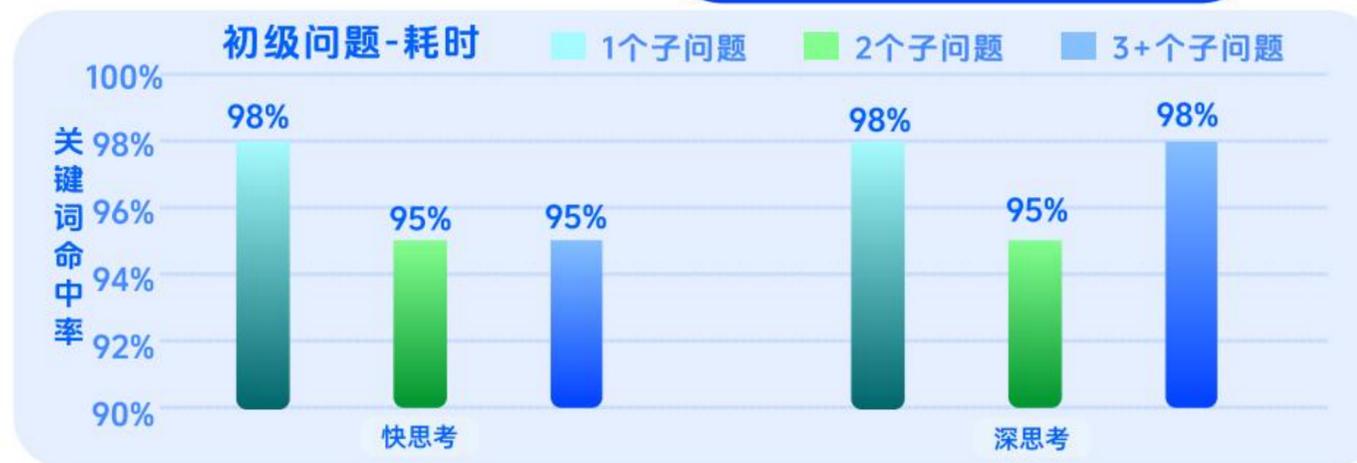
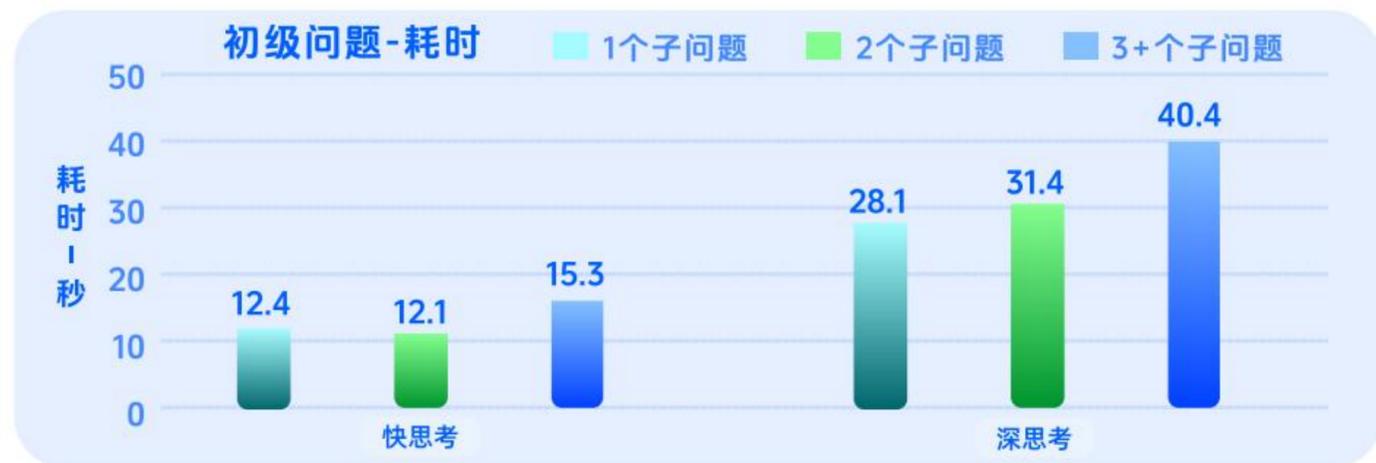


注：

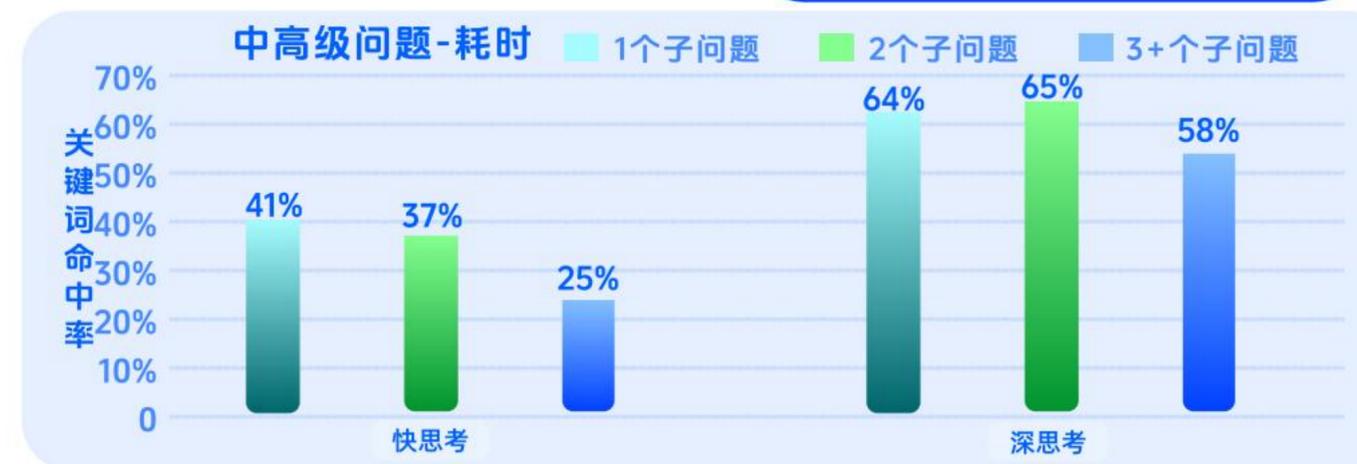
1. 正向知识库：存储问题、对应答案和对应的SQL，并向量化问题去进行相似度匹配。
2. 首召表：通过表描述快速召回的1-2个表结构。
3. 相似度的阈值：匹配度是通过智谱的Embedding-3模型生成的向量去做匹配，通过多轮问答，根据答案的正确程度，来调整相似度的阈值。

# 02 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 **（超约3倍速度）**



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**



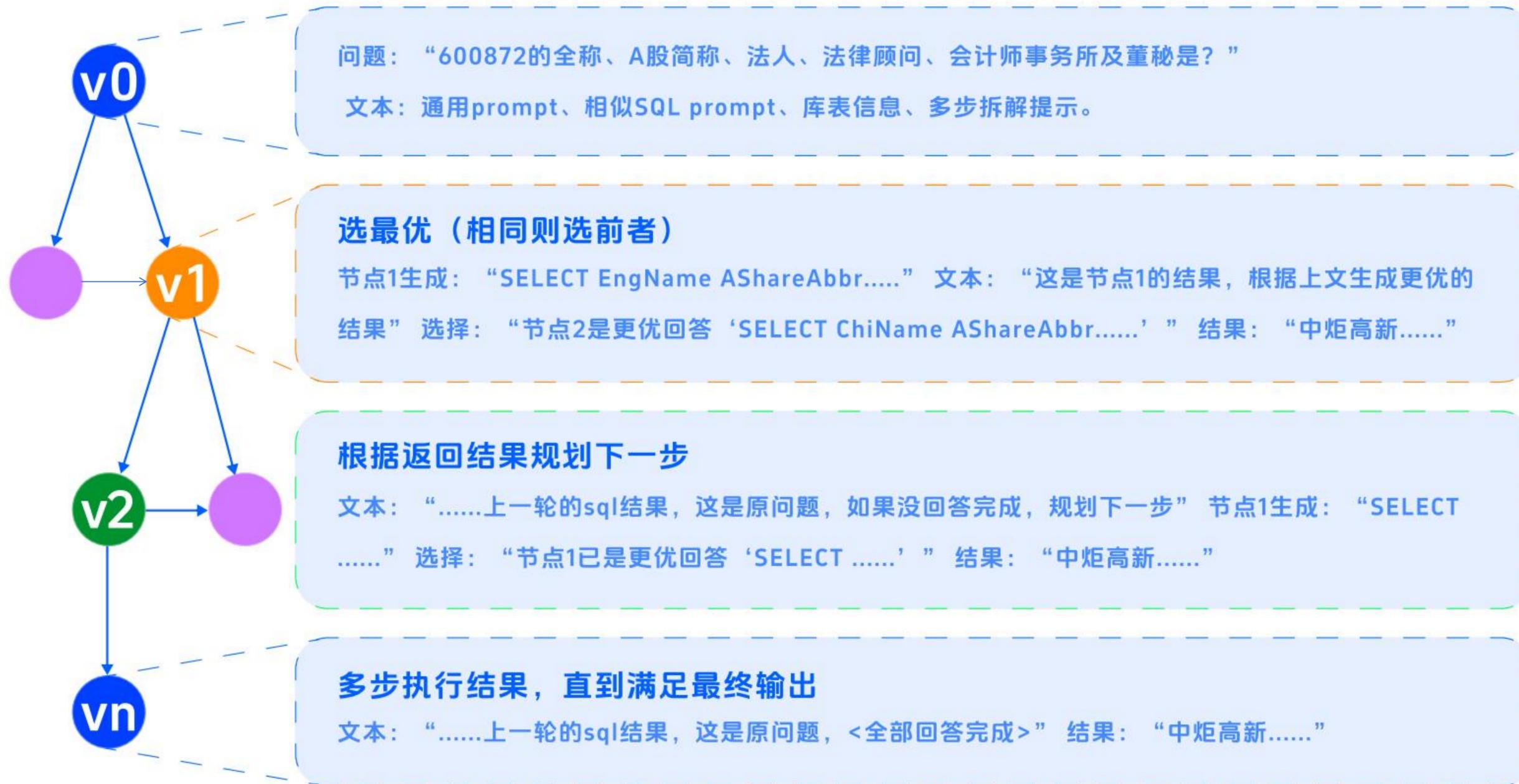
注：

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。

# 02 创新点-自纠正分解决策树

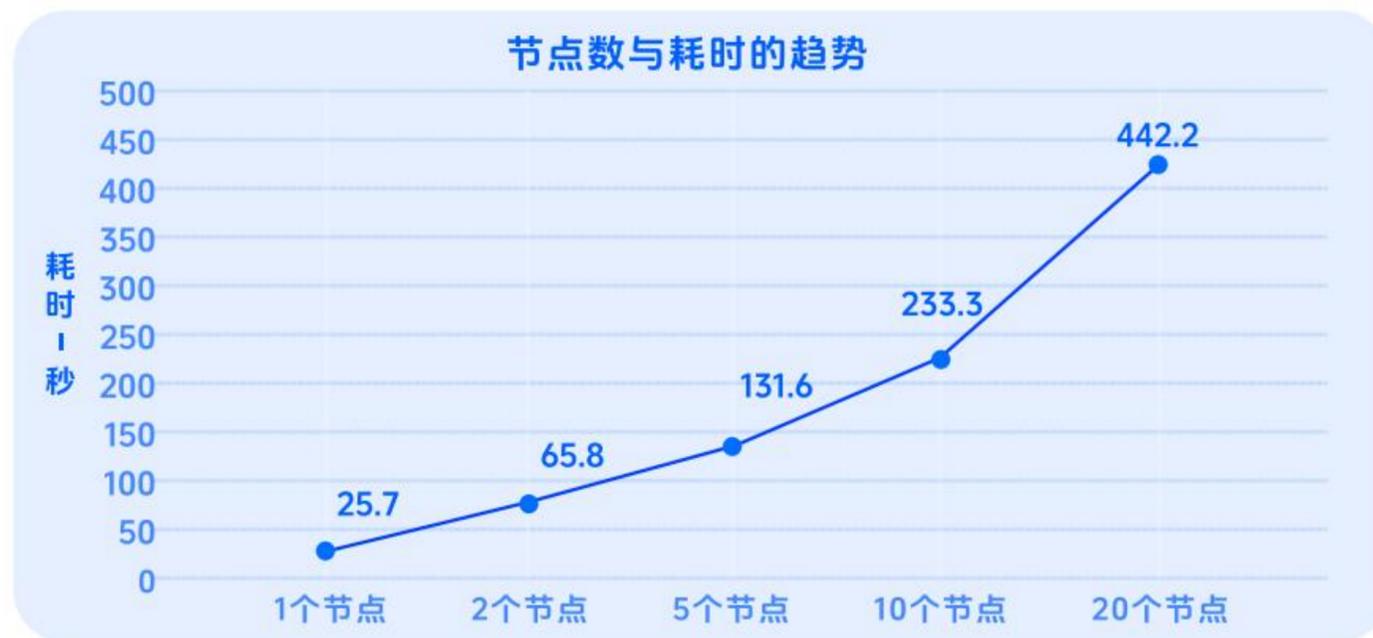
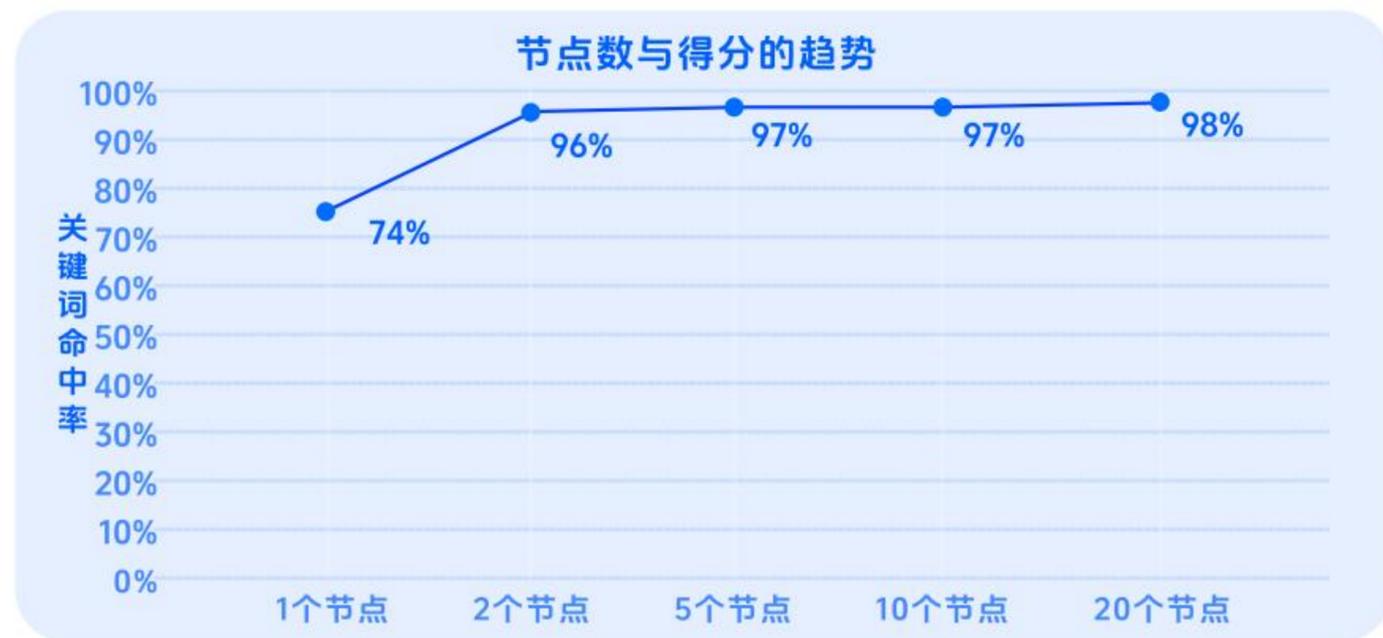
能够自我选择和校验，从而选到更优的答案。

后剪枝：推断  
选择更优的。



# 02 创新点-自纠正分解决策树-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。



Method	Model	Score	Avg Time
仅微调	GLM-4-9B	65%	25.7
自纠正分解决策树	GLM-4-9B	79%	67.2
自纠正分解决策树	GLM-4-Plus	96%	65.8

注：

1. 上述数据为34个正确答案样本的线下测试结果。
2. 可能因为数据量较小，微挑的GLM-4-9B正常情况下也不如Plus

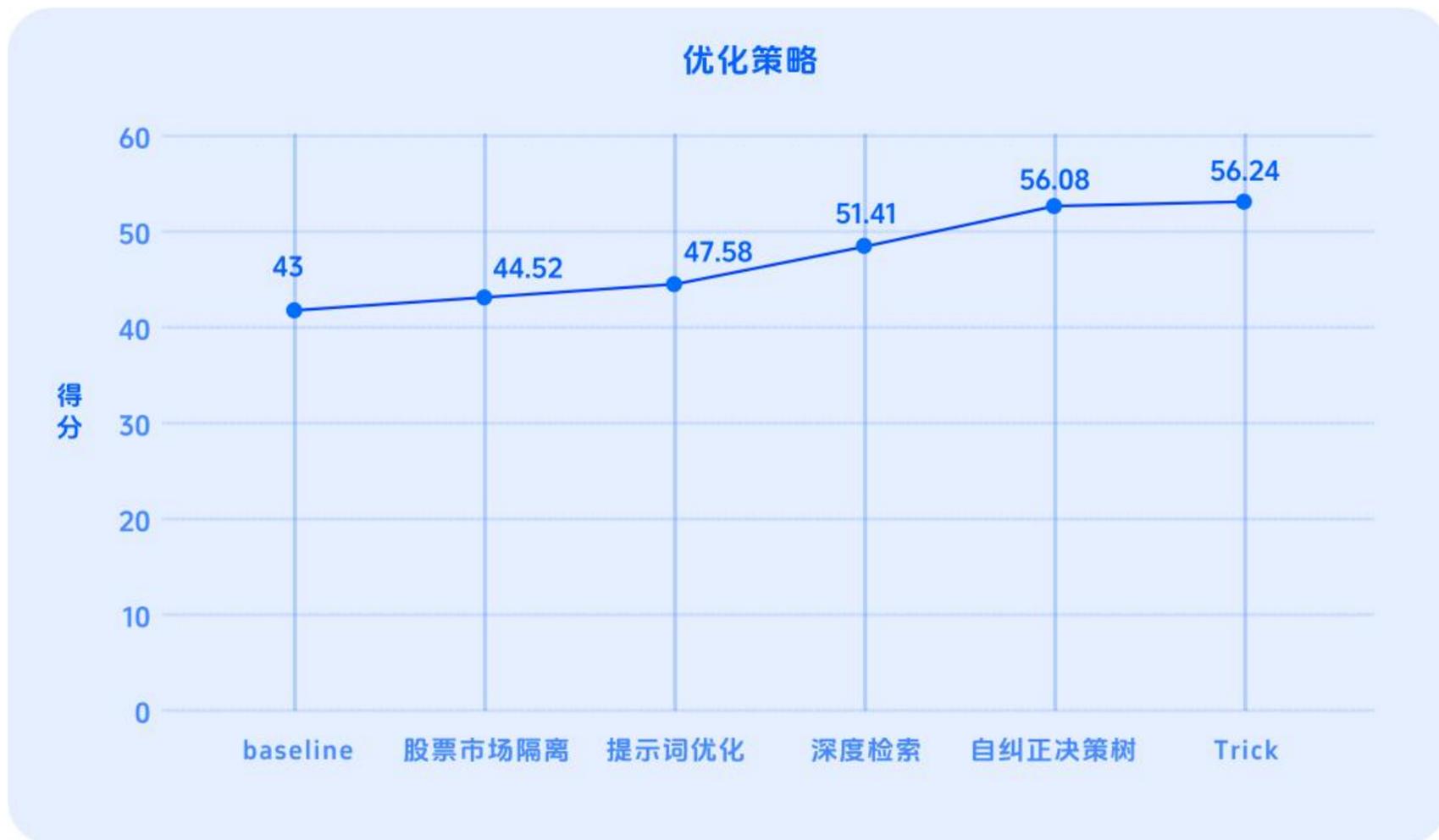
03

性能和效能指标



# 03 性能指标

以B榜为主的分数涨分路径：



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。

## 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

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Embedding匹配等。

## 自纠正决策树

参考第11页。

## Trick

二次回答、去掉时间的回答等。

## 03 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)	总分涨跌值	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
<b>我队</b>	<b>58.68</b>	<b>58.79</b>	<b>0.11</b>	<b>0.16</b>
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. 总分涨跌值 = 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

## 03 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。

### 耗时情况



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

04

业务价值和应用前景



# 04 业务价值和应用前景

## 准确性保障 ①

部分结果100%，有效缓解关键汇报人对数据错误的担忧。

## 快速落地部署 ②

无需大模型微调，支持轻量部署和快速接入现有数据库系统，加快企业数智化转型。

## 极大降低成本 ③

通过自适应思维机制降低人力与算力资源浪费。

### 业务价值

## 广泛业务适配性

可应用于政务经济分析、城市治理、产业监测、企业运营等多类场景，推动数据驱动决策的智能化升级。

## ★ 重构BI查询体验

从传统拖拽式分析升级为自然语言对话式查询，服务领导决策、日报分析、趋势洞察等业务需求。

## 🗣️ 多智能体协同发展

可拓展接入推荐系统、图表生成、策略优化等智能体，构建企业级“认知中台”，实现复杂任务智能协作。

### 应用前景

Thanks

# 让我们一起迈向AGI

演讲人：郭学威

商务合作

[service@zhipuai.cn](mailto:service@zhipuai.cn)

北京总部

北京市海淀区搜狐网络大厦10层

