



中国科学院大学

University of Chinese Academy of Sciences



中国科学院高能物理研究所

Institute of High Energy Physics

Chinese Academy of Sciences

Shale-Agent: 一种用于加速科研 知识发现的Text-to-Query智能体架构

—————• 第二十一届全国科学计算与信息化会议 •—————

报告人: 易杨

单 位: 高能所

日 期: 2025年8月25日

- 一、背景与挑战
- 二、研究内容
- 三、实验验证
- 四、总结与展望

科研中的“信息迷航”



- 绝大多数前沿科学知识以非结构化文本（如PDF）的形式存在，其指数级增长导致关键数据被“禁锢”于海量文档中，难以进行程序化访问。
- 实验条件、样本背景、微妙的因果结论等隐性知识，常以叙事性语言描述，缺乏统一范式。
- 基于关键词的传统检索技术无法深刻理解科研人员的复杂查询意图（如“对比A和B在C条件下的差异”），导致返回结果相关性差，需要耗费精力进行二次人工筛选。

**科学文献数量呈指数级增长，知识获取难度大
科研人员花费大量时间进行重复性的文献阅读、数据查找与对比。**

场景：中国散裂中子源的页岩孔隙结构研究

任务： “基于SAS探究不同岩相页岩孔径分布或连通性特征”

传统流程

1.检索

- 在论文库中用关键词搜索，下载几十篇相关PDF

2.阅读定位

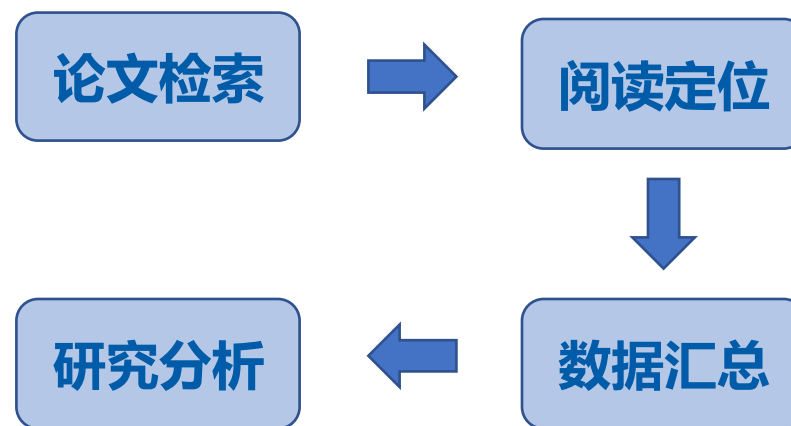
- 逐篇打开，快速浏览，找到描述实验和结果的章节

3.手动复制

- 将文字、表格中的数据手动复制到Excel

4.清洗对齐

- 统一不同的单位、命名法和格式



一个简单的数据对比任务，可能耗费一整天甚至更久！

研究现状

数据库中有28篇文章，回答只给出了3篇文章中的结果

根据提供的知识库内容，以下是所有提到的地层名称 (Formation Name)：

Formation Names

FORMATION NAME	DOCUMENT ID	SOURCE
Barnett Shale	ID=02	Ruppert-et-al-USANSSANS-Study-of-the-Accessibility-of-Pores-in-the-Barnett-Shale-to-Methane-and-Water
Marcellus Shale	ID=02	Ruppert-et-al-USANSSANS-Study-of-the-Accessibility-of-Pores-in-the-Barnett-Shale-to-Methane-and-Water
Woodford Shale	ID=02	Ruppert-et-al-USANSSANS-Study-of-the-Accessibility-of-Pores-in-the-Barnett-Shale-to-Methane-and-Water
Longmaxi Shale	ID=03	Zhang-et-al-Characterization-of-Closed-Pores-in-Longmaxi-Shale-by-Synchrotron-Small-Angle-X-ray-Scattering
Bakken Formation	ID=03	Zhang-et-al-Characterization-of-Closed-Pores-in-Longmaxi-Shale-by-Synchrotron-Small-Angle-X-ray-Scattering
Wolfcamp Formation	ID=05	Neil-et-al-Small-Angle-Neutron-Scattering-Investigation-of-Oil-Recovery-in-Mineralogically-Distinct-Wolfcamp-Shale-Strata

Sources

Each document contains references specific geological formations studied:

Barnett/Marcellus/Woodford Formations

From [Ruppert-et-al.-A-USANSSANS-Study-of-the-Accessibility-of-Pores-in-the-Barnett-Shale-to-Methane-and-Water]

Longmaxi/Bakken Formations

列举出数据库中所有的formation name

列举出数据库中所有的sample

根据提供的知识库内容，以下是从相关文档中提取的所有样本 (Sample) 信息：

“胡言乱语”

Samples Listed Across Documents

From Document Fragments Provided

```
markdown
# Samples Identified In Database Files
1
2
3 + Raw shale sample (from Hjelm et al., Flow-through compression cell study)
4 + Empty compression cell background (from Hjelm et al., Flow-through compression cell study)
5 + Sample corrected using MT oedometer method (from Hjelm et al., Flow-through compression cell study)
```

Additional Samples Found In Other Documents Not Included Initially But Relevant To Query Scope

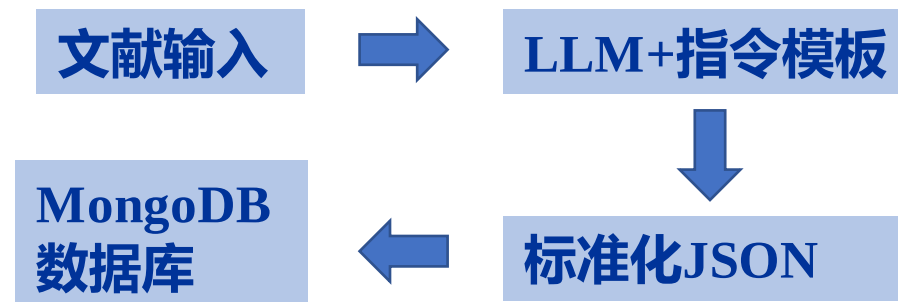
```
markdown
1 vick dike sill vein lode ore deposit mineral resource natural reserve wildlife habitat ecosystem biome b
2
3 v
4
```

主流知识库——大模型+RAG向量数据库检索模式回答问题会有遗漏且容易编造内容

Shale-Agent设计思路

阶段一：知识结构化 (文献 -> 知识)

- 利用大语言模型（LLM）的强理解力，将非结构化文本自动、精准地转化为结构化知识库。



阶段二：智能查询交互 (提问 -> 答案)

- 理解用户复杂的、模糊的自然语言提问，并从数据库中查询、推理、整合并以最友好的方式呈现答案



阶段一：知识结构化

挑战： 如何让LLM在抽取信息时不产生幻觉、不遗漏关键点？

方法： 设计一个结构化指令模板，直接定义要抽取的实体，定义它们之间的嵌套关系和数据格式

优势：

- 高精度度： 强约束减少了模型“自由发挥”的空间。
- 零样本/少样本： 无需为新领域重新训练模型，只需设计新模板

```
⊖{
  "experiment": ⊖{
    "SAS_experiments": ⊖{
      "small_angle_neutron_scattering": ⊖{
        "wavelength": 8,
        "Δλ/λ": 0.15,
        "Q_range": ⊕Array[2],
        "pore_size_range": ⊕Array[2]
      },
      "ultra_small_angle_neutron_scattering": ⊕Object{...},
      "small_angle_X_ray_scattering": ⊕Object{...},
      "sample_shape": "powder",
      "sample_thickness": null,
      "contrast_fluids": null,
      "fitting_models": "Polydisperse spheres model",
      "error_analysis_method": null
    },
    "mercury_intrusion_porosimetry": ⊖{
      "pressure_range": ⊕Array[2]
    }
  },
  "sample": ⊖{
    "sample_info": ⊖{
      "sample_label": "Ziliujing Formation shale sample",
      "sample_source": ⊕Object{...},
      "sample_depth": null,
      "mineral_composition": ⊕Object{...},
      "TOC": 0.64,
      "kerogen_type": null
    },
    "experiment_results": ⊕Object{...}
  },
  "paper": ⊖{
    "controlling_factors_of_pore_accessibility": "Thermal maturity and fluid accessibility significantly control pore accessibility in shale.",
    "the_effects_of_temperature_on_pore_size_distribution": "Pore volume and surface area first increase then decrease with rising temperature, and larger pores become more prevalent.",
    "the_effects_of_pressure_on_pore_size_distribution": null,
    "conclusions": ⊖{
      "conclusion_1": "Small angle neutron scattering captures closed and open pores, unlike fluid-based methods.",
      "conclusion_2": "Shale pore volume is mainly concentrated in the 5 to 100 nm range, and SANS yields higher porosity values than MICP and nitrogen adsorption.",
      "conclusion_3": "Differences among methods are largely due to fluid accessibility constraints.",
      "conclusion_4": "Elevated thermal maturity reduces the fraction of closed pores, enhances connectivity, and alters pore-size distribution."
    }
  },
  "title": ⊖{
    "title_of_the_paper": "The Multitype Pore Structure Evolution Characteristics of Ziliujing Formation Shale from the Northeastern Sichuan Basin in China: Experimental Study Using Small Angle Neutron Scattering Technology"
  }
}
```

nts

-ray and

'sorption

us_rank

tering

ssure

istration

I_FIB-SE

techniqu

ase_stu

阶段二：智能查询交互

一、语义理解 (LLM1)

- 利用大语言模型强大的自然语言理解能力，将用户问题和对话上下文，转化为一个结构化的、包含查询意图的句。

二、查询解析与执行 (Code)

- 通过代码中确定性的查询解析器，精确映射为数据库的完整字段路径，生成最终的MongoDB Query并执行。

三、结果生成 (LLM2)

- 将从MangoDB数据库中获取的原始数据，再次交由LLM2进行总结提炼和格式化，以最符合用户问题解答的方式呈现最终答案。

该样品的实验结论

OI shale_agent

智能体执行过程 (点击展开) ^

[阶段 1/3] 正在结合上下文，生成查询意图...

```
json
1 {
2   "filter": {
3     "sampleId": 696571
4   },
5   "projection": {
6     "结论1": 1,
7     "结论2": 1,
8     "结论3": 1,
9     "结论4": 1,
10    "_id": 0
11  }
12 }
```

[阶段 2/3] 正在将查询意图翻译成真实MongoDB查询...

```
"翻译后的真实查询: {'filter': {'sampleId': '696571'}, 'projection': {'_id': 0, 'sampleId': 1, 'sampleCharacteristics.paper.conclusions.conclusion_1': 1, 'sampleCharacteristics.paper.conclusions.conclusion_2': 1, 'sampleCharacteristics.paper.conclusions.conclusion_3': 1, 'sampleCharacteristics.paper.conclusions.conclusion_4': 1, 'sampleCharacteristics.sample.sample_info.sample_label': 1}}"
```

[阶段 3/3] 正在执行数据库查询...

"在 'Sample' 集合中查询到 1 条记录。"

[阶段 4/4] 正在为您总结查询结果...

阶段二：智能查询交互

一、语义理解 (LLM1)

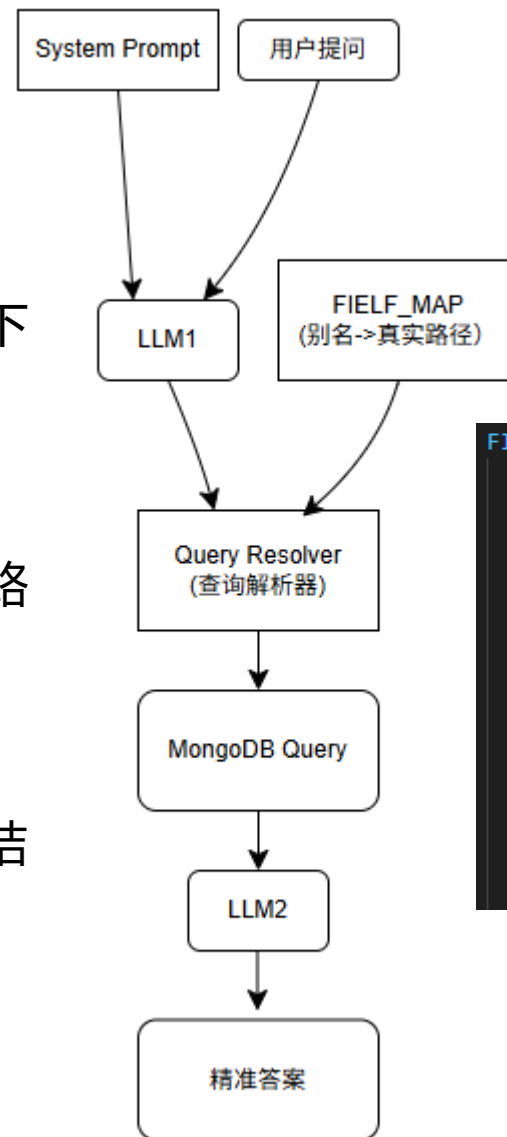
- 利用大语言模型强大的自然语言理解能力，将用户问题和对话上下文，转化为一个结构化的、包含查询意图的语句。

二、查询解析与执行 (Code)

- 通过代码中确定性的查询解析器，精确映射为数据库的完整字段路径，生成最终的MongoDB Query并执行。

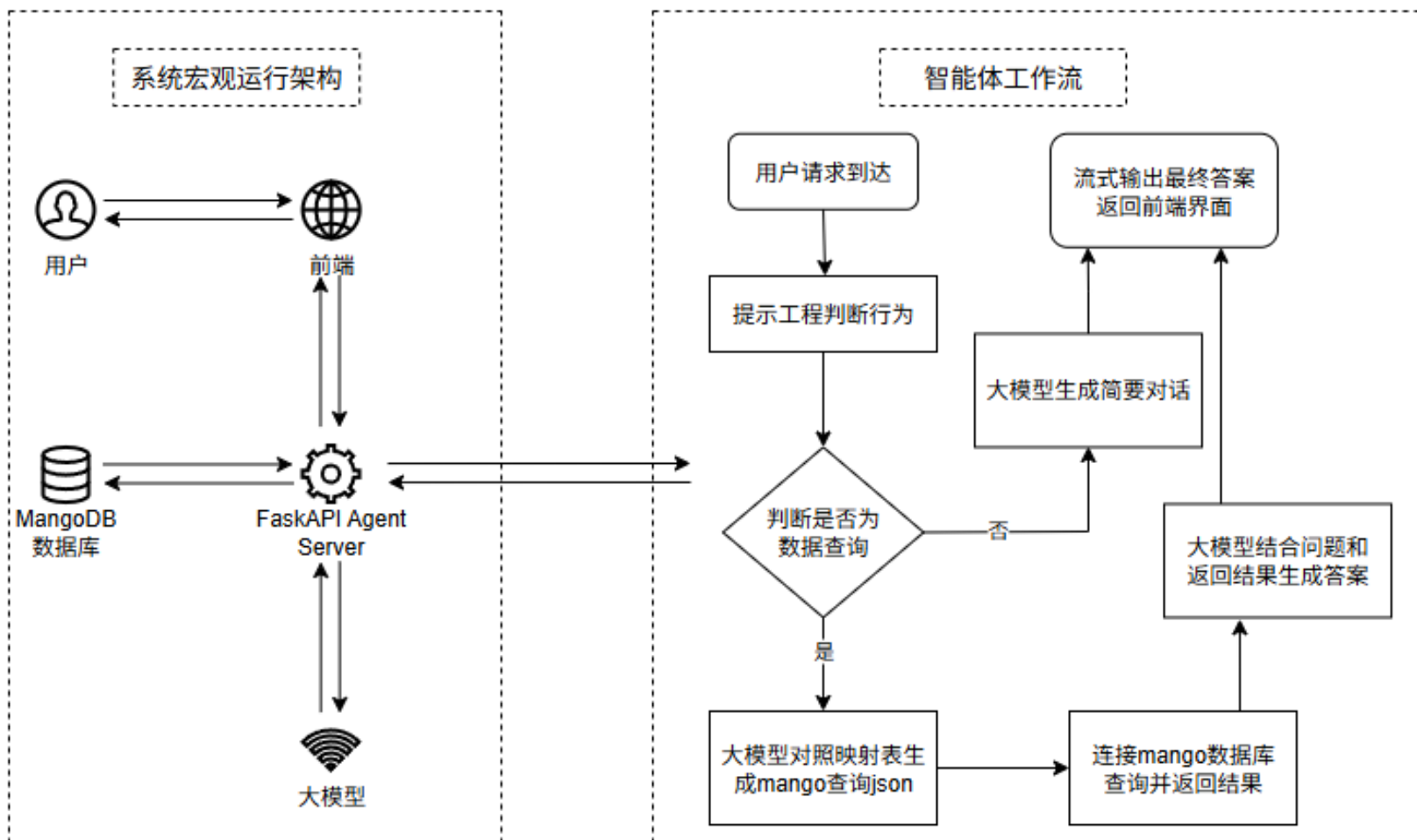
三、结果生成 (LLM2)

- 将从MongoDB数据库中获取的原始数据，再次交由LLM2进行总结提炼和格式化，以最符合用户问题解答的方式呈现最终答案。



```
FIELD_MAP = {  
    # 顶层ID  
    "sampleId": "sampleId",  
    # 实验方法与参数 (experiment)  
    "SANS波长": "sampleCharacteristics.expe",  
    "SANS波长散度": "sampleCharacteristics.",  
    "SANS_Q值范围": "sampleCharacteristics.",  
    "SANS孔径范围": "sampleCharacteristics.",  
    "USANS孔径范围": "sampleCharacteristics.",  
    "SAXS孔径范围": "sampleCharacteristics.",  
    "样品形态": "sampleCharacteristics.expe",  
    "样品厚度": "sampleCharacteristics.expe",  
    "对比流体": "sampleCharacteristics.expe",  
    "拟合模型": "sampleCharacteristics.expe",  
    "温度分布方法": "sampleCharacteristics."
```

Shale-Agent总体架构



Shale-Agent前端展示

列举出所有样品的中子散射孔隙度、石英含量、长石含量、黏土含量、碳酸盐含量、TOC和文章题目

01 shale_agent

智能体执行过程 (点击展开) ▾

根据您提供的JSON数据，以下是对所有样品的清晰、有条理的总结。所有信息均严格来源于数据源。

总体说明

- 数据共包含 1000 个样品记录。
- 大部分样品的 `sampleCharacteristics` 字段为空。
- 仅有少数样品 (约30个) 提供了完整的或部分的特征信息。

具有完整或部分数据的样品信息

以下是所有 `sampleCharacteristics` 字段不为空的样品信息。

样品ID (SAMPLEID)	中子散射孔隙度 (NEUTRON SCATTERING POROSITY)	石英含量 (%)
574243	5.48	-
852696	4.0	22.22

关键观察与总结

- 数据稀疏性:** 在1000个样品中，仅有 31个 样品提供了任何数据。绝大多数样品 (969个) 的 `sampleCharacteristics` 字段为空，因此无法提供其任何特征信息。
- 数据完整性:** 即使是这31个有数据的样品，其信息也不完整。例如：
 - 样品 552604 (Mar-1) 有矿物含量和TOC，但孔隙度为 `null`。
 - 样品 574243 (BARNETT) 有孔隙度和TOC，但无任何矿物含量。
 - 样品 774412 (New Albany Shale) 有孔隙度范围，但无其他任何数据。
- 数据格式多样性:** 提供的数据类型多样，包括：
 - 数值: 如 5.48 , 22.22 。
 - 范围: 如 "1.5-9.1" , "3.0-9.4" 。
 - 描述性文本: 如 "increased after SC-CO2" 。
 - 字符串数值: 如 "52" , 这可能是数据录入的格式问题。
- 文章题目:** 所有提供数据的样品都附带了文章题目，这为数据来源提供了明确的文献依据。

gas reservoirs using USANS/SANS, gas adsorption, and mercury intrusion
Hierarchical Pore Morphology of Cretaceous Shale: A Small-Angle Neutron Scattering and Ultrasmall-Angle Neutron Scattering Study

端到端问答性能评测

测评数据集：该数据集包含抽选自中国散裂中子源理论计算模拟平台的 30 篇关于页岩的核心文献，并基于此人工构建了涵盖 100 个问题的问答对

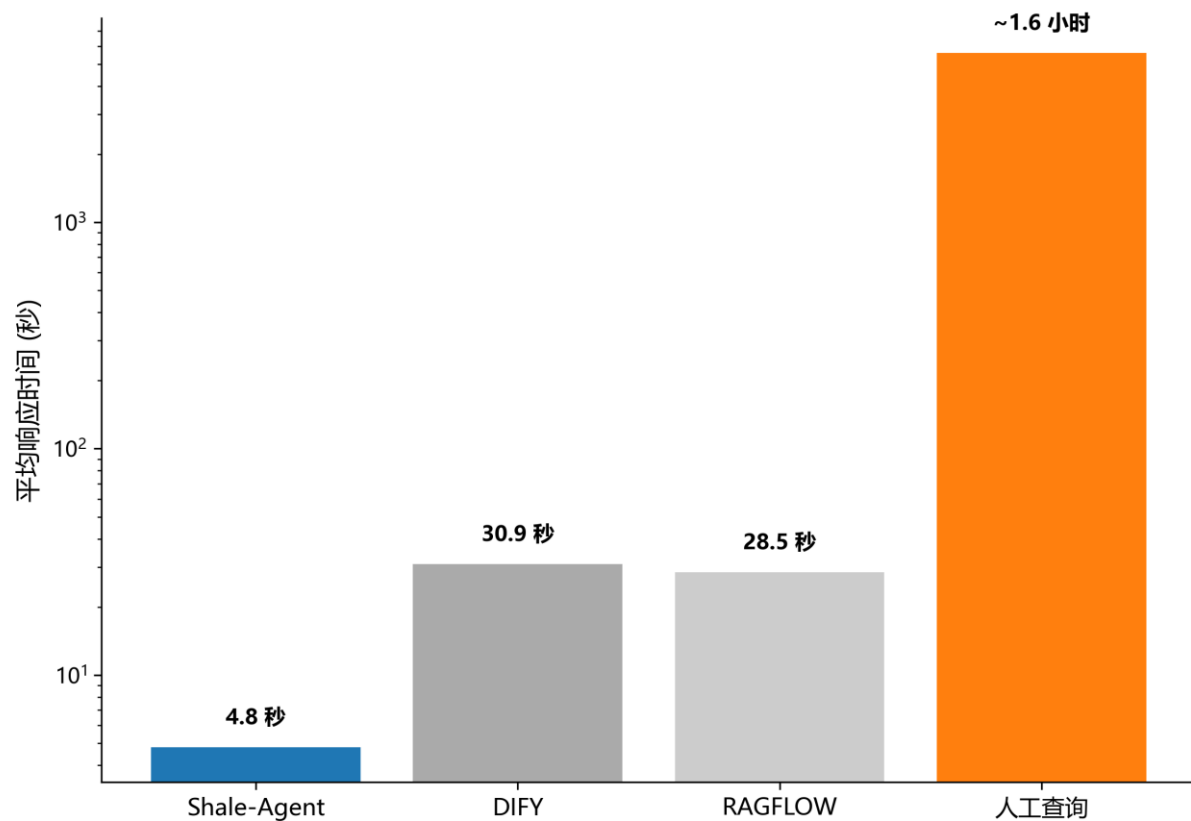
- 事实检索型：查询特定实体的精确属性。
- 条件筛选型：筛选满足单个或多个条件的实体。
- 对比分析型：对比两个或多个实体的属性。
- 聚合计算型：进行排序、计数等聚合操作。

```
{
  "question": "样品'Mar-1'的石英含量是多少？",
  "answer": "66 wt%"
},
{
  "question": "找出所有石英含量大于30%的样品名称。",
  "answer": "Mar-1, S9, S11, 1#, 3#, SH3, SH9, SH15"
},
{
  "question": "找出所有镜质体反射率（Ro）大于1.0%的样品名称。",
  "answer": "Mar-1, Mar-2, Mar-3, C1, C2, C3, C4, 296"
},
{
  "question": "对比样品'Mar-1'和'Mar-2'的TOC含量。",
  "answer": "Mar-1: 5.64 wt%, Mar-2: 9.60 wt%"
},
{
  "question": "石英含量最低的三个样品是哪些？",
  "answer": "SH15 (5.2 wt%), 1# (30.7 wt%), SH3 (26.1"
}
```

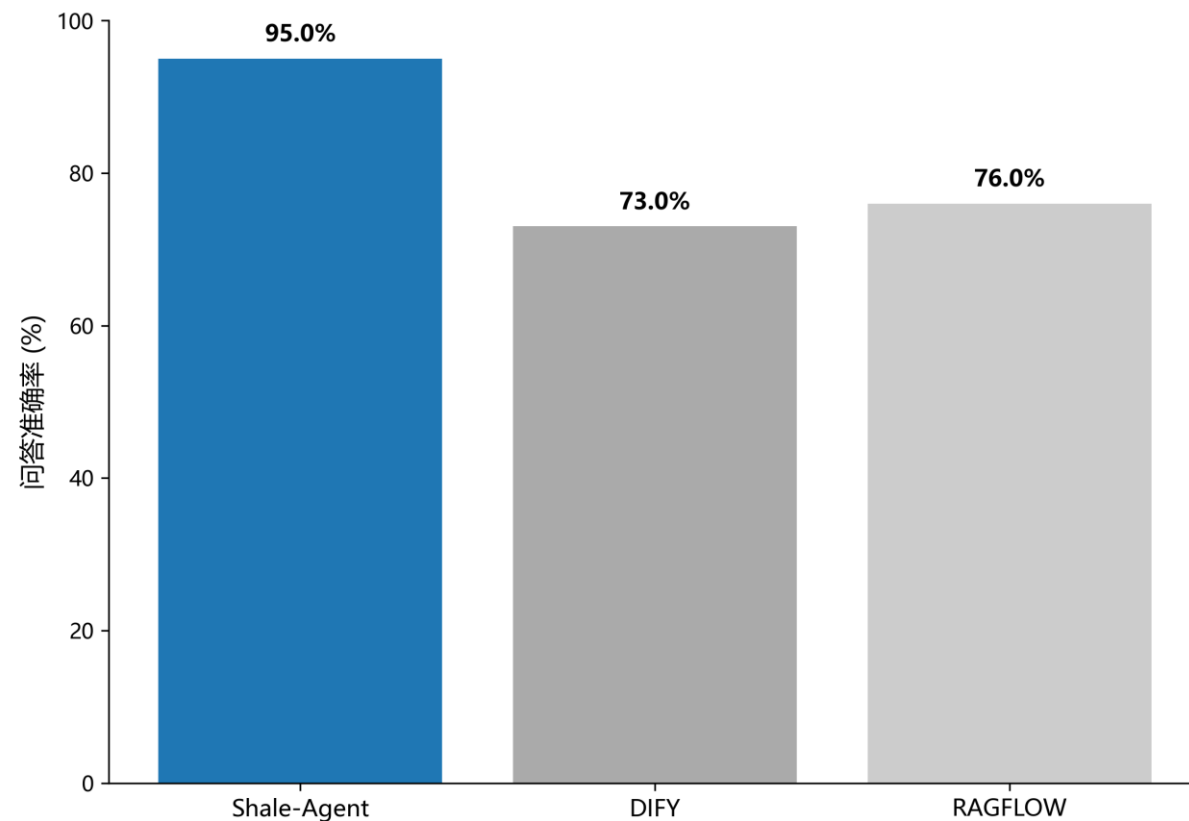
评测指标	回答准确率	平均响应时间
基线模型	Ragflow知识库	DIFY workflow

实验结果

模型与人工的平均响应时间对比 (对数尺度)



模型端到端问答准确率对比



结论：

- 提出了一种Text-to-Query智能体架构，通过知识结构化与智能查询两个阶段，为解决特定科研领域的自动化知识发现问题，提供了一个系统性的解决方案。
- 设计了一种“语义-执行”解耦的查询处理流程，利用大模型生成中间语义表示，再通过确定性的代码逻辑将其解析为可执行查询，有效弥合了自然语言的灵活性与数据库查询的严格性之间的鸿沟。
- 实验结果验证了所提架构的有效性与优越性，在端到端问答准确率上超越了现有的RAG知识库模式，证实了其作为下一代科研知识管理工具的巨大潜力。

未来展望：

- **融合多模态知识抽取：** 扩展Agent能力，使其能够解析科学文献中的图表、公式等非文本信息，并将这些多模态数据与文本知识进行对齐与融合，构建更全面的知识体系。
- **增强复杂因果与假设推理：** 推动Agent从“信息检索者”向“科研协作者”演进，使其不仅能回答事实性问题，还能基于已有数据进行因果推断、发现潜在关联，甚至生成有价值的科学假设。
- **Agent自主演进与自适应学习：** 研究Agent的自我优化机制，使其能够根据用户反馈和新发表的文献，自动学习和迭代其知识抽取模板与查询优化策略，以降低对人工维护的依赖。

愿景：将Shale-Agent从一个领域专用工具，升维为面向多学科、可定制、智能化的科学知识即服务平台。

- **平台化与可定制化：** 开放底层架构，支持用户上传任意领域的文献集，并通过低代码/无代码界面自定义知识抽取模式，实现从“页岩领域”到“全科学领域”的泛化。
- **多模态知识融合引擎：** 集成先进的计算机视觉与图表识别技术，实现对科研文献中图表、化学式、流程图等复杂多模态信息的自动解析与结构化，构建文、图、表一体的立体知识网络。
- **跨文献知识图谱与推理：** 将抽取出的孤立知识点，自动构建为大规模动态知识图谱。在此基础上，开发高级推理引擎，主动发现跨文献、跨领域的潜在科学关联，实现从“信息检索”到“知识发现”的跃迁。
- **交互式分析与主动推荐：** 打造可视化的数据探索与分析仪表盘，并基于研究者的兴趣画像，主动推送可能被忽略的关键文献、潜在的实验方向或跨学科合作机会，成为科研人员的“主动式”智能伙伴。

谢谢大家！
请各位专家批评指正！



中国科学院大学

University of Chinese Academy of Sciences



中国科学院高能物理研究所

Institute of High Energy Physics

Chinese Academy of Sciences