

从 RAG 到 KAG： 结构化思考范式下的复杂推理

王昊奋

OpenKG 轮值主席

腾讯云 TVP



CONTENTS / 目录

01

**LLMs垂域
复杂推理问题**

02

**LLMs应用
外挂知识库**

03

**KAG框架
结构化推理**

04

**KAG-Thinker
结构化思考**

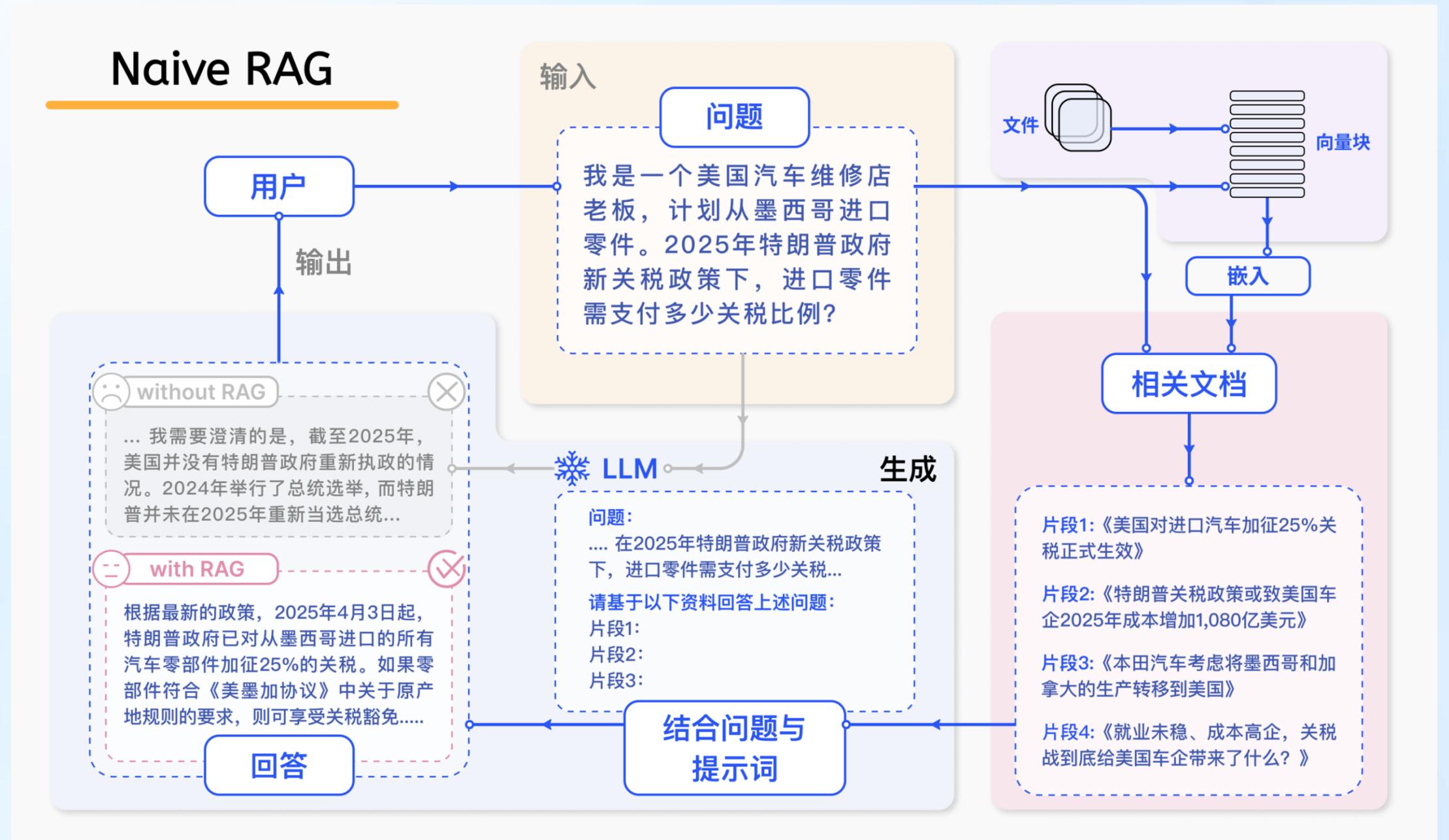
大模型受语料限制，无法跟进最新事件。
例如，在不依赖外部知识库的情况下，LLM不知道特朗普执政期间的情况。

LLM的缺陷

- 幻觉
- 信息过时
- 参数化知识效率低
- 缺乏专业领域的深度知识
- 推理能力弱

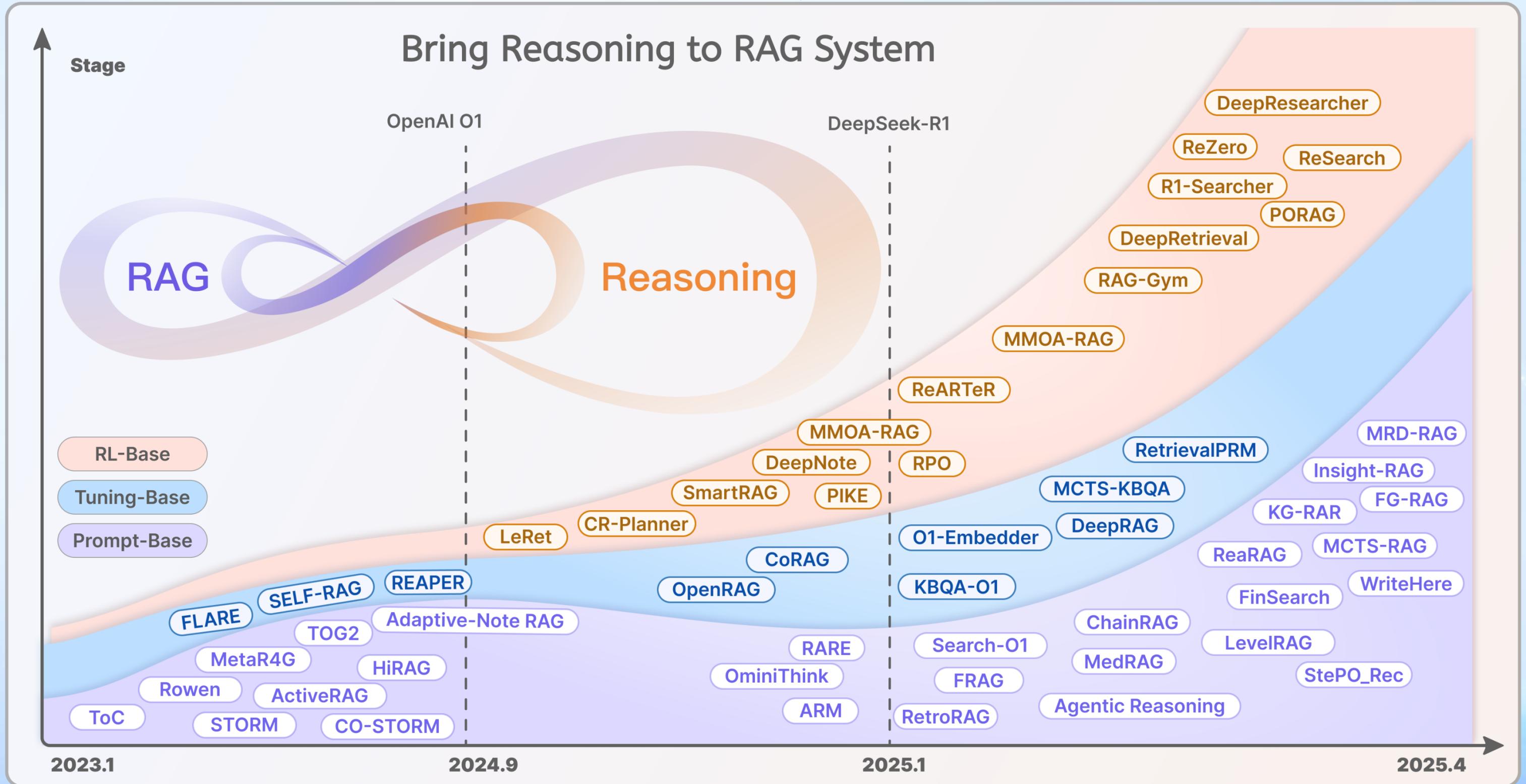
实际应用的需求

- 领域精准问答
- 数据频繁更新
- 生成内容可解释可溯源
- 成本可控
- 数据隐私保护



RAG通过语义检索为LLM提供**额外知识**。
在案例中，检索到特朗普2025年新的关税政策，新增25%的232条款，并提到美加墨贸易协定可能豁免。但答案仍有不足，如未提及基础税率，未明确豁免的具体要求，也未区分汽车零部件税率。仅靠简单RAG无法满足复杂问题需求，我们需要的是决策辅助者，而非仅是知识助手。

在OpenAI O1和DeepSeek-R1等慢思考模型兴起后，将RAG与推理能力结合的研究不断出现

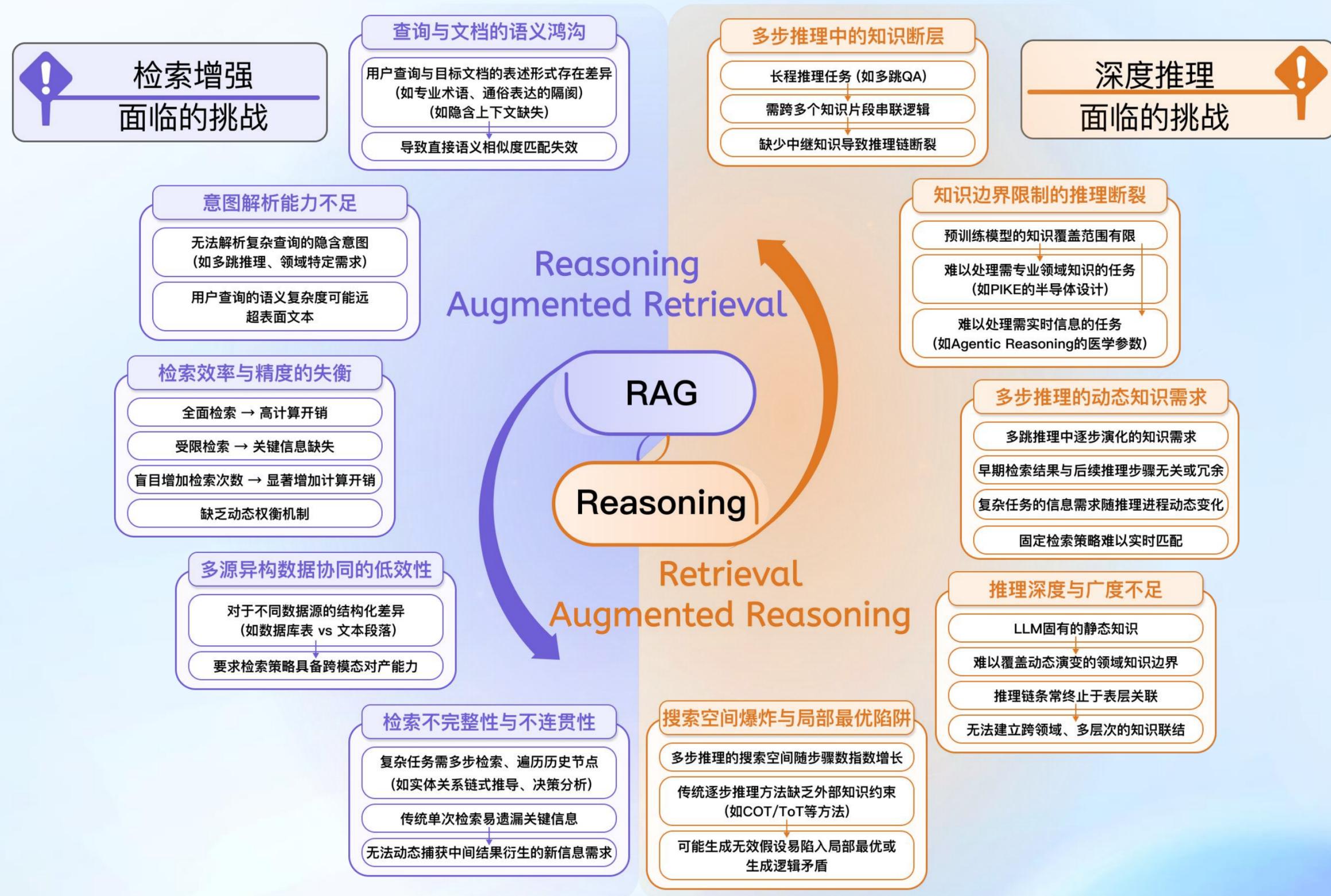


RAG的局限性

- 意图解析能力不足
- 查询与文档的语义鸿沟
- 多源异构数据协同的低效性
- 检索不完整性与不连贯性
- 检索效率与精度的失衡

Reasoning的局限性

- 多步推理中的知识断层
- 知识边界限制的推理断裂
- 搜索空间爆炸与局部最优陷阱
- 多步推理的动态知识需求
- 推理深度与广度不足



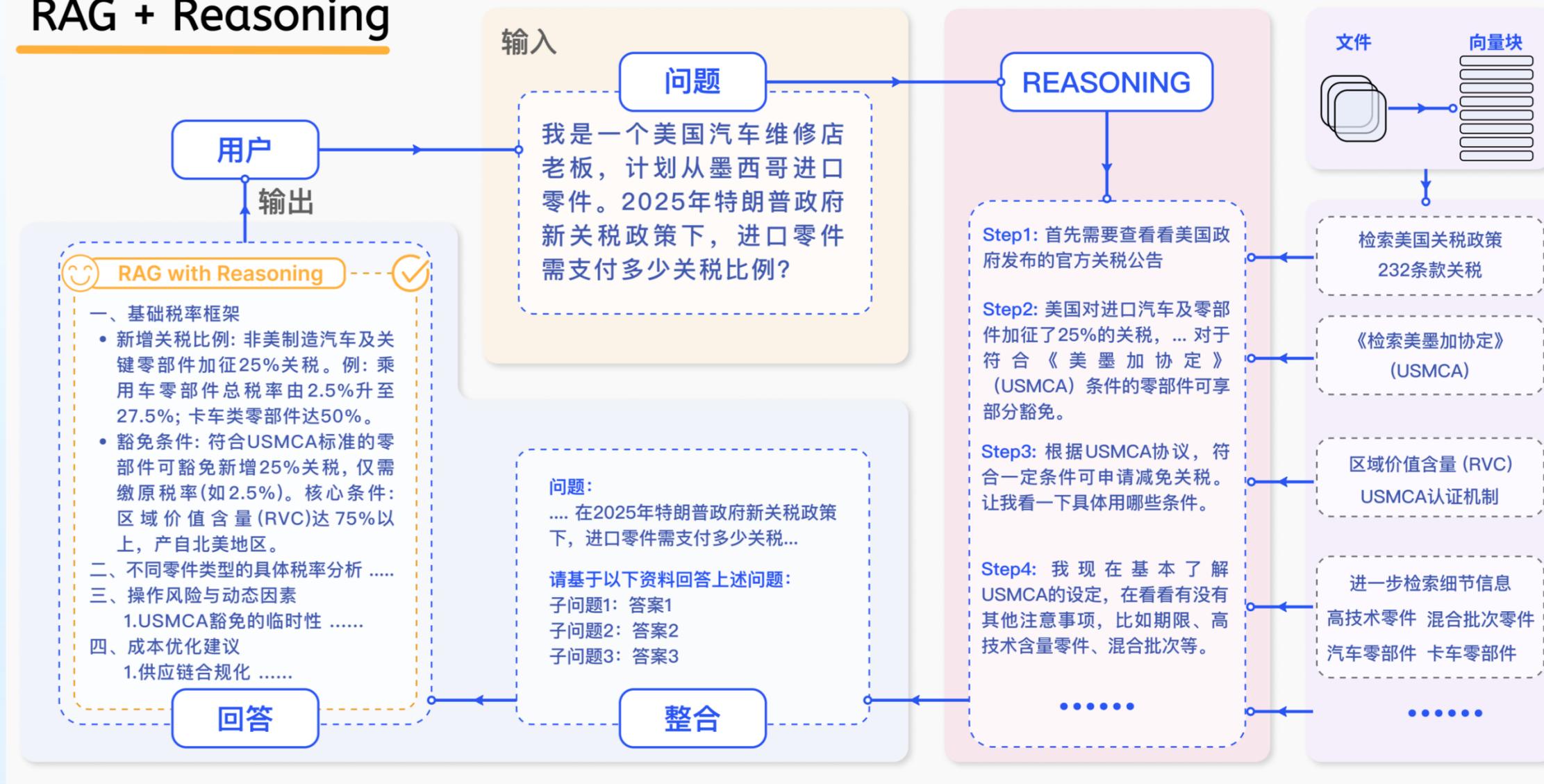
朴素RAG面临的缺陷

- 意图理解弱
- 复杂推理 [多跳] 能力差
- 信息覆盖度不足
- 决策链路不透明

实际场景提出更高的要求

- 语义匹配 → 逻辑驱动的精准确索
- 信息堆砌 → 逻辑自治的上下文构建
- 盲目的检索 → 智能资源分配
- 单一问答 → 系统性决策支持
- 被动的知识助手 → 主动的认知伙伴

RAG + Reasoning



在该案例中，通过多步推理，将复杂问题分解后分别检索。与传统RAG直接堆砌答案不同，RAG+reasoning通过任务拆分、检索、理解，最终形成逻辑自治的推理链路。最终答案更细致，包含基础税率、汽车和卡车零部件税率差异、USMCA豁免条件、风险分析和优化建议，是一个较满意的回答。

私域文档&数据

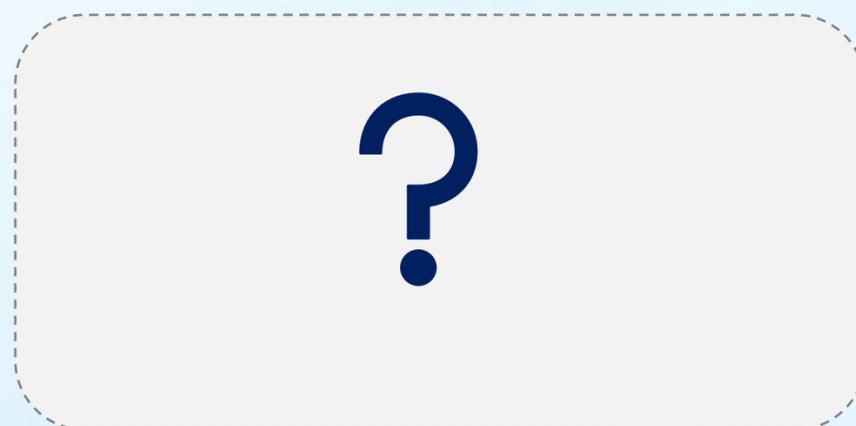


研报

政策文件

项目	2018年
现金类资产(亿元)	1.70
资产总额(亿元)	107.89
所有者权益(亿元)	71.97
短期债务(亿元)	3.60
长期债务(亿元)	25.54

建索引 + 检索 + 推理 + 生成



专业性知识服务

专业问答
法律、政务、医疗、科学

写作助手
新闻稿、研报、分析

场景

知识精准
知识完备
逻辑严谨
时间敏感
数值敏感

要求

1. 错误定性或错误逻辑

2. 事实性错误或无依据

3. 时间、数值不敏感

4. 张冠李戴

5. 不能区分重要性

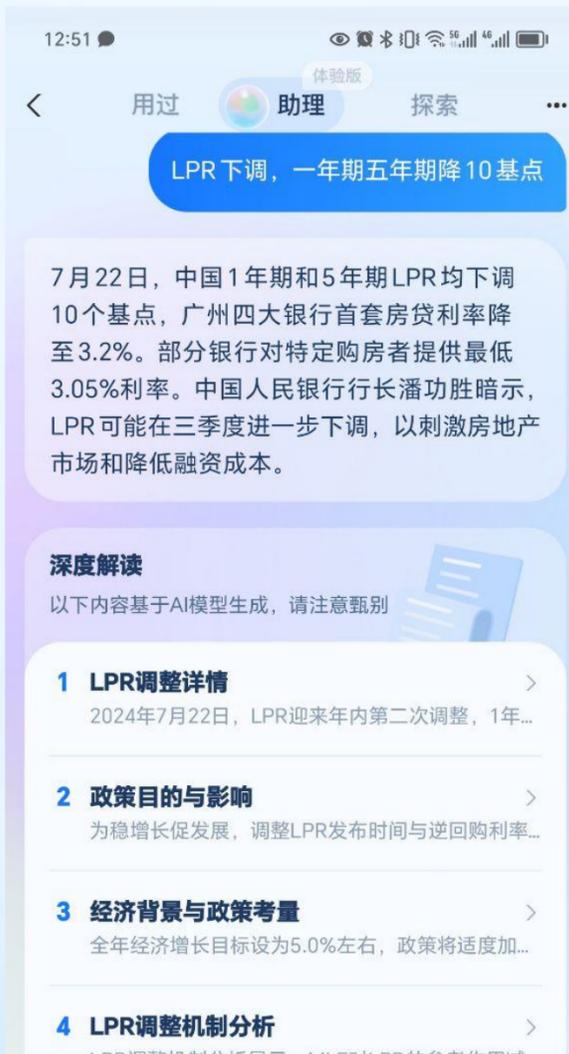
6. 语义不精准

7. 召回不完备



分析写作

热点事件解读

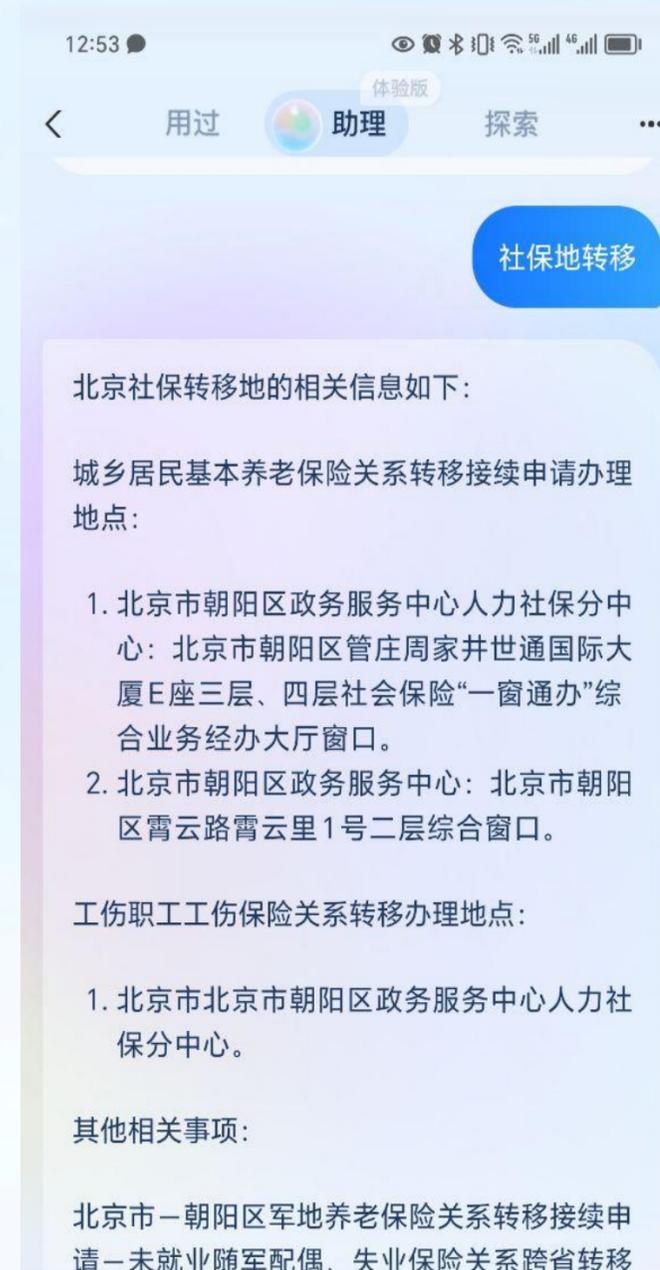


银行风险分析



知识问答

政务办事问答



医疗健康问答



保定市提取公积金需要哪些材料

购买自住住房提取住房公积金

不全，提取公积金的12种不同情况

生育险在哪里看

生育津贴支付

检索错误，参保人员参保信息查询

舟山市怎么查房产证

未找到相关信息

遗漏，不动产权属证明网上查询

社保月缴费多少

职工参保登记

检索错误，没有事项

	有果 准确率	召回率
基本RAG	0.55	0.37
知识增强RAG	0.91	0.71

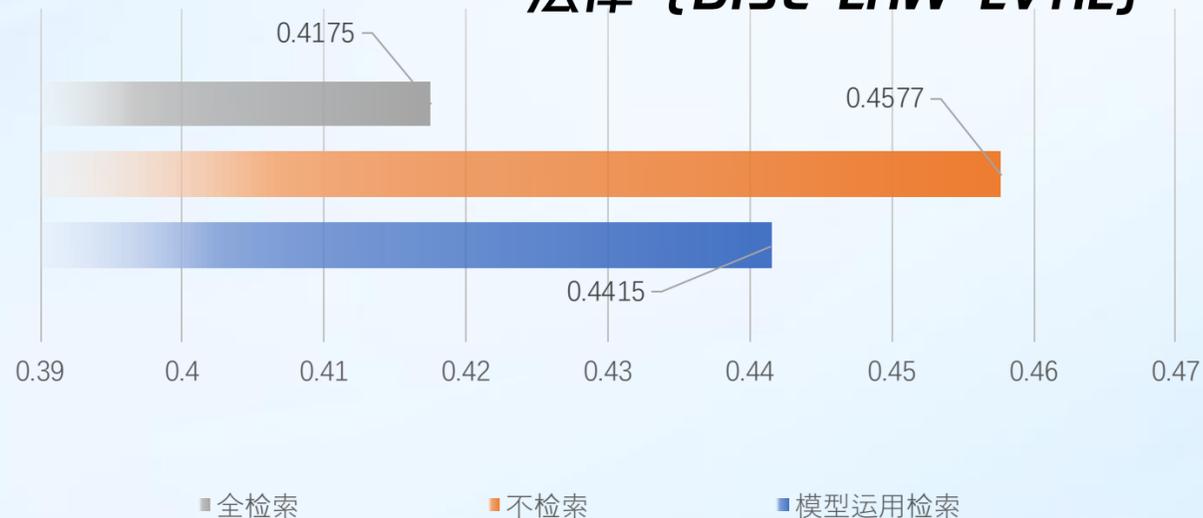
我买房申请了500万公积金贷款，计划20年还清，当前月利率是0.02，每个月应该还多少钱

已知公积金贷款月供计算公式为[贷款本金×月利率×(1+月利率)^{还款月数}]÷[(1+月利率)^{还款月数}-1]
算式：500万×0.02×(1+0.02)^{20×12}÷[(1+0.02)^{20×12}-1]

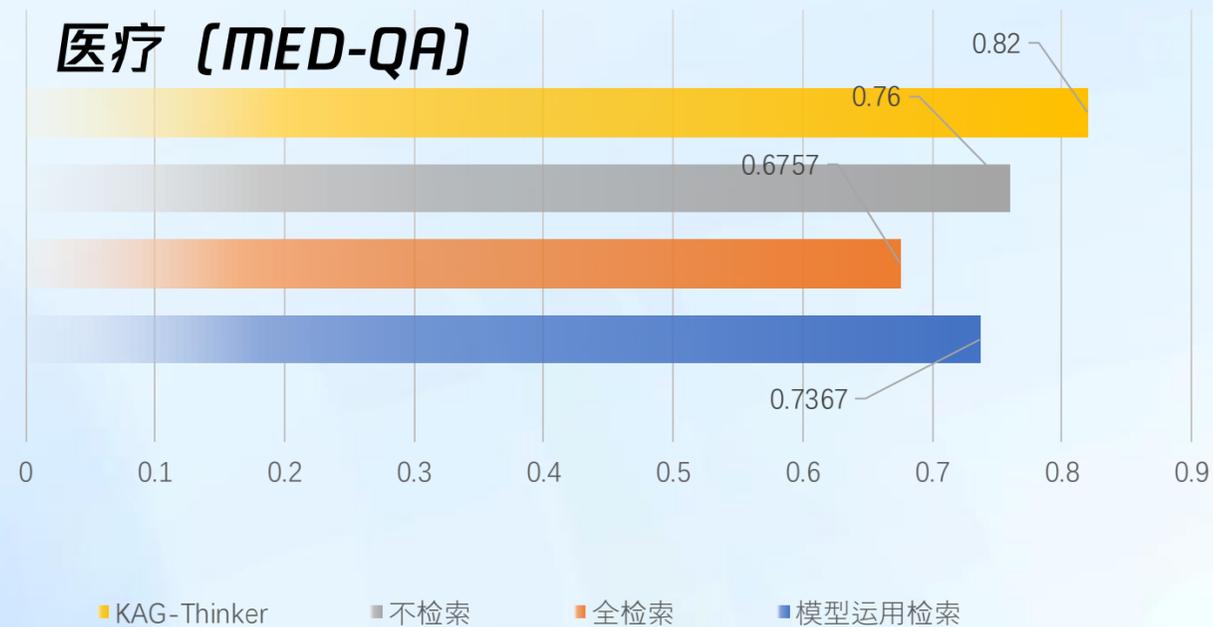
01

LLMs垂域落地的典型问题

法律 (DISC-LAW-EVAL)



医疗 (MED-QA)



男孩, 3岁。体重16kg, 双下肢被开水烫伤。查体: BP 85 / 60mmHg, 烦躁不安。双下肢 [包括臀部] II°烧伤, 尿量15ml/h。第一个24小时应补充胶体液的量为 []。 \n"options": {"A": "500ml", "B": "800ml", "C": "400ml", "D": "600ml", "E": "700ml"}

Qwen3-235B-A22B

根据烧伤补液公式, 第一个24小时的胶体液补充量为:

$\text{胶体液量} = \text{II}^\circ、\text{III}^\circ\text{烧伤面积} (\%) \times \text{体重} (\text{kg}) \times 0.5$

3岁男孩双下肢 (包括臀部) II°烧伤, 烧伤面积约为26% (根据中国九分法, 儿童双下肢面积为26%)。体重为16kg, 因此胶体液量计算如下:

$\text{胶体液量} = 26 \times 16 \times 0.5 = 208 \text{ ml}。$

但根据国内通用公式, 每1%烧伤面积每千克体重补胶体液0.5ml (儿童可能略高), 再结合临床经验调整后, 通常会适当增加补液量。在选项中, 最接近合理值的是 500ml。

Web Dev Artifacts Image Generation

1. <https://www.360kao.com/shiti/16039.h...>

男孩, 3岁, 体重16kg。双下肢被开水烫伤。查体: BP 85/60mmHg

单选题) 男孩, 3岁, 体重16kg。双下肢被开水烫伤。查体: BP 85/60mmHg, 烦躁不安, 双下肢 (包括臀部) II°烧伤。尿量15ml/h。第一...

2. 百度

男孩, 3岁。体重16kg, 双下肢被开水烫伤。查体: BP 85 - 百度

男孩, 3岁。体重16kg, 双下肢被开水烫伤。查体: BP 85 / 60mmHg, 烦躁不安。双下肢 (包括臀部) II°烧伤, 尿量15ml/h。第一个24小时...

3. <https://www.doctor-...>

烧伤补液 (Brooke公式)

伤后第一个24h内, 成人每1%II度、III度烧伤面积补给等渗乳酸钠林格溶液1.5ml, 胶体0.5ml, 同时补给基础水分2000ml。此公式按实际烧...

LLM 在使用KB时存在: 检索问题不合适 (过于宽泛、涉及案例细节)、抑制自身知识、受检索噪音干扰 等问题, 业界也有较多优化方案

优化路线	主要特性	业界代表方法	KAG方法
框架优化	知识索引及框架	IRCOT [24]、HippoRAG [24]、LightRag [24]、GraphRAG [24]	KAG-V1 [24]
模型优化	建模LLM内在认知	ReSearch [25]、Search-R1 [25]、ZeroSerach [25]、StepSearch [25]	KAG-Thinker [25]

Query: 介绍下大模型Ling-Plus、Ling-Lite的关键进展和核心优势? 为求解这个问题, 请帮完成思维链拆分, 只输出每个步骤就可以, 不需要输出任何冗余信息。

Ling-latest-MoE

1. 确定大模型Ling-Plus和Ling-Lite的基本信息。
2. 收集Ling-Plus和Ling-Lite的关键进展。
3. 分析Ling-Plus和Ling-Lite的核心优势。
4. 整合信息, 形成完整的回答。



5.5 s

第一次

1. 确定大模型Ling-Plus和Ling-Lite的最新动态和更新。
2. 识别Ling-Plus和Ling-Lite在技术实现上的创新点。
3. 分析这两个模型在实际应用中的表现和用户反馈。
4. 概括Ling-Plus和Ling-Lite的核心优势, 包括性能、效率、适用性等方面。
5. 整合信息, 形成对Ling-Plus和Ling-Lite关键进展和核心优势的介绍。



6.6 s

第二次

Qwen3-235B-A22B

1. 确定Ling-Plus和Ling-Lite是研发的大模型产品。
2. 收集Ling-Plus的关键技术进展, 包括但不限于参数规模、训练数据、应用场景等方面。
3. 收集Ling-Lite的关键技术进展, 聚焦其轻量化、高效推理、部署能力等方向。
4. 分析Ling-Plus的核心优势, 如多模态处理能力、语言理解深度、生成质量等。
5. 分析Ling-Lite的核心优势, 例如低资源消耗、快速响应、适配移动端或边缘设备等。
6. 对比两者在定位、性能、适用场景等方面的差异与互补性。
7. 总结Ling-Plus和Ling-Lite在AI战略中的角色和意义。

第一次

1. 确定Ling-Plus和Ling-Lite的基本定位和应用场景
2. 收集Ling-Plus和Ling-Lite的技术参数和模型架构信息
3. 分析Ling-Plus和Ling-Lite在性能指标上的关键进展
4. 比较Ling-Plus与Ling-Lite之间的差异与适用场景
5. 总结Ling-Plus和Ling-Lite的核心技术优势
6. 提取其在实际应用中的具体成果或案例
7. 综合信息, 整理出结构化的答案框架

第二次

两次思考过程差别较大、逻辑不一致

介绍下大模型Ling-Plus、Ling-Lite的关键进展和核心优势?

两次思考过程相对稳定, 依然存在关键信息的缺失

• Step1: 大模型Ling-Plus的关键进展是什么?

∨ logical-form-chain

Action1: Retrieval(s=s1:model[`Ling-Plus`], p=p1:keyAdvancement, o=o1:advancement)

• Step2: 大模型Ling-Lite的核心优势是什么?

∨ logical-form-chain

Action2: Retrieval(s=s2:model[`Ling-Lite`], p=p2:coreAdvantage, o=o2:advantage)

• Step3: 输出 #1 和 #2

∨ logical-form-chain

Action3: Output(o1, o2)

• Step1: 大模型Ling-Plus的关键进展和核心优势是什么?

∨ logical-form-chain

Action1: Retrieval(s=s1:model[`Ling-Plus`], p=p1:keyAdvancementsAndCoreAdvantages, o=o1:advantages)

• Step2: 大模型Ling-Lite的关键进展和核心优势是什么?

∨ logical-form-chain

Action2: Retrieval(s=s2:model[`Ling-Lite`], p=p2:keyAdvancementsAndCoreAdvantages, o=o2:advantages)

• Step3: 输出#1和#2

∨ logical-form-chain

Action3: Output(o1, o2)

资产负债信息中流动资产最高的子项是哪个？同时也计算占总流动资产的比例



正确答案 38.65%

根据中芯国际2024年第三季度报告中的合并资产负债表，2024年9月30日的流动资产中，货币资金为29,878,544千元，是流动资产中最高的一项。流动资产合计为77,297,667千元。

货币资金占总流动资产的比例计算如下：

$$\text{比例} = \left(\frac{29,878,544}{77,297,667} \right) \times 100\%$$

比例 $\approx 38.64\%$

因此，流动资产最高的子项是货币资金，占总流动资产的比例约为38.64%。

复制 再试一次 分享



货币资金占总流动资产的比例计算公式如下：

$$\text{货币资金占比} = \left(\frac{\text{货币资金}}{\text{流动资产合计}} \right) \times 100\%$$

将给定的数值代入公式中：

$$\text{货币资金占比} = \left(\frac{29,878,544}{77,297,667} \right) \times 100\% \approx 38.68\%$$

根据中芯国际集成电路制造有限公司2024年第三季度报告中的合并资产负债表，流动资产最高的子项是“一年内到期的非流动资产”，其数值为17,834,516千元。

总流动资产的数值为77,297,667千元。

为了计算“一年内到期的非流动资产”占总流动资产的比例，我们可以使用以下公式：

$$\text{比例} = \left(\frac{\text{一年内到期的非流动资产}}{\text{总流动资产}} \right) \times 100\%$$

将具体数值代入公式：

$$\text{比例} = \left(\frac{17,834,516}{77,297,667} \right) \times 100\%$$

比例 $\approx 23.07\%$

因此，“一年内到期的非流动资产”占总流动资产的比例大约是23.07%。



根据中芯国际集成电路制造有限公司2024年第三季度报告中的合并资产负债表，流动资产最高的子项实际上是“货币资金”，金额为29,878,544千元。

总流动资产的数值为77,297,667千元。

为了计算“货币资金”占总流动资产的比例，我们可以使用以下公式：

$$\text{比例} = \left(\frac{\text{货币资金}}{\text{总流动资产}} \right) \times 100\%$$

将具体数值代入公式：

$$\text{比例} = \left(\frac{29,878,544}{77,297,667} \right) \times 100\%$$

比例 $\approx 38.67\%$

因此，“货币资金”占总流动资产的比例大约是38.67%。

Bob is cold. Bob is not nice. Bob is not quiet. Charlie is blue. Charlie is round. Gary is rough. Harry is cold. Rough things are blue. Red, quiet things are blue. All rough, blue things are nice. If something is quiet and round then it is not nice. If Harry is red then Harry is rough. If something is nice then it is round. If something is round then it is red. Red things are not quiet.

Based on the above information, is the following statement true, false, or unknown? Gary is red.

False. Gary is rough, and according to the given information, rough things are blue, not red. Therefore, Gary cannot be red.



6.6 s

Gary is rough

Gray is **rough**

rough things are blue

rough things are blue

blue things are nice

blue things are nice

If something is nice then it is round

nice things are round

If something is round then it is red

round things are red

证据链

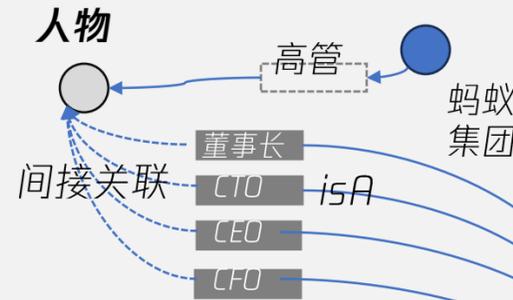
逻辑链

信息检索：以信息加工、信息获取为主

认知推理：基于获取的信息，使用推理框架完成特定任务

问题 大模型Ling-Plus主要对标的国内大模型是哪个公司发布的?

问题 蚂蚁集团的高管有哪些?



多跳关联 当阿里一开始砸钱搞AI全家桶时，DeepSeek却用开源代码炸翻了整个行业，先后发布DeepSeek-V2.5, DeepSeek-V3, DeepSeek-R1等模型！这两家中国AI巨头的技术路线，...

概念语义隐性关联 蚂蚁集团总裁韩歆毅于今年3月1日起接任CEO，负责全面业务及日常管理，向蚂蚁集团董事长井贤栋和董事会汇报。

2025年2月5日蚂蚁集团任命刘政担任新CFO，全面负责集团财务工作。2月5日，界面新闻获悉，蚂蚁集团候任CEO、总裁韩歆毅...

回答 大模型Ling-Plus主要对标的国内大模型包括阿里通义Qwen2.5-72B-Instruct和DeepSeek-V2.5 ...。

回答 蚂蚁集团高管包括董事长井贤栋、CEO韩歆毅、CFO刘政、CTO何征宇等

事件 小前提 张三去申请xx银行贷款，查询得到其信用评分580分

规则 大前提 X银行规定：信用评分<600的申请人不得发放贷款

结论 张三不符合xx银行规定，不得发放贷款

事件 小前提 某甲于2024年5月，在某市一家超市内趁人不备，将一部价值人民币3000元的手机藏入背包并带出店外。经监控确认，甲并未付款，且无任何正当理由占有该手机。

法规 大前提 根据《中华人民共和国刑法》第二百六十四条规定：“盗窃公私财物，数额较大或者多次盗窃、入户盗窃、携带凶器盗窃、扒窃的，处三年以下有期徒刑、拘役或者管制，并处或者单处罚金。”

结论 依据上述法律规定，甲的行为构成盗窃罪，且数额达到“较大”标准【通常各地标准略有不同，但3000元一般属于“数额较大”】，应判处三年以下有期徒刑、拘役或管制，并处罚金。

前提 2025年4月7日，A股三大股指大跌，上证指数下跌7.34%，失守3100点关口；深证成指下跌9.66%，创业板指跌幅更高，达12.5%，全市场近3000只个股跌停。随后在5月20日和21日，市场出现连续两日上涨，沪指分别上涨0.38%和0.21%。

结论 A股市场在经历暴跌后出现了初步企稳迹象

前提 你计划投资一笔资金 $P=10,000$ 元，年利率为 $r=5\%$ (即 0.05)，按复利计息。如果每年复利一次，请问经过 $t=10$ 年后，这笔投资的终值是多少?

依赖知识点：复利计算

```
import sympy as sp
# 定义符号变量
P, r, t = sp.symbols('P r t')
# 复利终值公式
FV = P * (1 + r)**t
# 代入具体数值
P_value = 10000 # 初始本金
r_value = 0.05 # 年利率
t_value = 10 # 投资年限
# 计算终值
final_FV = FV.subs({P: P_value, r: r_value, t: t_value})
# 输出结果，保留两位小数
print(f"复利终值为: {final_FV.evalf(2)} 元")
```

Level-1: 显性事实
How to locate explicit fact?

Level-2: 隐性事实
How to connect series of implicit facts?

Level-3: 规则推理
[演绎推理]
How to follow external rationale?

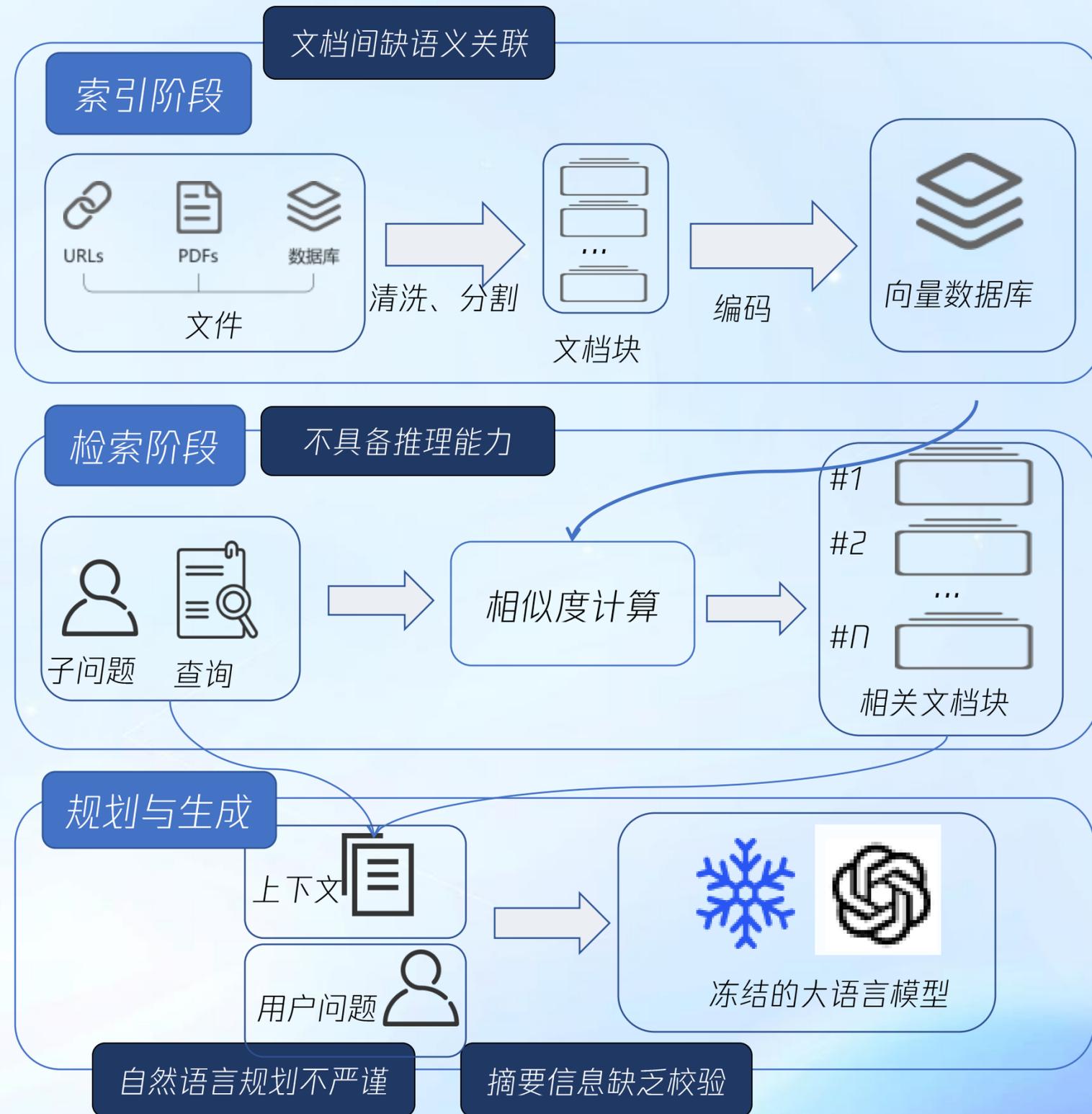
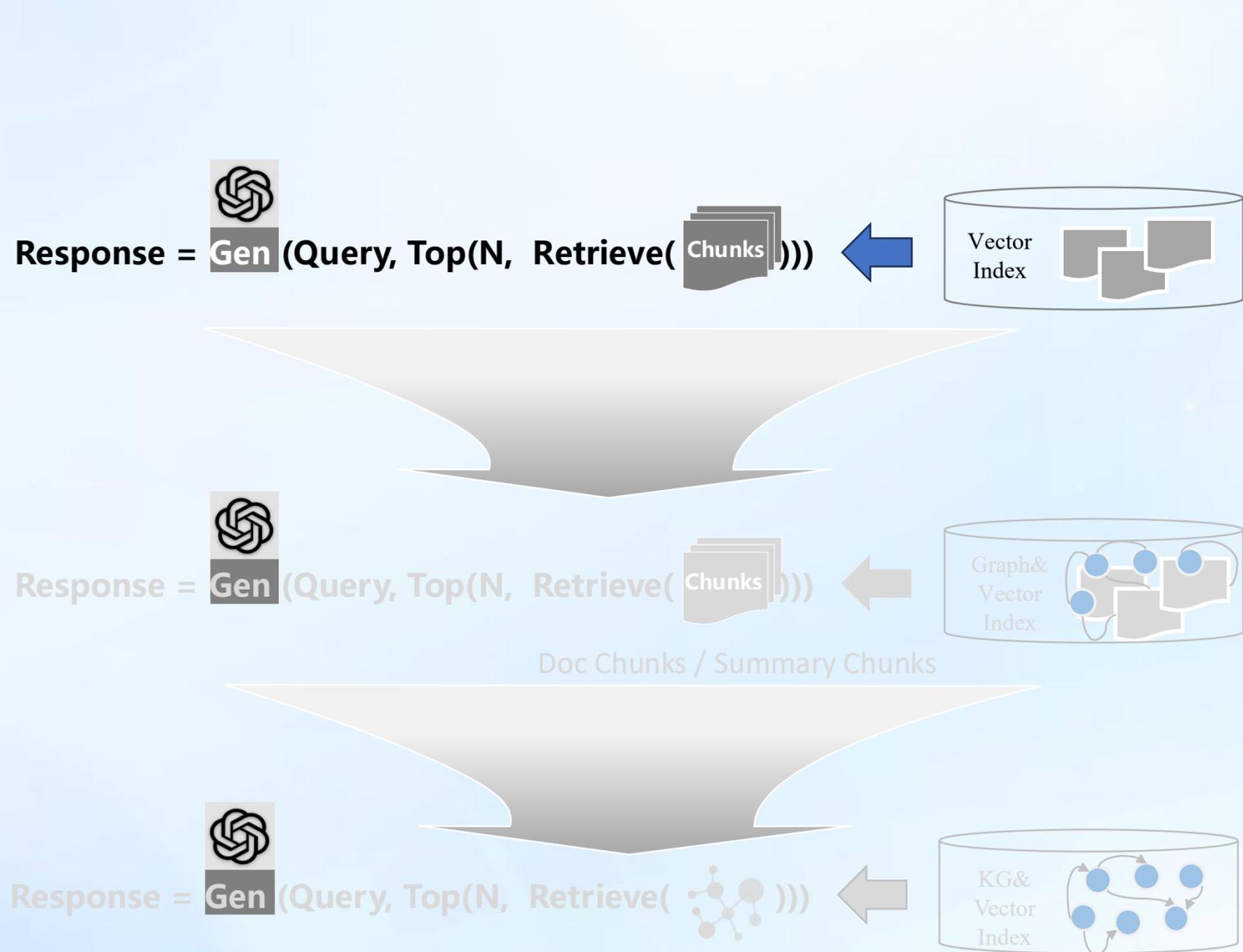
Level-4: 结果推断
[类比、归纳、溯因推理]
How to discover hidden rationale?

02

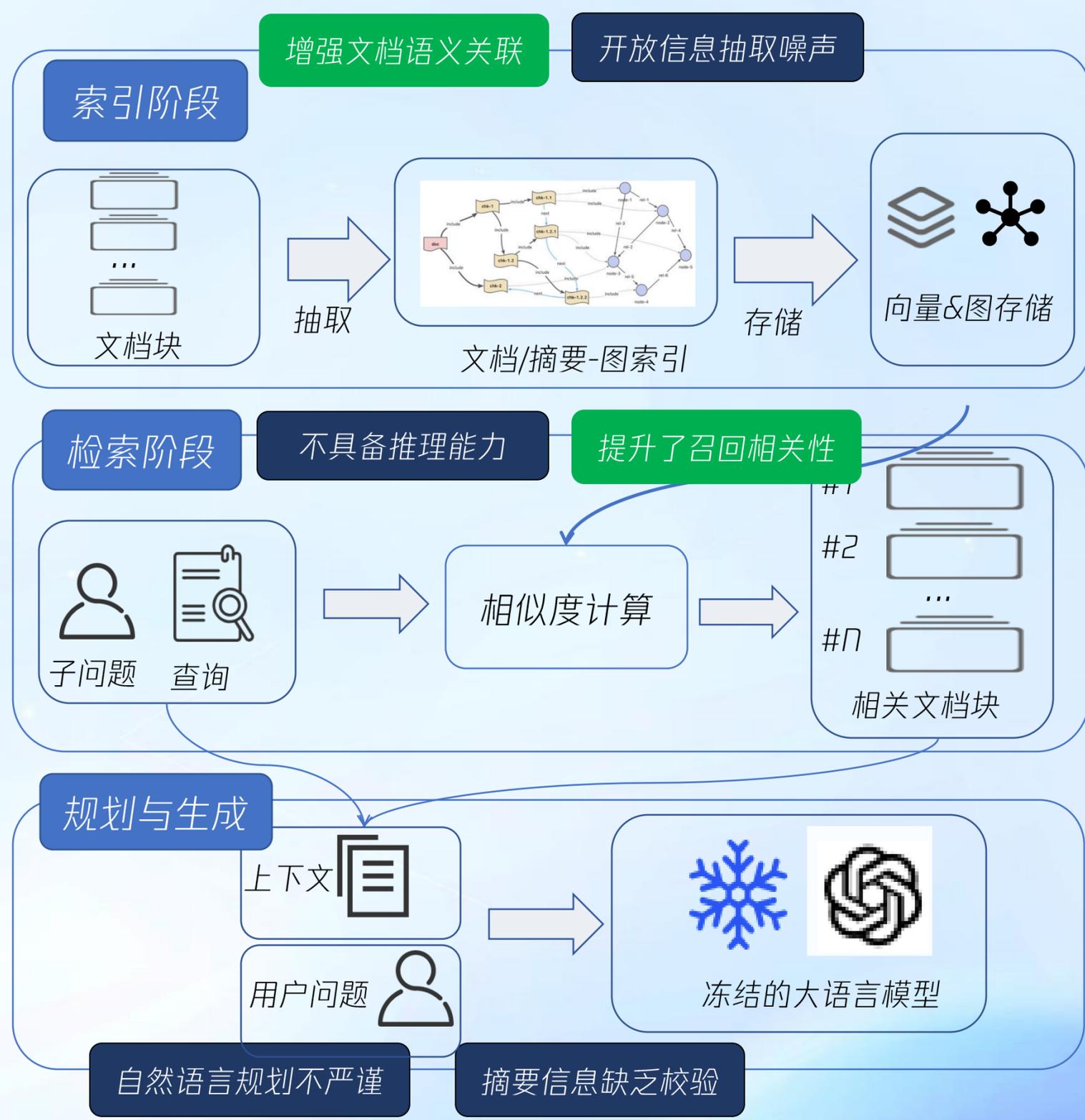
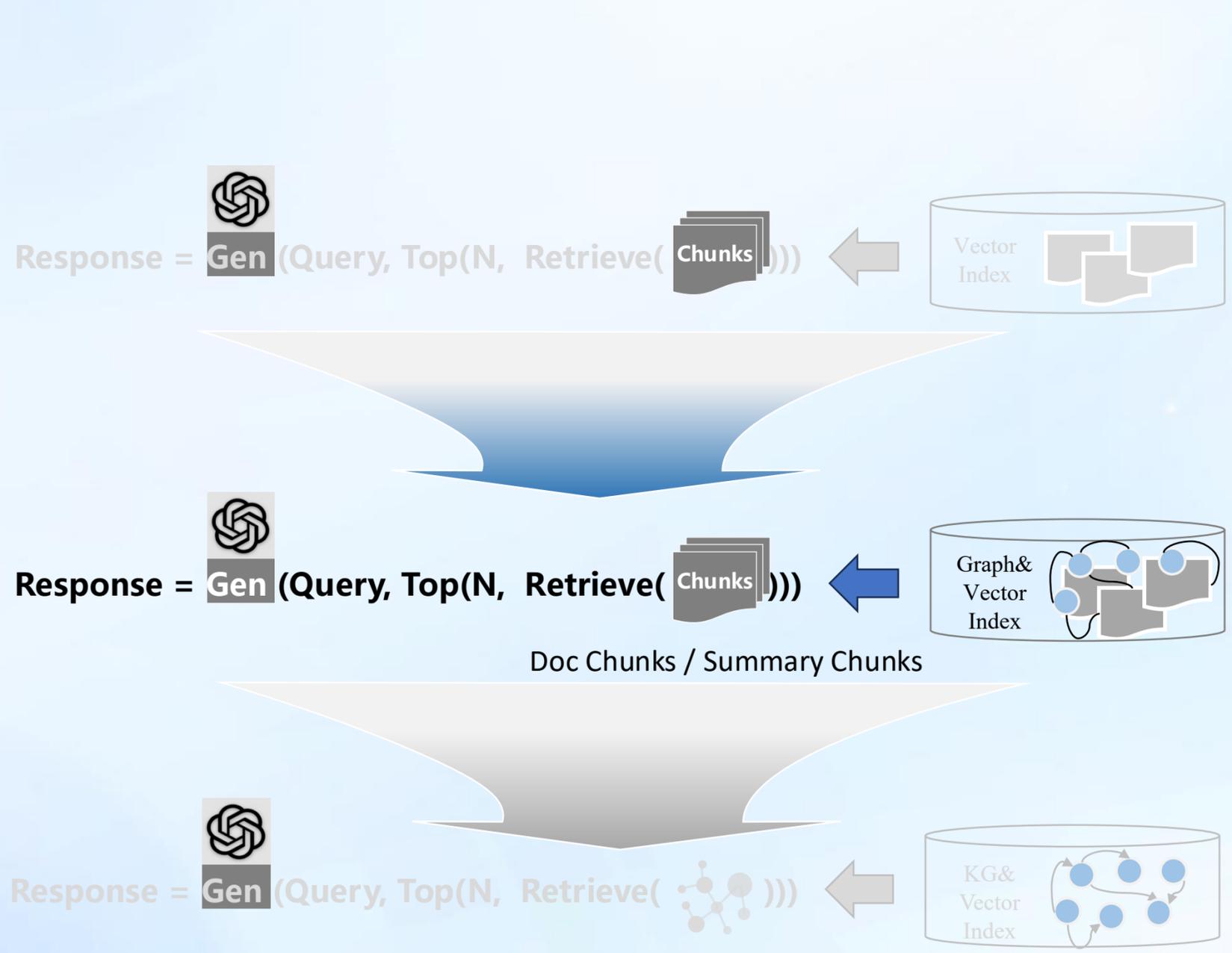
大模型外挂知识库路线

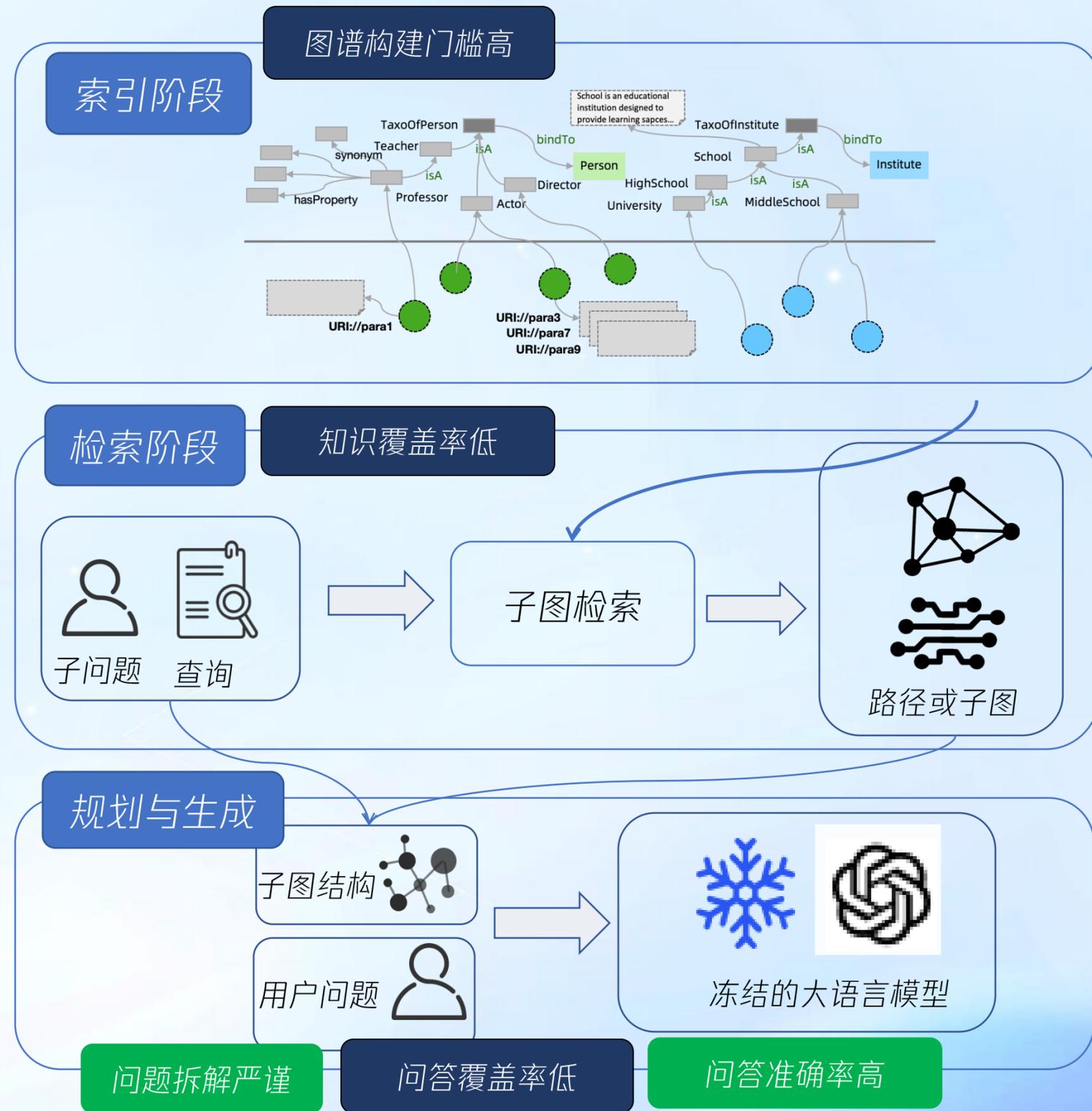
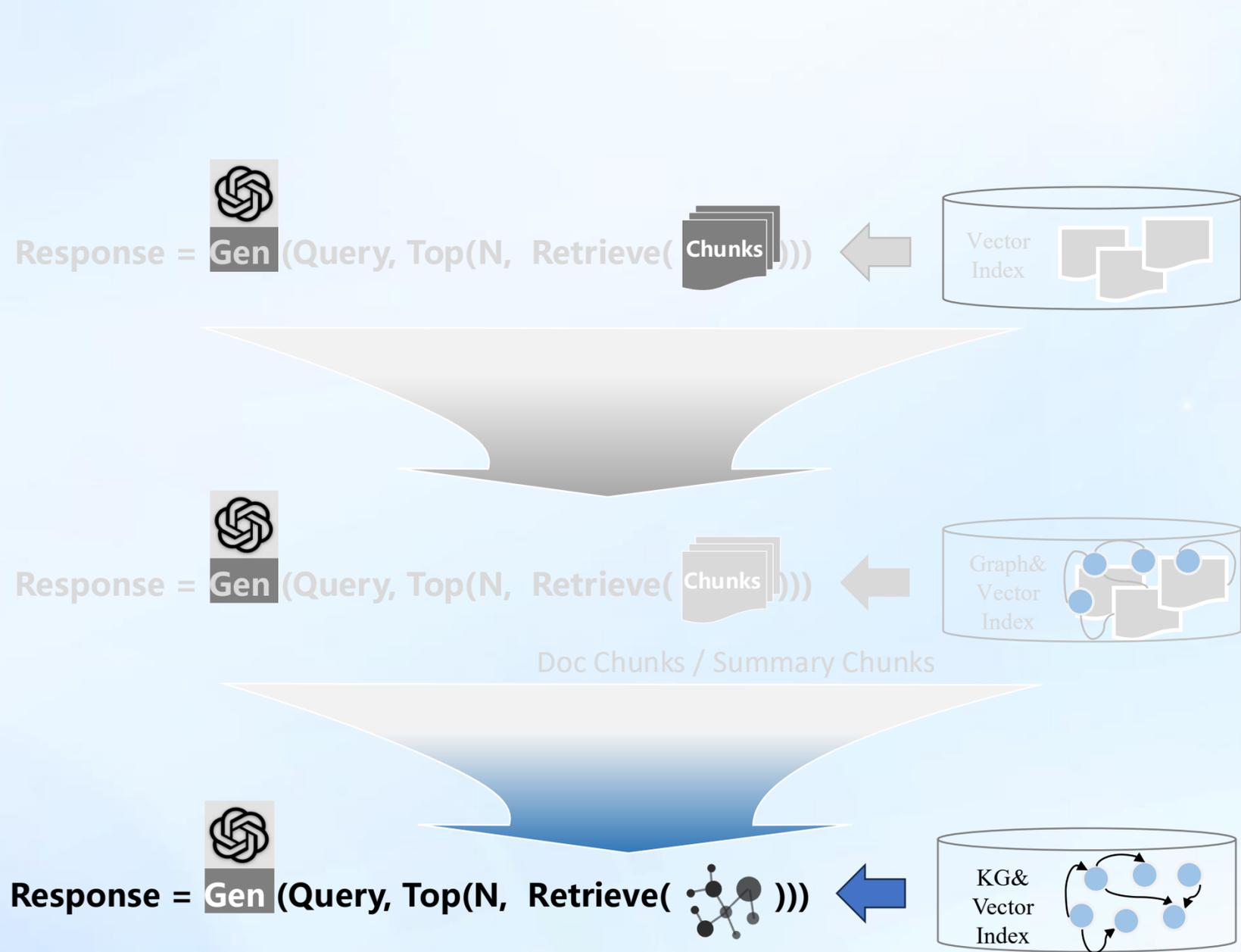
框架化集成外部知识库与大模型，使用专用检索器对接外部知识库，使用通用大模型完成问题规划和答案生成

AI 外挂知识库技术路线之RAG - 搜索引擎技术大模型外延



AI 外挂知识库技术路线之GraphRAG - 增加文档间语义关联





	<i>Indexing</i>	<i>Planning</i>	<i>Retrieve</i>	<i>Generation</i>	优点	缺点
RAG	Chunk 向量化	自然语言CoT	向量检索	QFS	构建成本低 效率高 信息完备	精准性不足 拆解不严谨 逻辑性差
GraphRAG	Chunk 向量化 图	自然语言CoT	向量检索 PPR	QFS	使用门槛低 文档语义关联 信息完备	精准性不足 拆解不严谨 逻辑性差
KBQA	实体抽取 SPO抽取	Logical-form	结构化查询语句 实体链指 语义推理	引用原事实	结构严谨 逻辑清晰 置信度高 时间和数值敏感	构建成本高 信息有损 知识缺失 可读性差

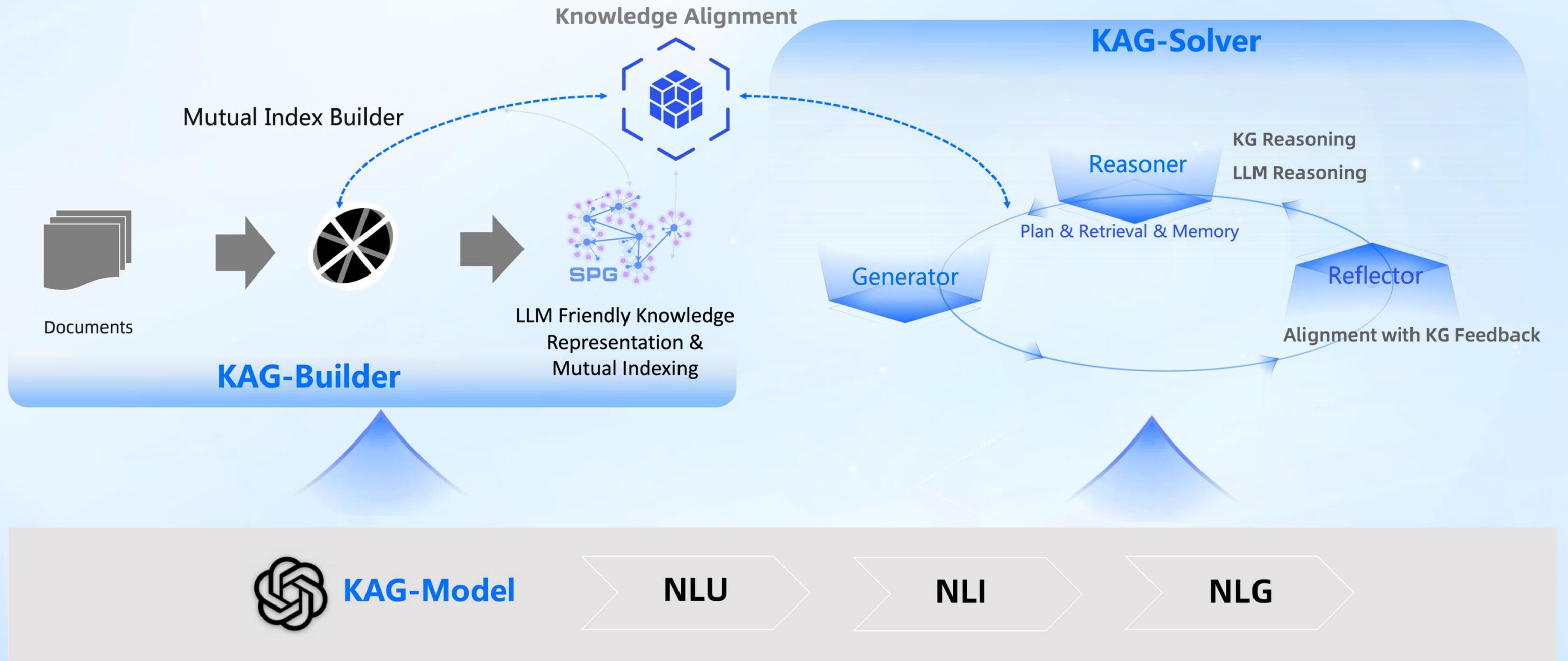
03

知识增强生成KAG路线

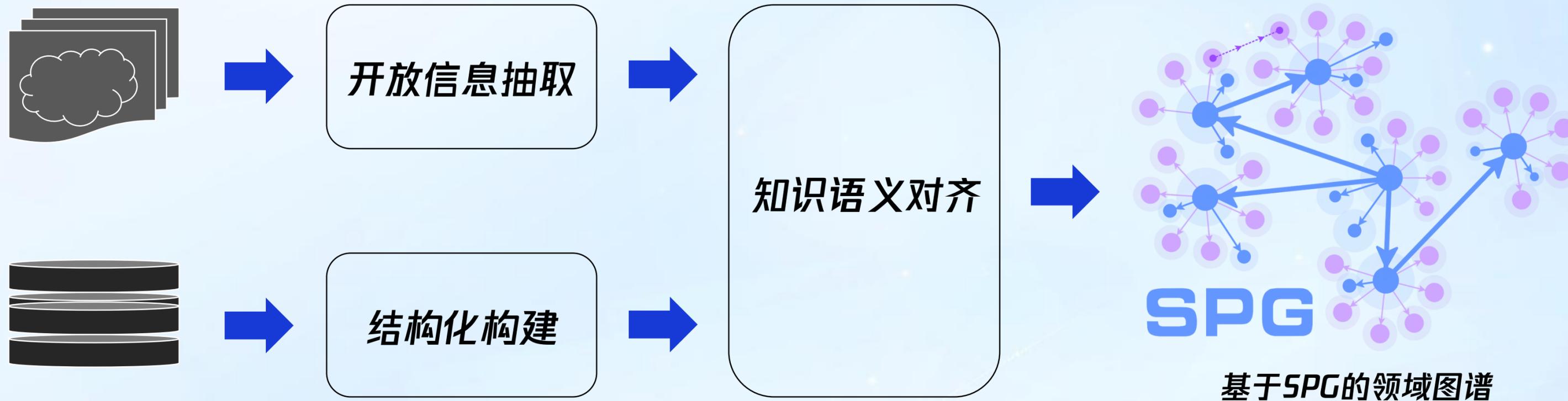
充分利用KG结构严谨的优势：通过KG-Text互索引来组织知识索引，通过KG语义引导构建结构化思考和推理

AI 专业领域对推理问答的逻辑严谨、事实正确等有更高的要求





关键能力 ① 严谨结构化知识与非结构化信息共存于同一实体空间，确保信息无损

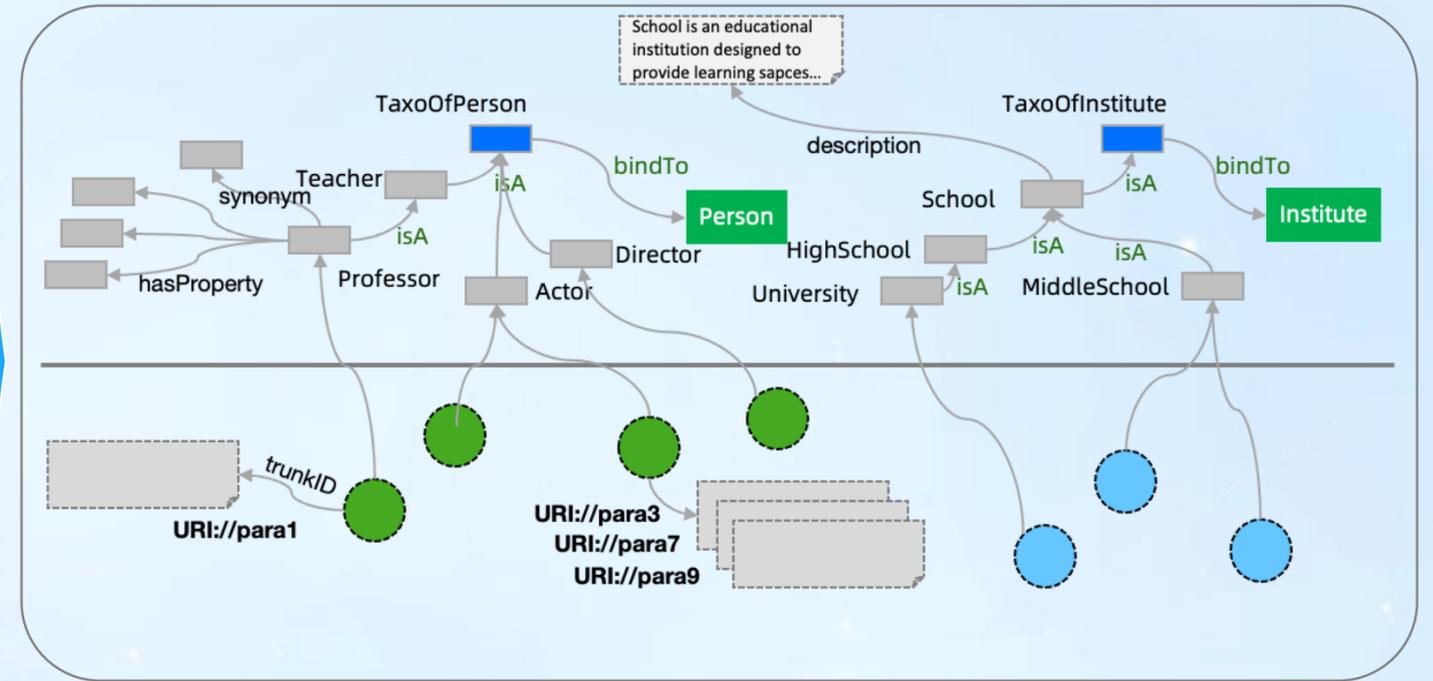
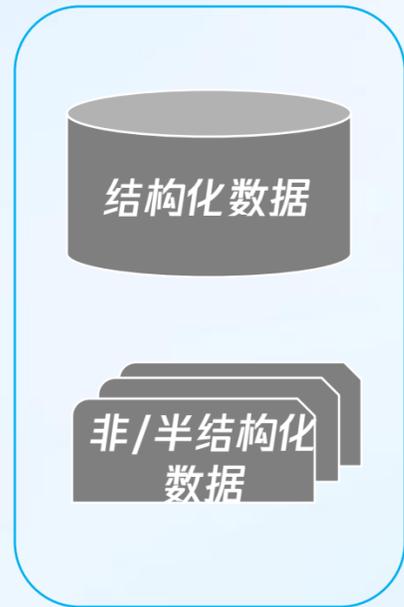


大幅降低知识图谱的构建门槛

基于语义对齐平衡信息抽取(低门槛)与专业构建

通过语义对齐缓解开放信息抽取引入的噪声问题

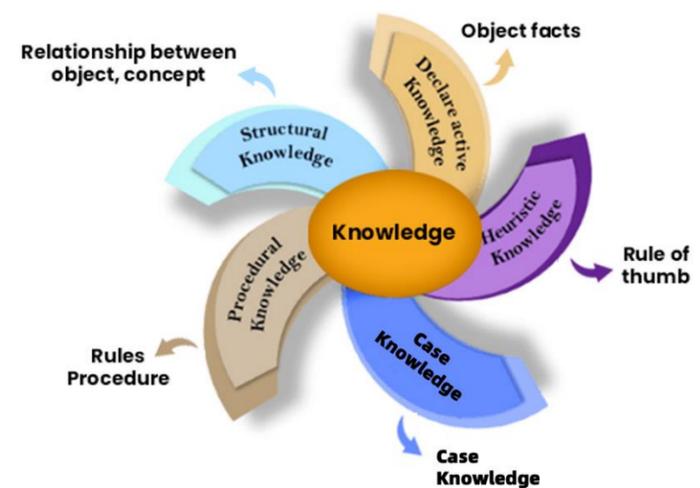
关键能力 1 严谨结构化知识与非结构化信息共存于同一实体空间，确保信息无损



[半]自动化知识建模、LLM增强的知识抽取、以知识点为中心的构建

术语对齐增强领域一致性、概念挂载增强知识联通性、实体消歧降低知识噪声

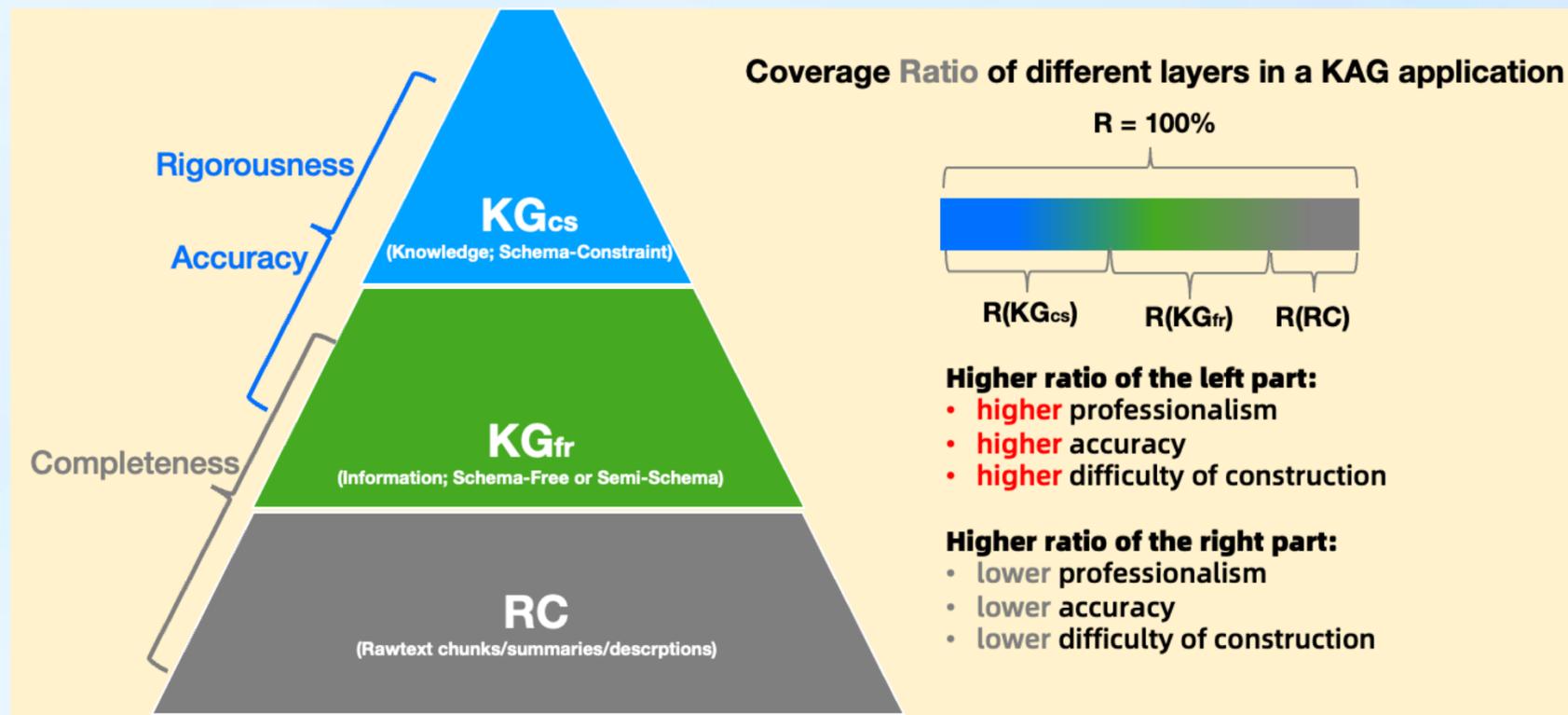
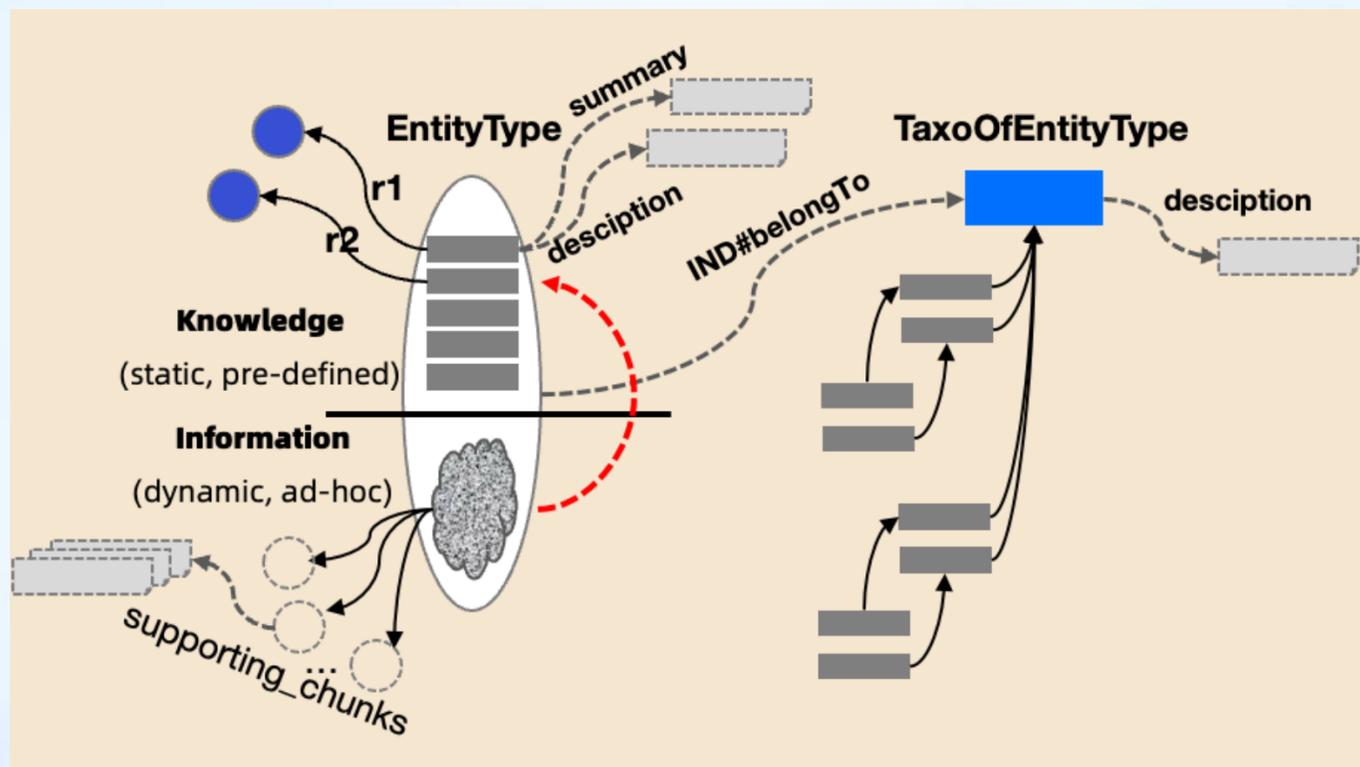
LLMs友好的知识表示(KG-Text互索引结构)、基于KGFabric的大规模知识管理



知识分类

- 结构化知识：结构化知识点：实体/事件/概念等
- 陈述性知识：知识点的自然语言陈述
- 过程性知识：知识点的领域符号表示：规则/算法/code/数学等
- 案例类知识：结合具体情景的知识点：病例、判例、题目、事件等
- 推理类知识：利用知识点通过分析/计算/归纳/演绎等得出的观点或结论

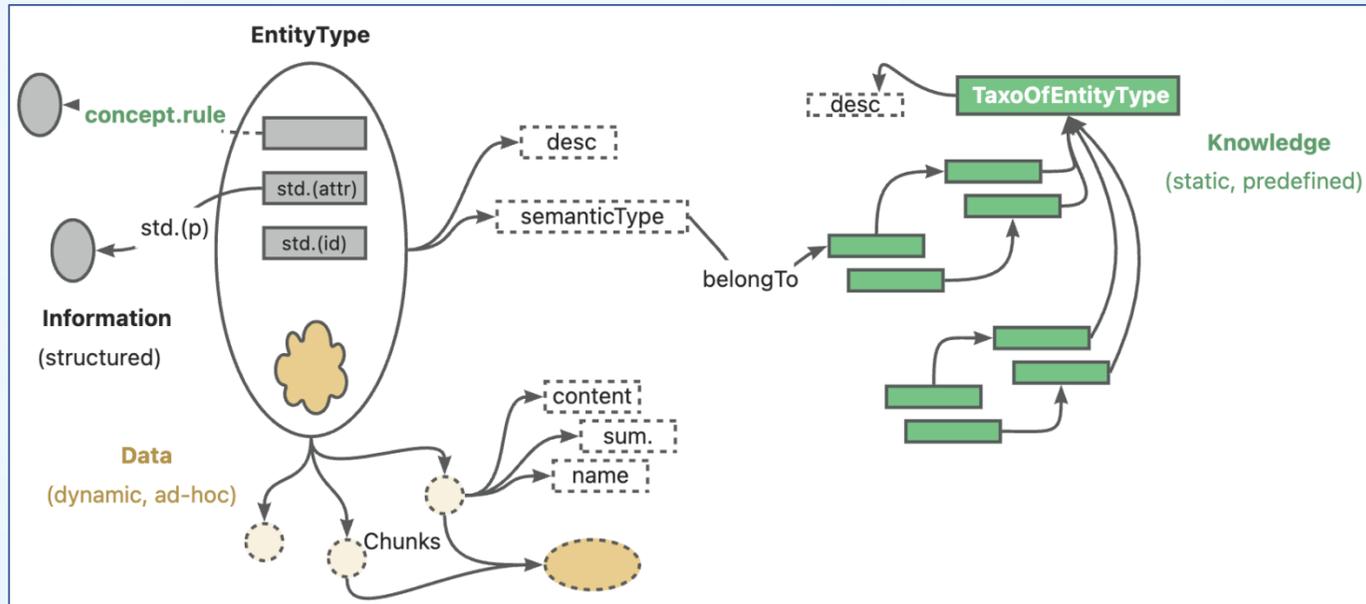
关键能力 1 严谨结构化知识与非结构化信息共存于同一实体空间，确保信息无损



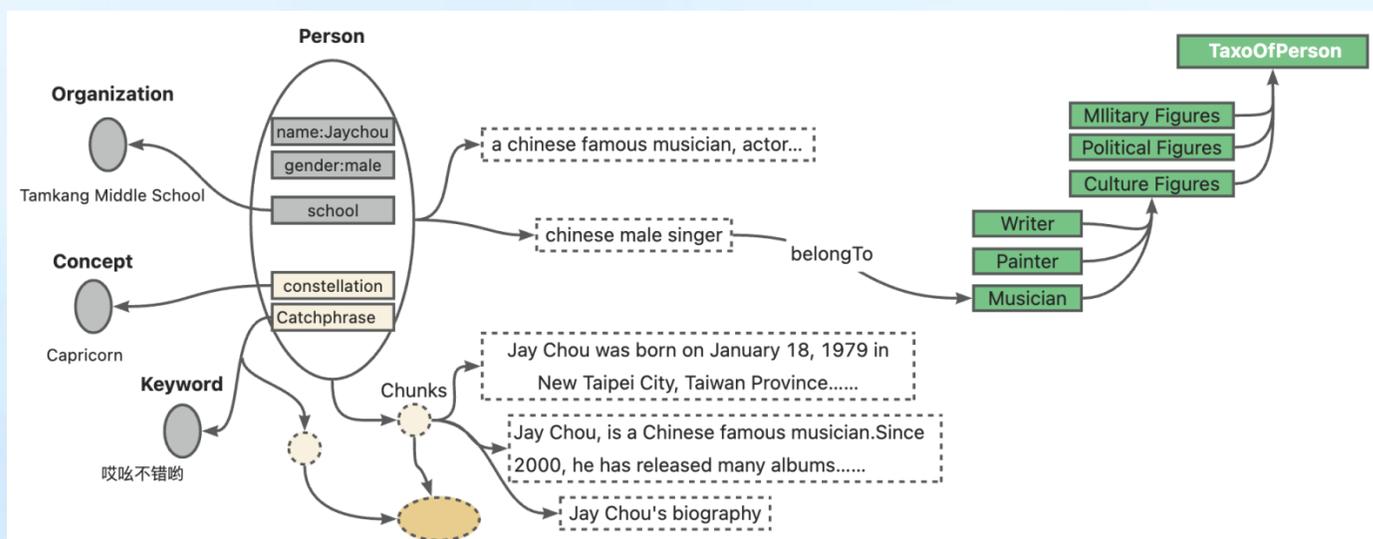
- 升级SPG为面向大模型友好的知识表示LLMFriSPG
- 兼容强Schema专业知识和弱Schema开放信息
- 图结构知识与文本知识的互索引结构
- 专业领域可平滑调节的专业决策与信息检索, 丰富知识完备性

关键能力 1 严谨结构化知识与非结构化信息共存于同一实体空间，确保信息无损

KAG – Indexing Structure



KAG – Indexing instance of Jay Chou



Organization[组织机构]: EntityType
 properties:
 id[主键]: Text
 index: TextAndVector
 name[机构名]: Text
 index: TextAndVector
 desc[描述]: Text
 index: TextAndVector
 semanticType[语义类型]: Text

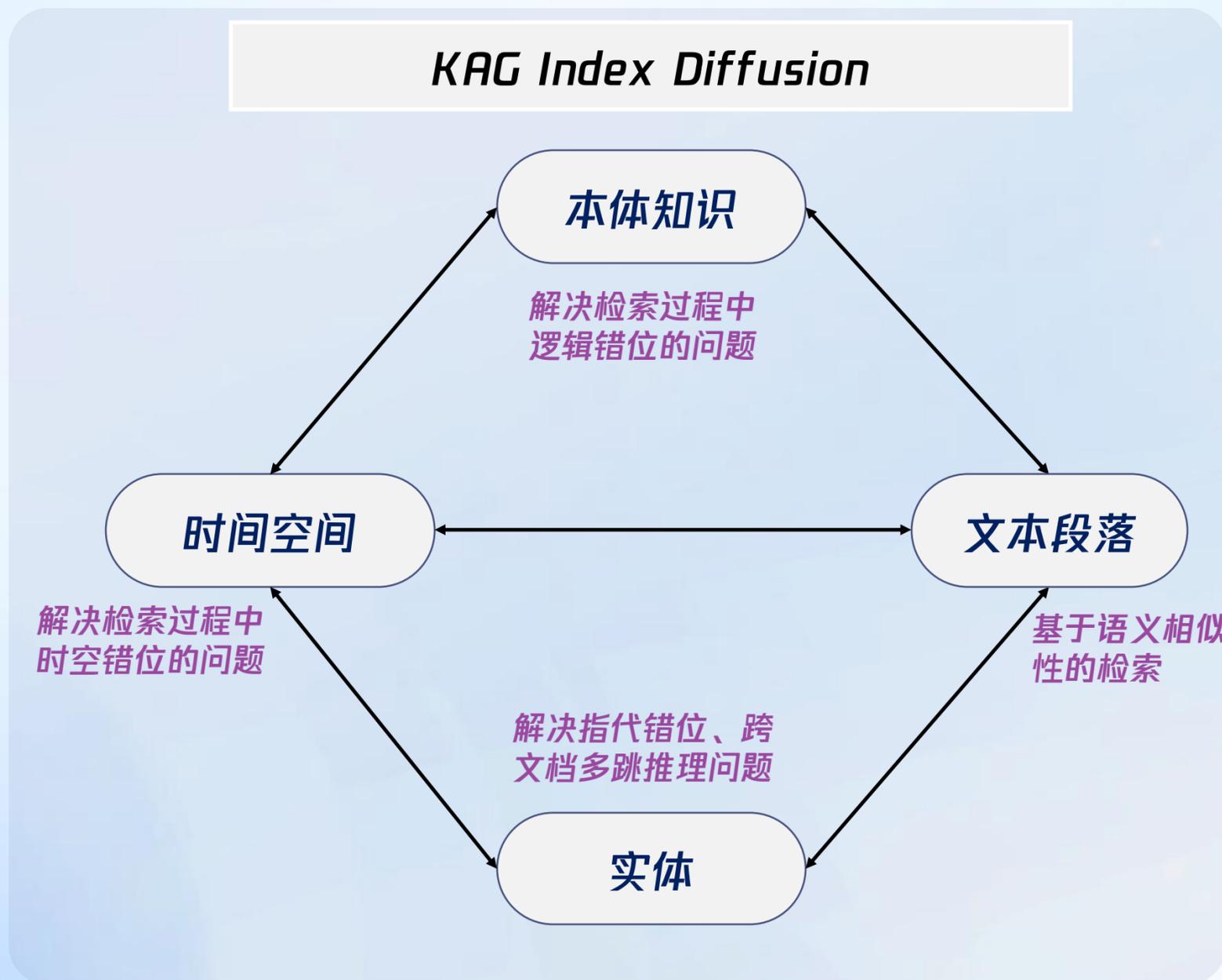
Person[人物]: EntityType
 properties:
 id[主键]: Text
 index: TextAndVector
 name[姓名]: Text
 index: TextAndVector
 desc[描述]: Text
 index: TextAndVector
 school[毕业院校]: Organization
 gender[性别]: Text
 semanticType[语义类型]: Text

Works[作品]: EntityType
 Concept[概念]: EntityType
 GeoLocation[地理位置]: EntityType

 Chunks[文章段落]: EntityType
 Others[其它]: EntityType

default.schema

关键能力 ① 严谨结构化知识与非结构化信息共存于同一实体空间，确保信息无损



问题类型	问题示例	描述
指代错位	1、俄罗斯总统访华 2、美国总统访华 3、普京抵达北京首都国际机场	<ul style="list-style-type: none"> • Bge-m3 $Simil.[1,2] = 0.85, Simil.[1,3] = 0.716$ • Qwen3-embedding $Simil.[1,2] = 0.740, Simil.[1,3] = 0.579$
时空错位	1、2024年5月30号 2、2023年5月30号 3、2024年6月1号	<ul style="list-style-type: none"> • Bge-m3 $Simil.[1,2] = 0.834, Simil.[1,3] = 0.811$ • Qwen3-embedding $Simil.[1,2] = 0.935, Simil.[1,3] = 0.604$
数值错位	1、刑法第71条 2、刑法第17条 3、偷窃190元 4、偷窃1900元 5、偷窃200元	<ul style="list-style-type: none"> • Bge-m3 $Simil.[1,2] = 0.62, Simil.[3,4] = 0.918, Simil.[3,5] = 0.841$ • Qwen3-embedding $Simil.[1,2] = 0.789, Simil.[3,4] = 0.990, Simil.[3,5] = 0.964$
逻辑错位	1、呼吸系统慢性疾病 2、哮喘 3、消化系统慢性疾病	<ul style="list-style-type: none"> • Bge-m3 $Simil.[1,2] = 0.60, Simil.[1,3] = 0.80$ • Qwen3-embedding $Simil.[1,2] = 0.558, Simil.[1,3] = 0.67$

关键能力 1 严谨结构化知识与非结构化信息共存于同一实体空间，确保信息无损

Document

During the Civil War, Secretary of the Navy, Gideon Welles, led the Northern Navy and decided to launch an attack on New Orleans.

Open-IE

(Gideon Welles, is, Secretary of the Navy)
 (Gideon Welles, lead, Northern Navy)
 (Northern Navy, attack, New Orleans)

Semantic Enhance

Ontology Labeling

(Northern Navy, belongTo, Organization)
 (New Orleans, belongTo, GeographicLocation)

Hyper-concept Generation

(Secretary of the Navy, isA, government officials)
 (Civil War, isA, war)

Concept connection

(government officials, has, political party)
 (war, has, casualties)
 (war, has, results)
 (political party, has, leader)

Synonym Generation

(Civil War, synonym, War between States)

Repository

Document 1: Abraham Lincoln took moderate positions designed to bring the South back into the union as quickly as possible, while Radical Republicans in Congress sought stronger measures to upgrade the rights of African Americans.

Document 2: Republicans regained control of the chamber they had lost in the 2006 midterm election ...



Question:

In the United States after the Civil War, the political party that favored harsh punishment of the southern states gained control of which institution in 2010?

Semantic Retrieval Syntax

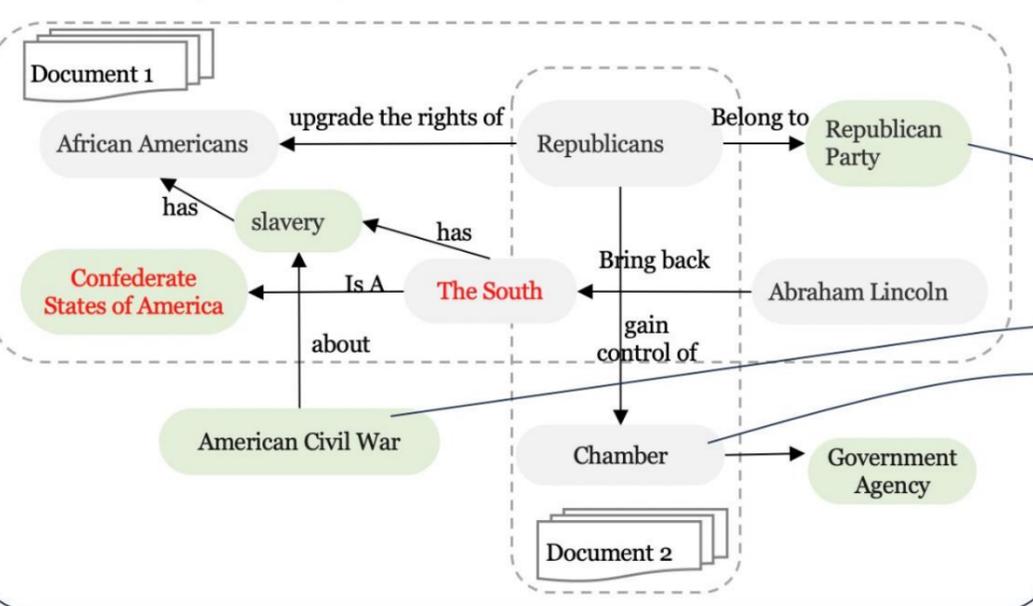
Step1: What's the name of the political party that that favored harsh punishment of southern states after the Civil War?

Action1: Retrieval(s=s1: political party, p=p1:punish, o=o1:Location[southern states])

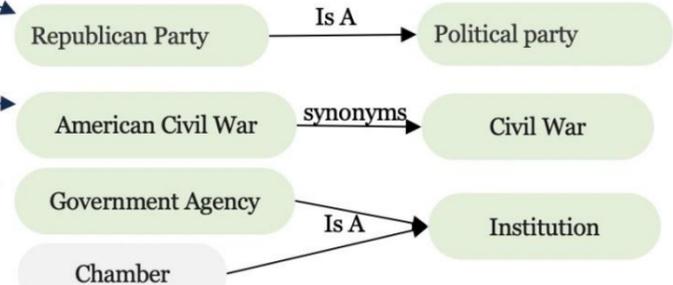
Step2: Which institution did the political party take control of in 2010?

Action2: Retrieval(s=s1, p=p2: control, o=o2: Institution)

Semantic Graph Indexing



Semantic Retrieval Scorer



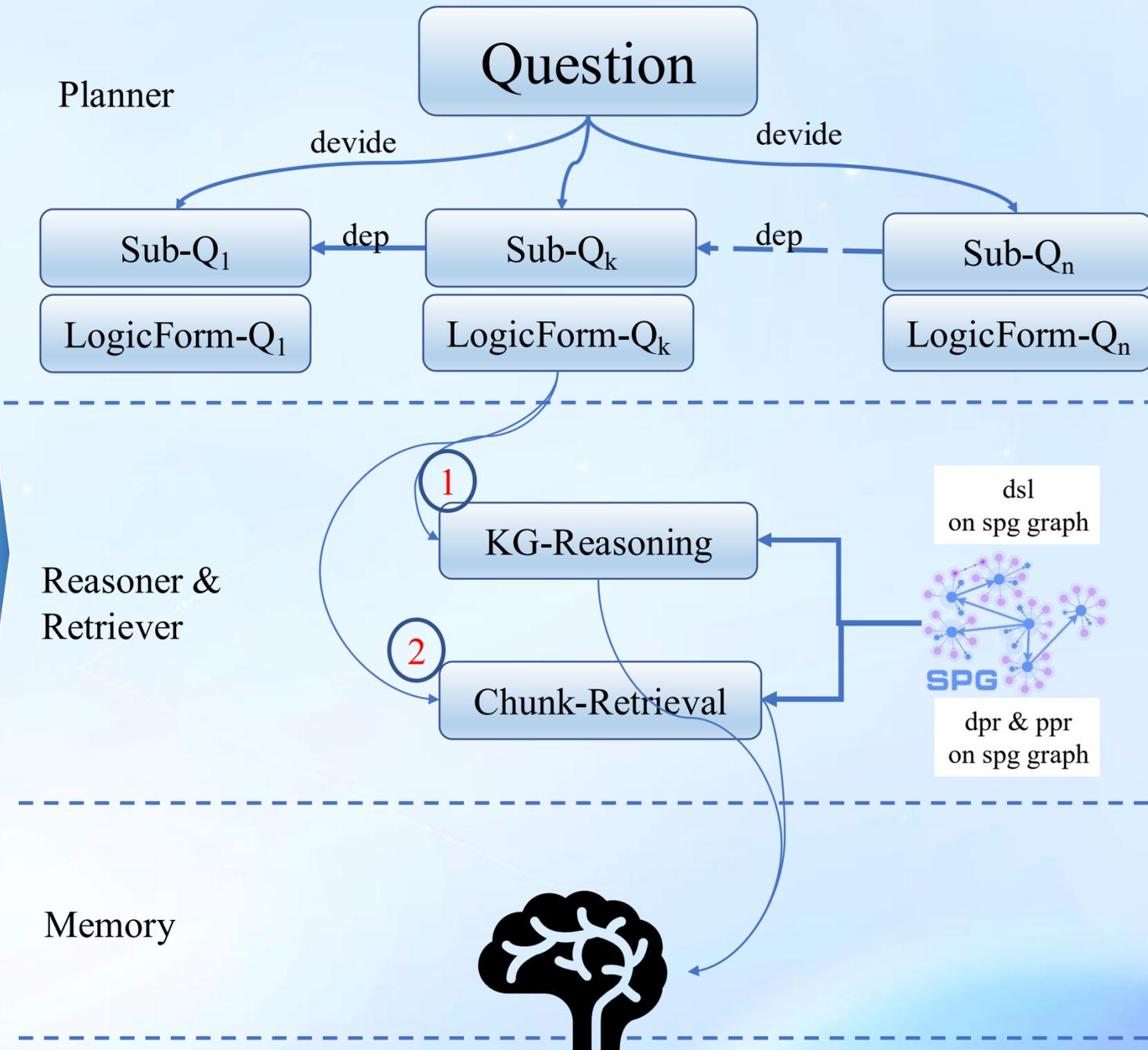
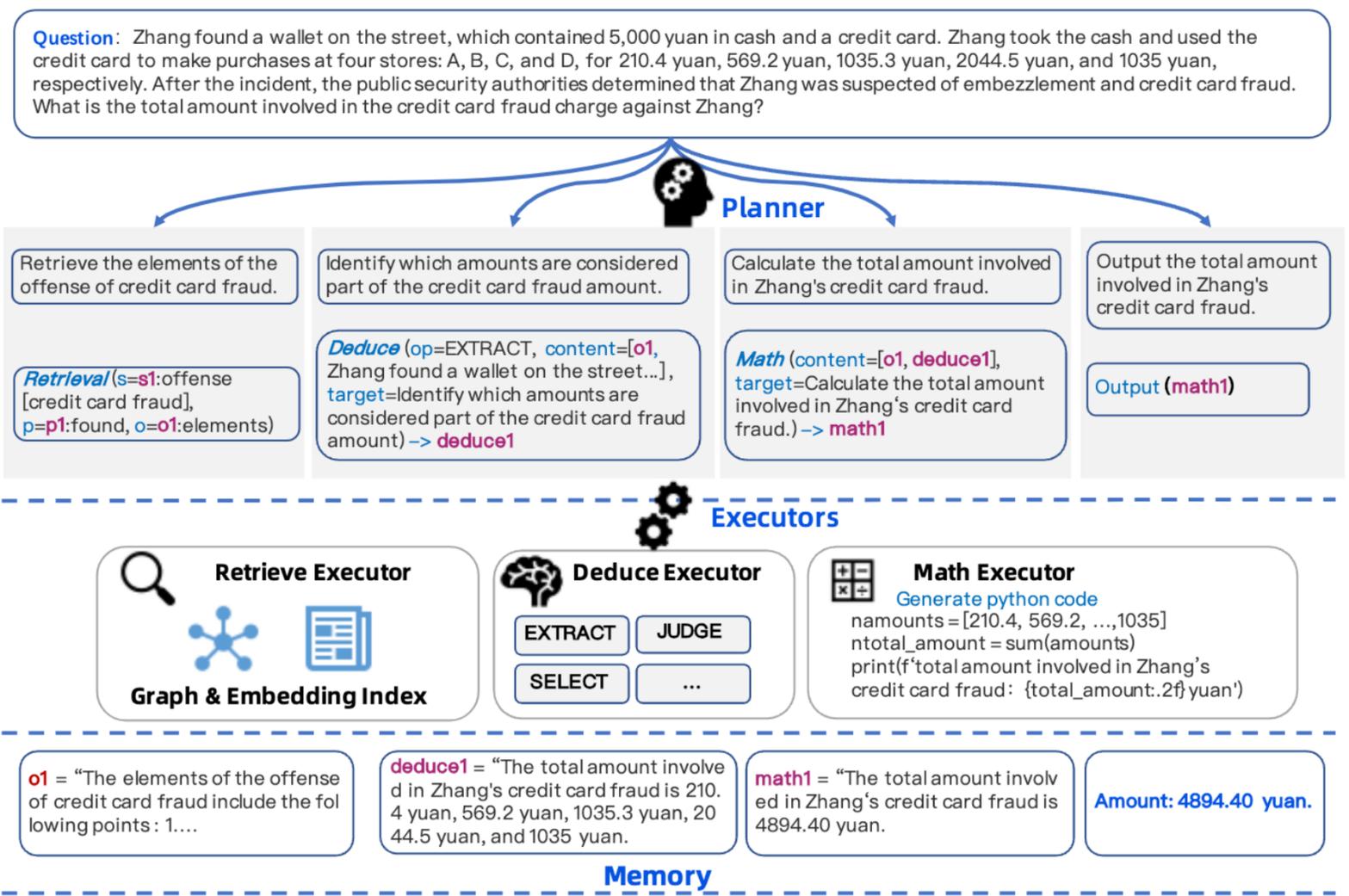
Generation

语义增强以提升连通性、降低噪声

基于Logical-form对齐用户问题与知识索引

关键能力 2

逻辑符号引导的结构化知识推理，拆解用户问题的逻辑结构依赖

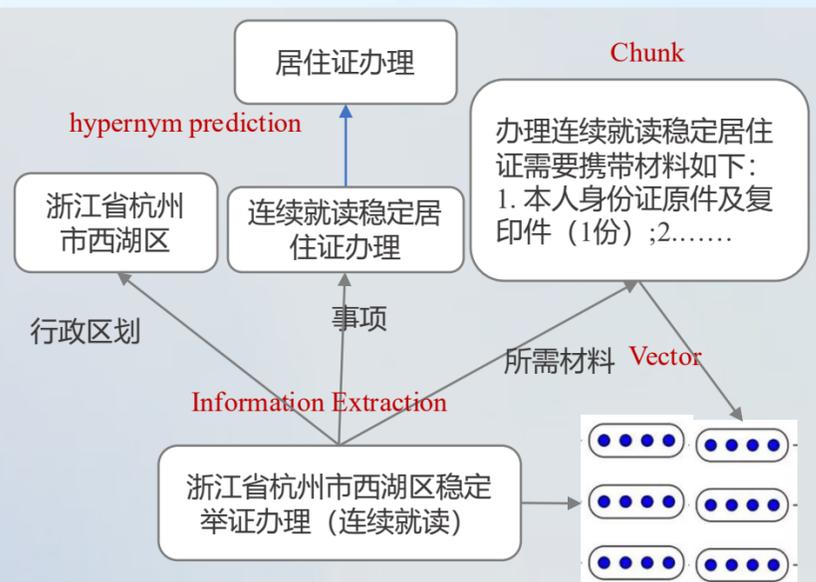


关键能力 2

逻辑符号引导的结构化知识推理，拆解用户问题的逻辑结构依赖

Function Name	Function Declaration
Retrieval	$Retrieval(s = s_i : type[name], p = p_i : edge, o = o_i : type[name], s.prop = value, p.prop = value, o.prop = value)$
Sort	$Sort(A, direction = min max, limit = n)$
Math	$math_i = Math(expr)$, expr is in LaTeX syntax and can be used to perform operations on sets. e.g. count: $\ A\ $, sum: $\sum A$
Deduce	$Deduce(left = A, right = B, op = entailment greater less equal judgement)$
Ask	$Ask(content = c, op = choice answer)$, Generates a query to the user based on the previous content. op represents the option to return a selective or answer query.
Output	$Output(A, B, C, \dots)$

Table 1: Functions of logical form.



在读博士申请居住证是否需要学校开具在读证明

Logic form = F(query + graph schema)

- 查找博士办理居住证的办事指南
Retrieval(s=s1:在读博士申请居住证|政务办事, p=p1:材料证明, o=o1:材料证明)
- 判断材料中是否需要读证明
Deduce(left=o1, right=在读证明, op=entailment)

用户问题拆解成严谨的多跳符号逻辑表示

UIE+概念推理+向量化 [建立互索引]

Did Olivia Wilde's work period start after 1983?

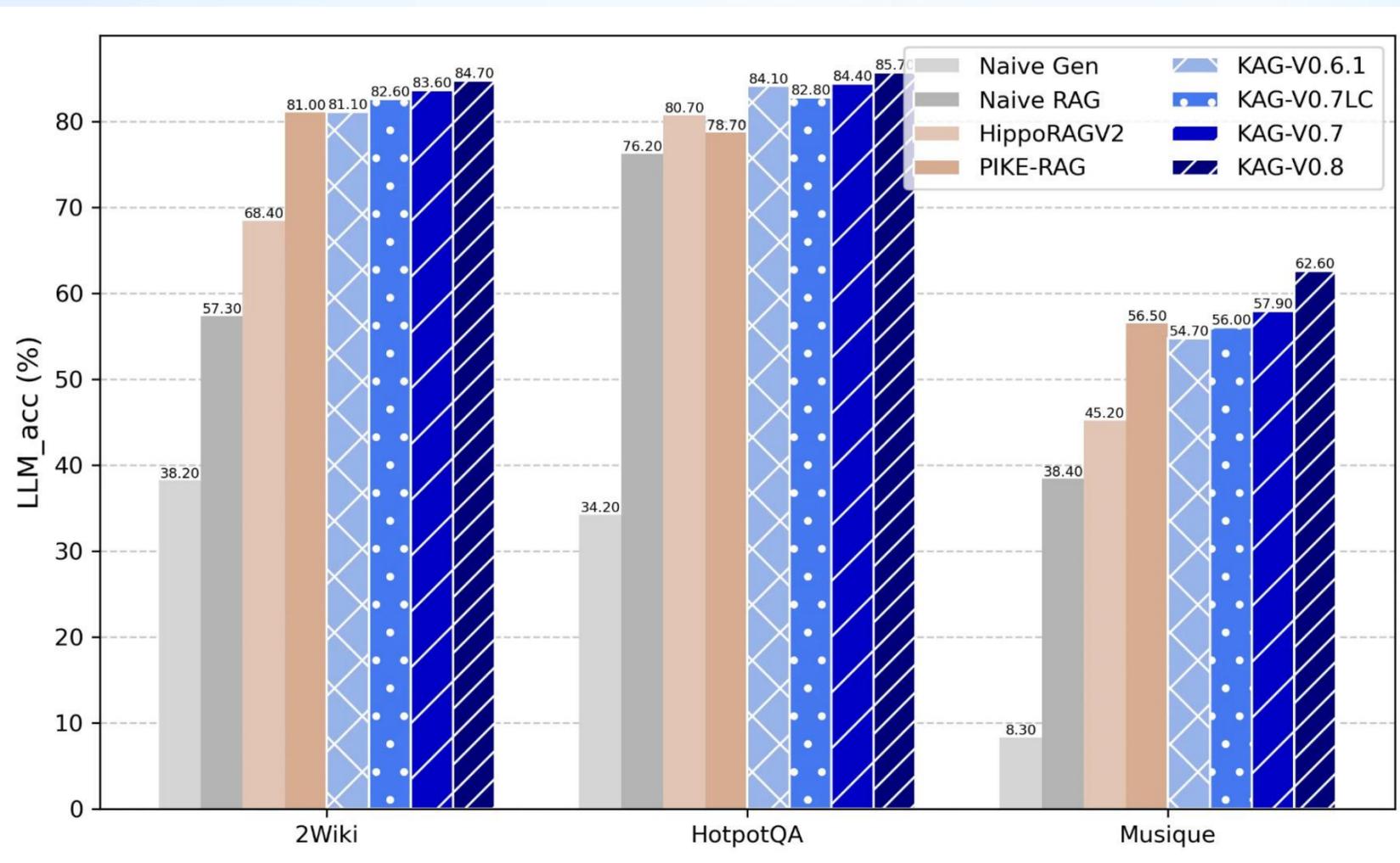
- Step1: Identify the work period of Olivia Wilde.
Action1: Retrieval(s=s1:Person[Olivia Wilde], p=p1:workPeriodStart, o=o1:Year)
Step2: Compare the work period start year to 1983.
Action2: Deduce(op=judgement)
- 示例：逻辑判定与LLM交叉验证

Doctor, I have been experiencing sudden and frequent panic attacks. I don't know what to do.

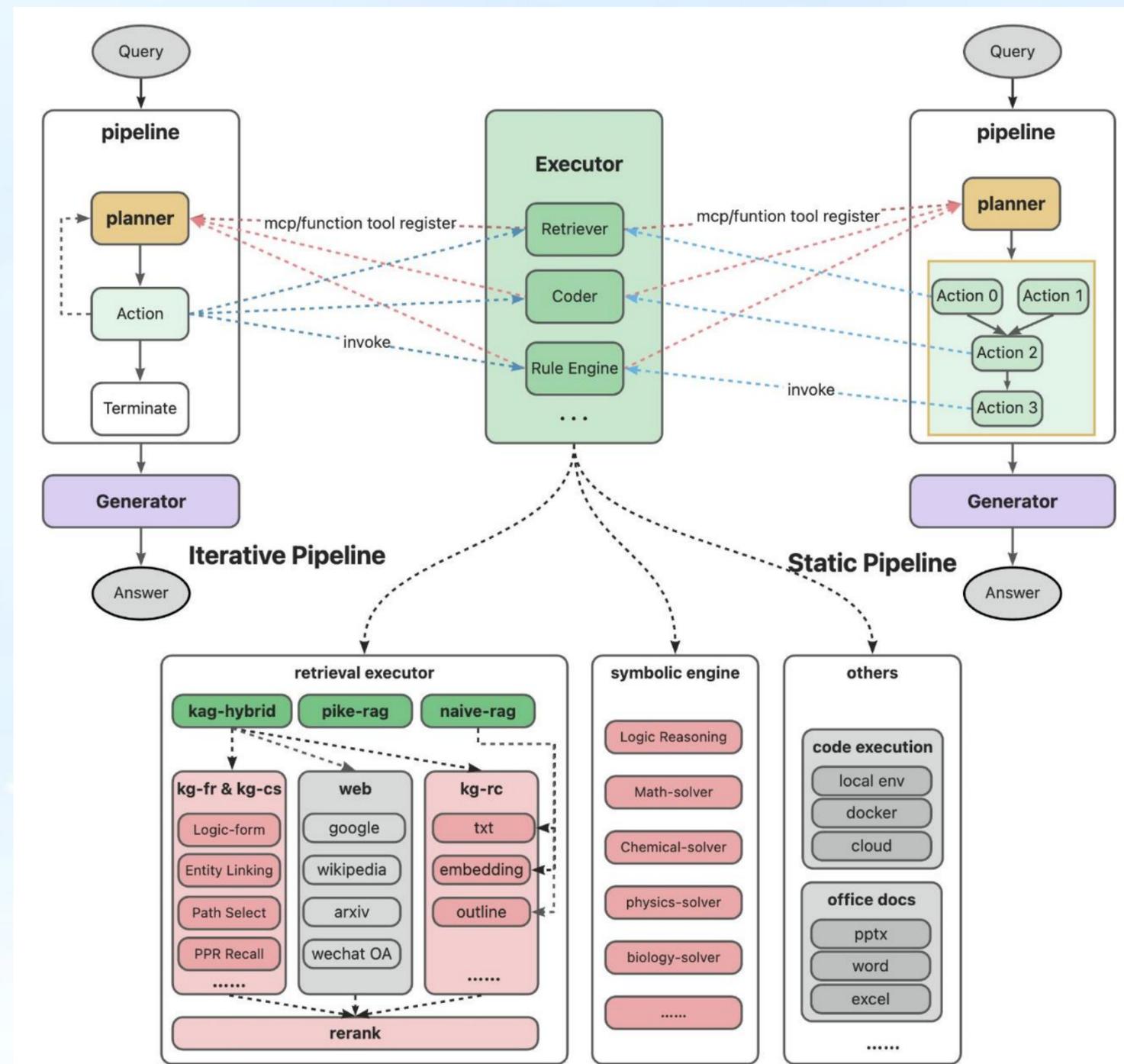
- Step1: Identify the symptoms experienced by the patient.
Action1: Retrieval(s=s1:Patient, p=p1:experiences, o=o1:Symptom[panic attacks], o.frequency=frequent, o.onset=sudden)
Step2: Determine the possible causes or triggers for panic attacks.
Action2: Retrieval(s=o1, p=p2:associatedWith, o=o2:Cause)
Step3: Suggest potential treatments or interventions for panic attacks.
Action3: Retrieval(s=o1, p=p3:treatedWith, o=o3:Treatment)
Step4: Output the suggested treatments.
Action4: Output(o3)
- 示例：医疗多跳推理问答

How many towns' postal code is VLT and area is less than 530 square miles?"

- Step1: Identify towns with postal code VLT.
Action1: Retrieval(s=s1:Town, p=p1:postalCode, o=o1:PostalCode[VLT])
Step2: Filter towns with area less than 530 square miles.
Action2: Retrieval(s=s1, p=p2:area, o=o2:area, o.area<530)
Step3: Count the number of towns that meet the criteria.
Action3: Math(count(s1))
Step4: Output the count.
Action4: Output(s1)
- 示例：检索中的数值计算



提供Lightweight Construct (LC) 模式，在保证效果损失较小的情况下，构建成本降低90%+

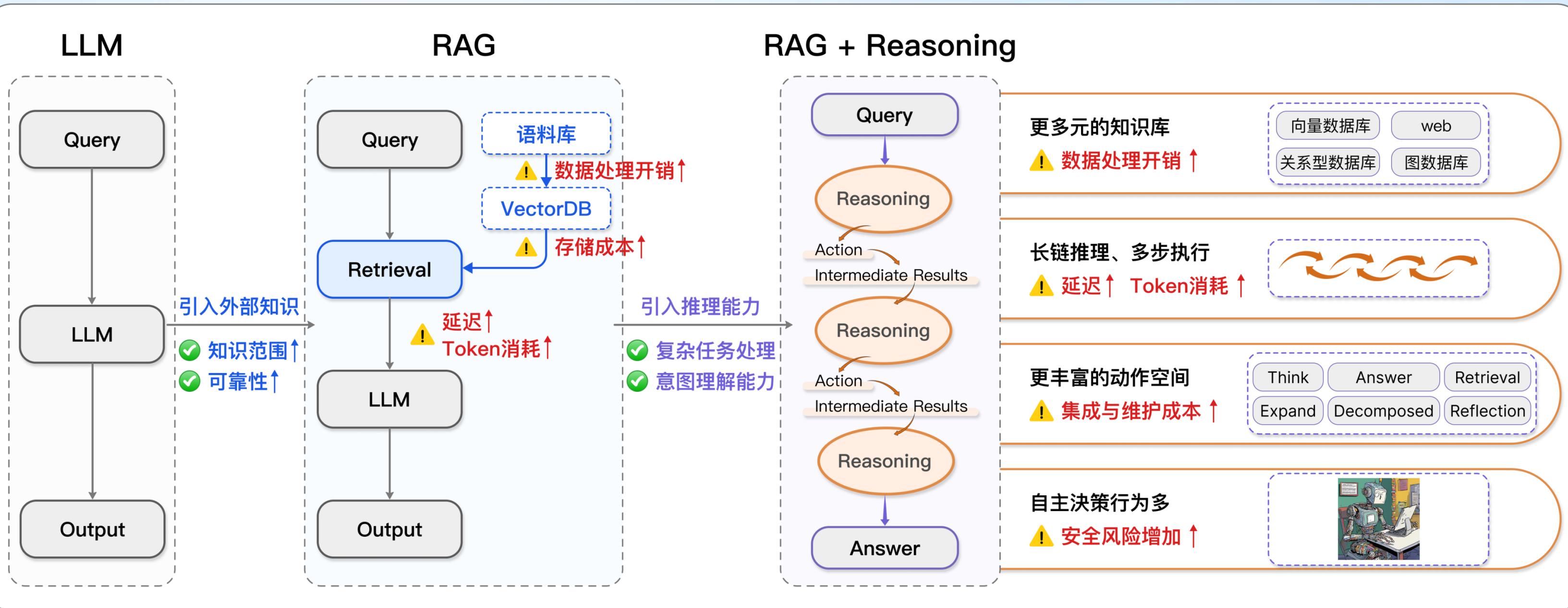


04

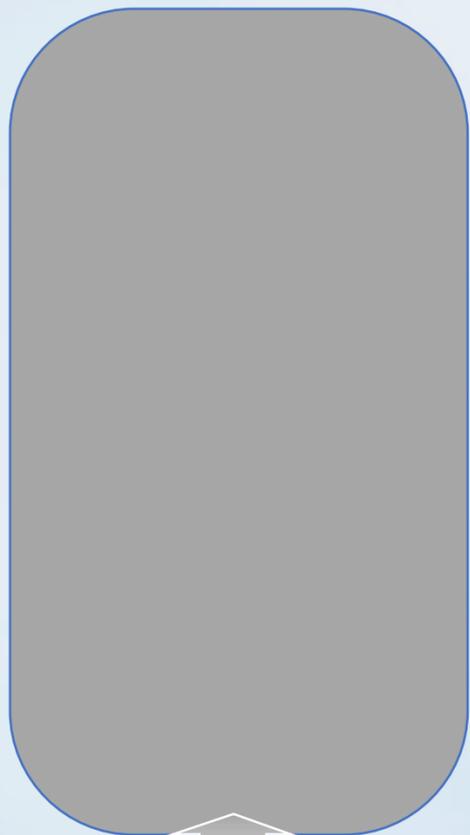
*KAG-Thinker*思考模型

将KAG框架能力转换成模型内生能力，使模型具备严谨思考和利用外部知识库的能力，训练得到KAG-Thinker

RAG与深度推理的结合，使得能够解决更加复杂的问题，但是也带来更高的成本，而这一点却常常被忽略
从Standalone LLM到RAG再到RAG+Reasoning，每次技术升级都伴随着成本



RL增强LLM检索



Search-R1/ReSearch/
ZeroSearch

KAG-Thinker的推理范式

逻辑性稳定性: 广度拆解和深度求解

检测知识边界: 推理反思及logits检测

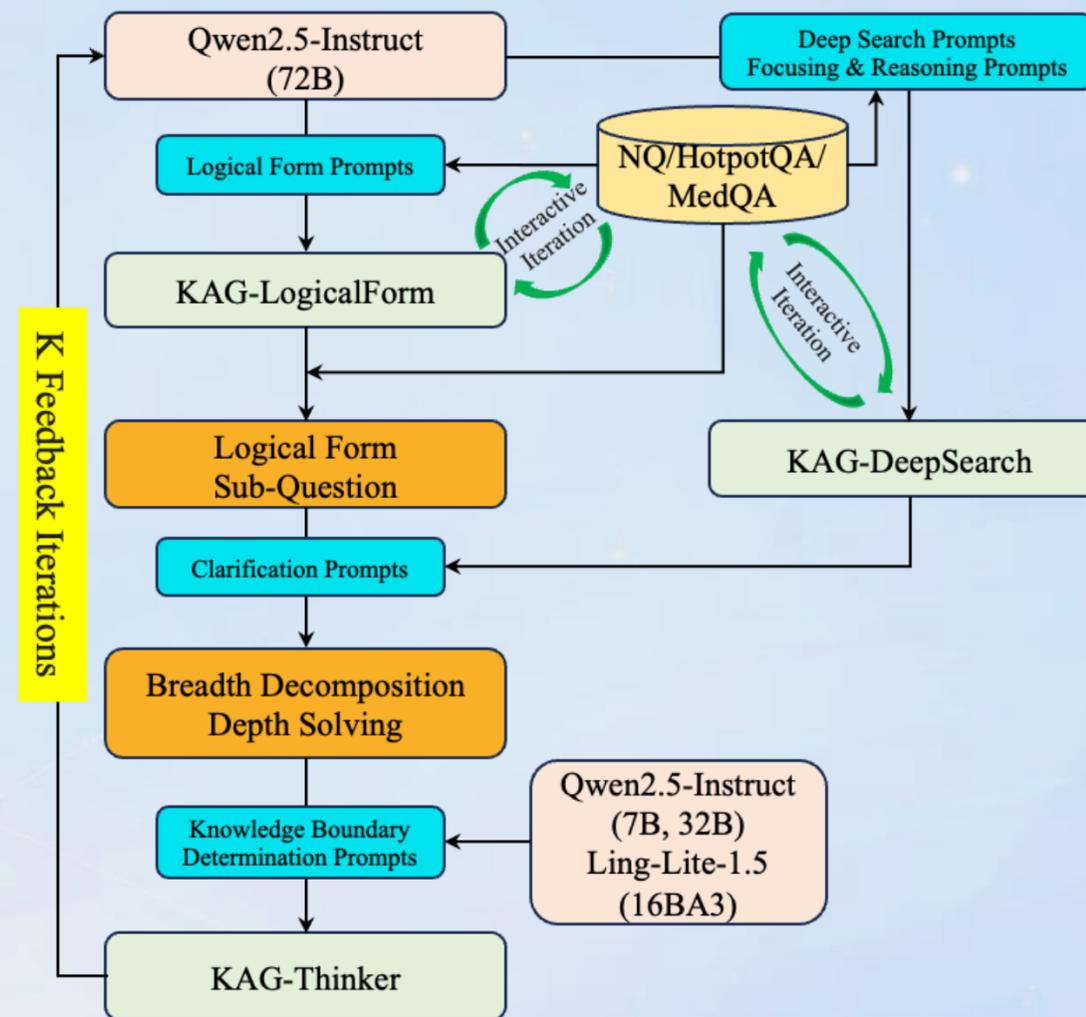
求解严谨性: 符号嵌入推理过程

检索噪音鲁棒性: Reference分析

过程约束

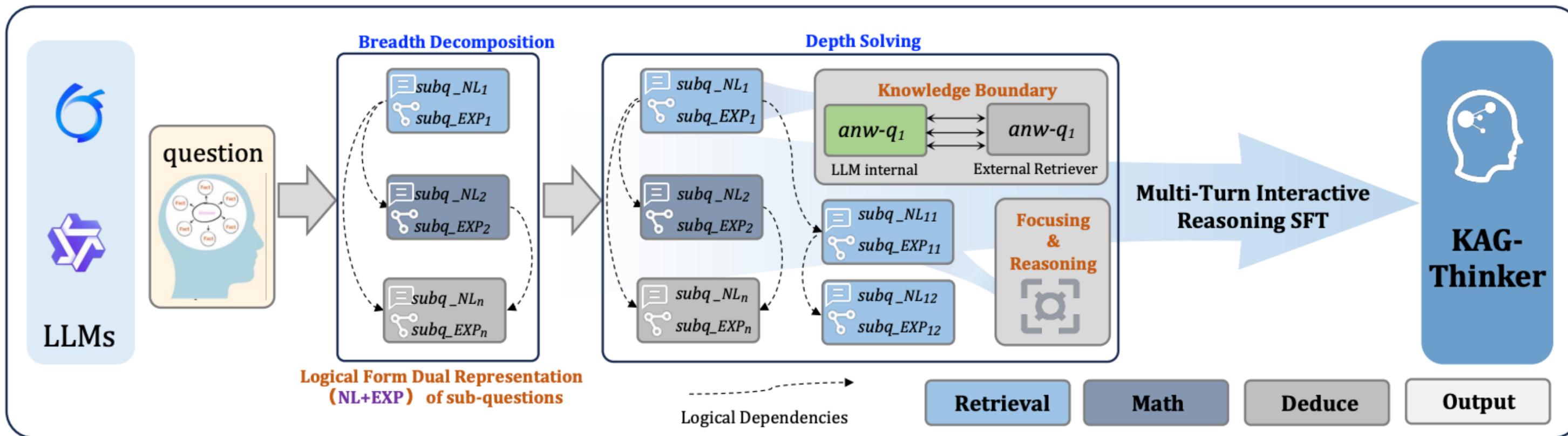
LLM+KB符号化思考范式

推理过程样本合成



合成SFT语料, 使模型学习思考范式

KAG-Thinker Model: Corpus Synthesis and Training Pipeline

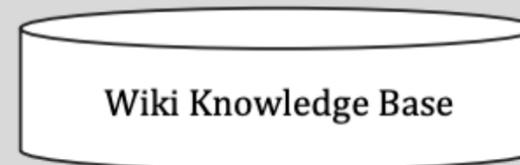
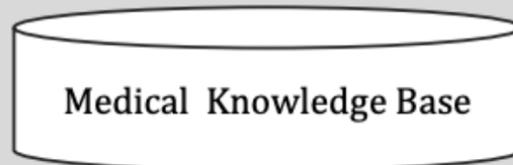
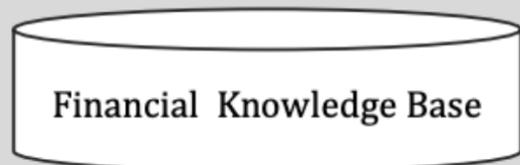
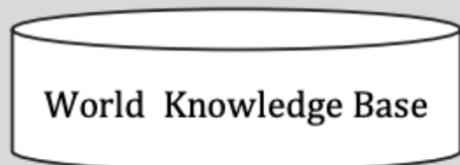


Hybrid Graph Retriever

BGE-M3 Retriever

KAG Index: Automated construction of domain knowledge

Text-based Vector Index

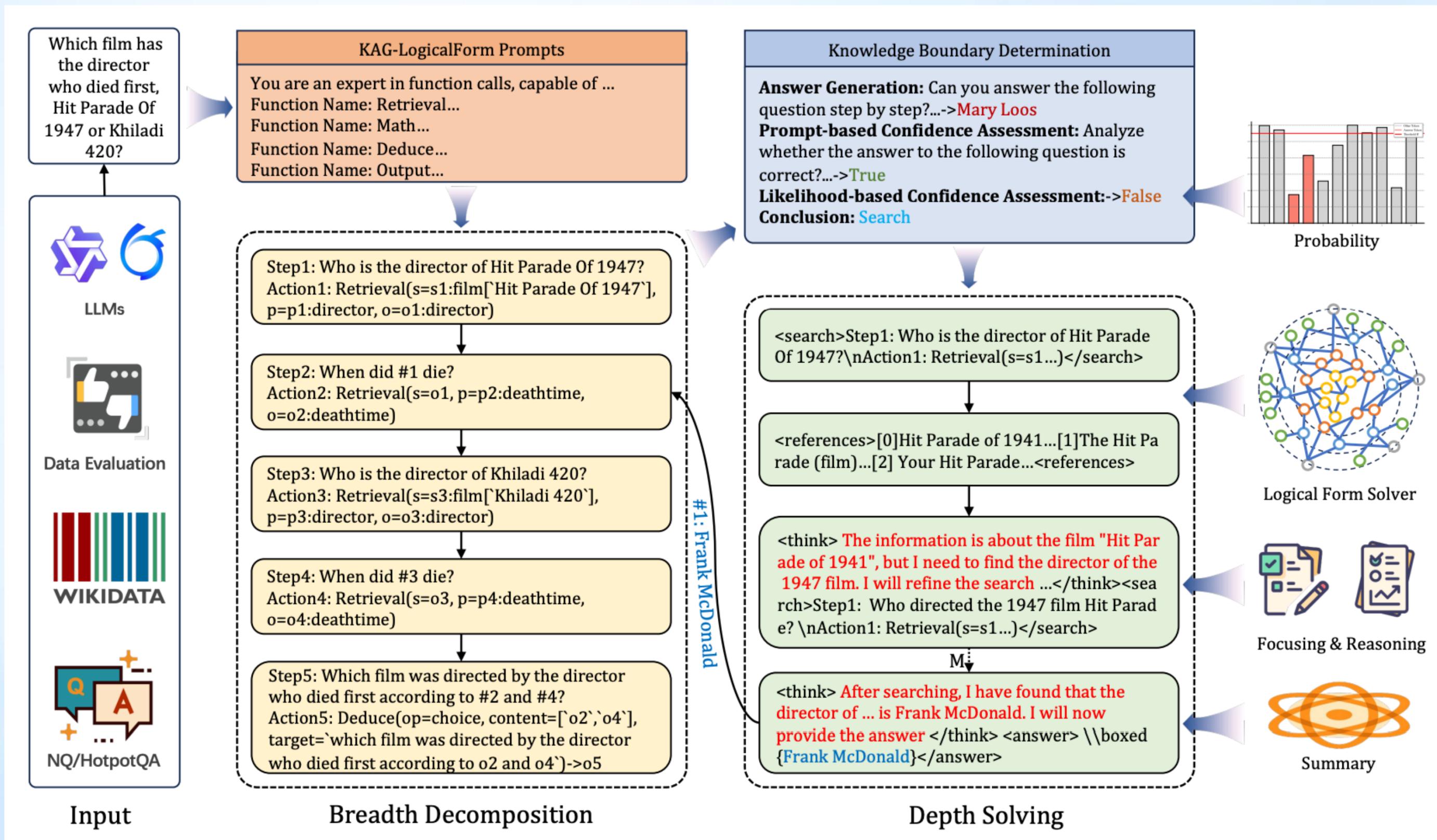


Encyclopedic/Conceptual Knowledge

Financial Research Reports / Events / Common Knowledge

Medical Textbooks / Websites

Encyclopedic



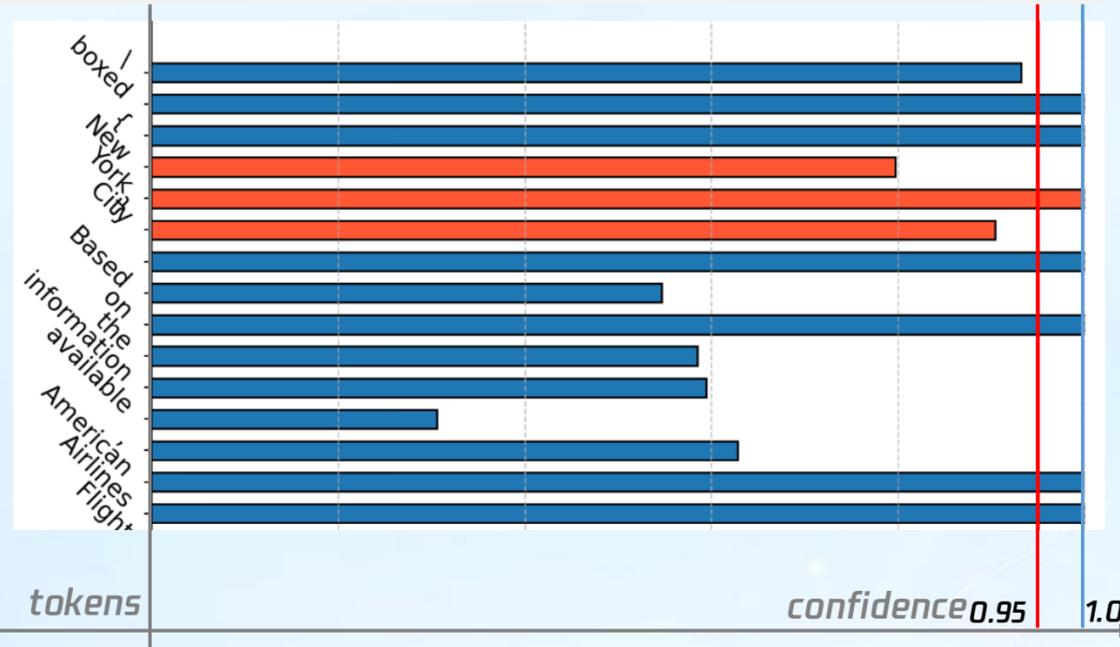
AI 模型能力：检索时机，提升LLM自身知识边界的判定能力



直接回答原始问题 或 子问题相对复杂

User: American Airlines Flight 11 was flying away from the airport in what major city?

LLM: **New York City**



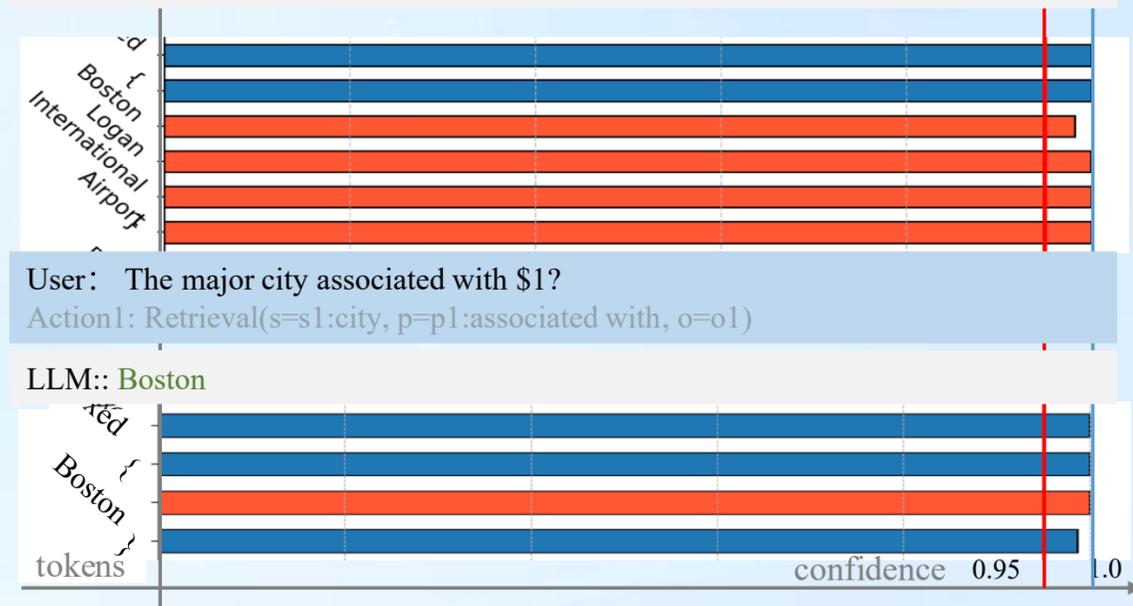
问题未被拆开，模型就直接回答，回答错误且confidence较低



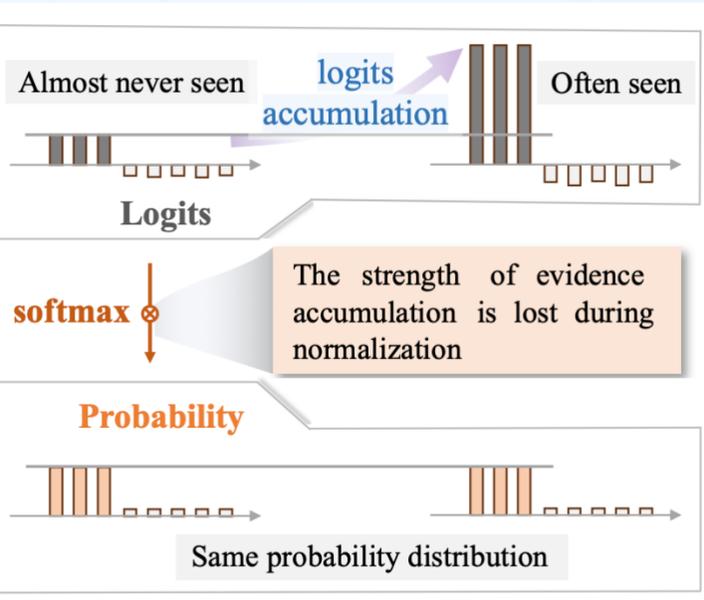
Thinker Logical-form约束子问题：单实体一跳检索

User: Step 1: The airport from which American Airlines Flight 11 was flying? Action1: Retrieval(s=s1:airline[`American Airlines Flight 11`], p=p1:fly away, o=o1:airport)

LLM: Logan International Airport



按Logical-form粒度约束拆分问题，回答正确且confidence较高



LLMs对训练阶段高频见过的知识表现出较高的累积logits

思考过程: Can you answer the following questions step by step? If you can, wrap your answer with <answer>\\boxed{your answer}</answer>. If you can't, just reply that based on my internal knowledge, I can't answer this question, I need to retrieve external knowledge. \\nQuestion: Step1: when does ellen's new game show start\\nAction1: Retrieval(s=s1:show[`Ellen's New Game Show`], p=p1:startDate, o=o1:date)

<think>based on my internal knowledge, I can't answer this question, I need to retrieve external knowledge.</think>

Cited from Estimating LLM Uncertainty with Logits 2025

AI KAG-Thinker 模型效果分析

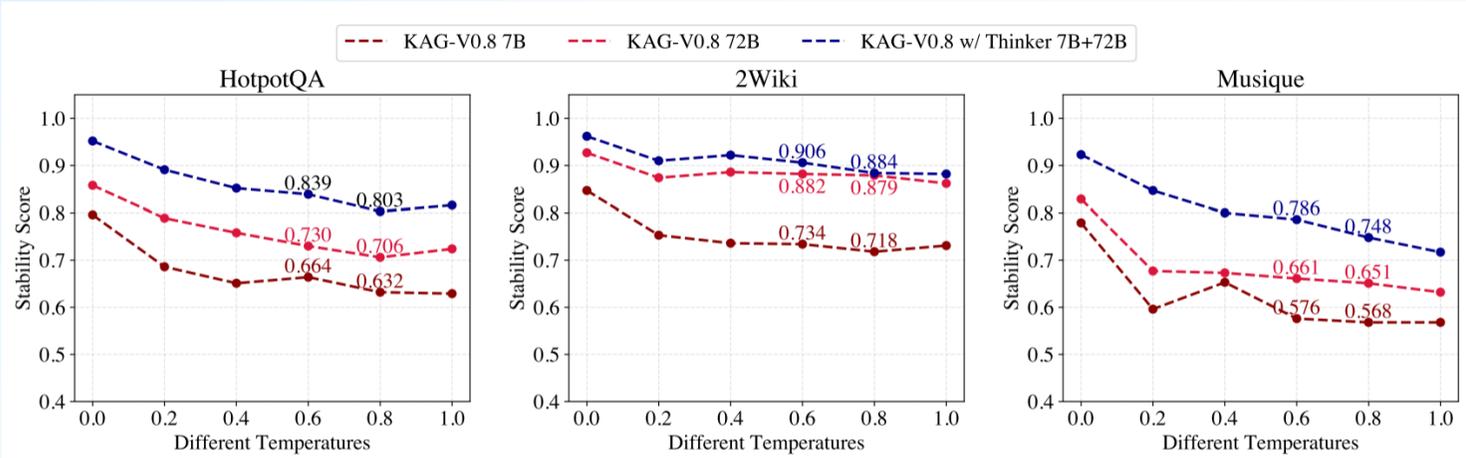


Figure 5: Stability under different temperatures.

	Single-Hop QA			Multi-Hop QA				Avg
	NQ [†]	TriviaQA [*]	PopQA [*]	HotpotQA [†]	2Wiki [*]	MuSiQue [*]	Bamboogle [*]	
Naive Generation	0.134	0.408	0.140	0.183	0.250	0.031	0.120	0.181
CoT	0.048	0.185	0.054	0.092	0.111	0.022	0.232	0.106
Search-o1	0.151	0.443	0.131	0.187	0.176	0.058	0.296	0.206
IRCoT	0.224	0.478	0.301	0.133	0.149	0.072	0.224	0.226
Naive RAG	0.349	0.585	0.392	0.299	0.235	0.058	0.208	0.304
R1-Gen	0.270	0.537	0.199	0.237	0.292	0.072	0.293	0.271
Search-R1	0.393	0.610	0.397	0.370	0.414	0.146	0.368	0.385
ZeroSearch	0.436	0.652	0.488	0.346	0.352	0.184	0.278	0.391
StepSearch	-	-	-	0.386	0.366	0.226	0.400	-
ReSearch	0.407	0.611	0.423	0.419	0.412	0.205	0.400	0.411
KAGThinker	0.450	0.642	0.484	0.421	0.469	0.221	0.480	0.452

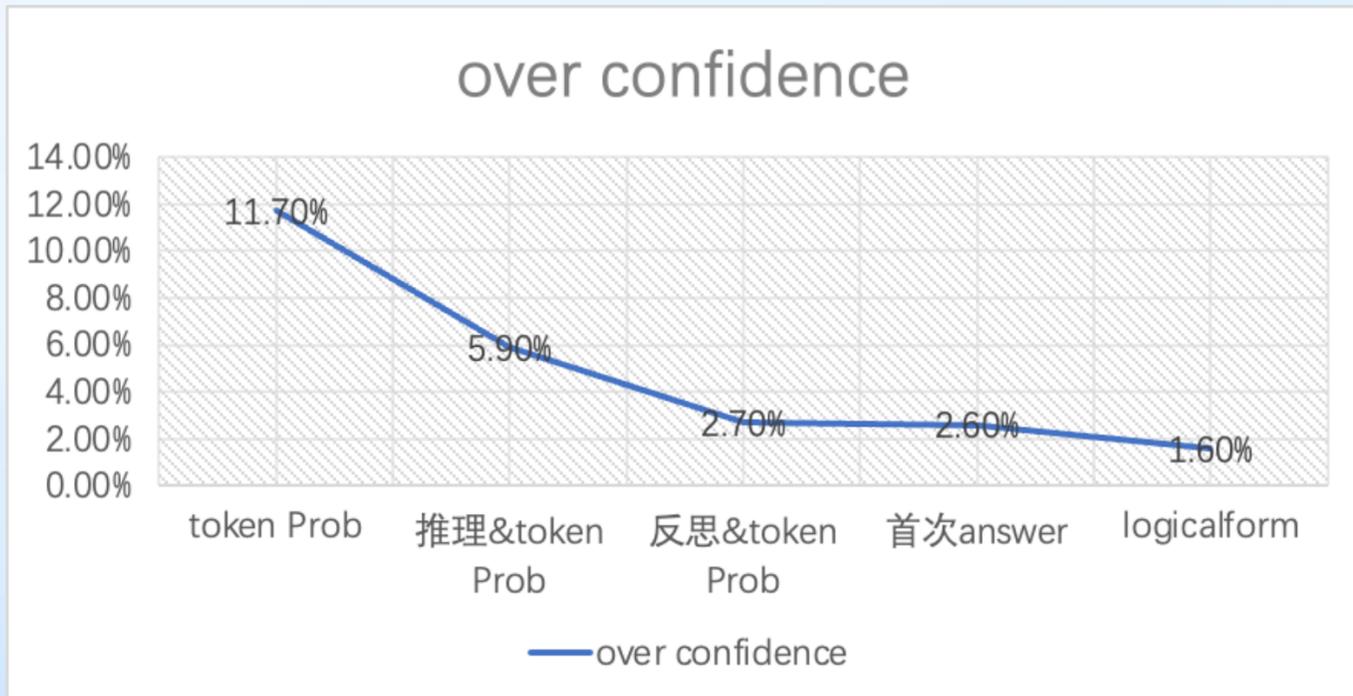
Table 7: EM performance of different models on Qwen2.5-7B-Instruct. The best performance is set in bold. [†]/^{*} represents in-domain / out-domain datasets. In contrast to other baselines, StepSearch and ReSearch employ the Musique dataset for training.

基于维基百科 [大库] 的QA数据集

LLM	Methods	MedQA
Meta-LLAMA-3.1-8B-Instruct	Naive Generation Qwen et al. (2025)	58.33
	Naive RAG Lewis et al. (2020)	61.54
	Naive RAG+adaptive Jeong et al. (2024)	61.33
	IRCoT Trivedi et al. (2023)	62.67
	ReAct Yao et al. (2023)	39.22
	KAG-Med-Thinker	74.00
Qwen2.5-14B-Instruct	Naive Generation Qwen et al. (2025)	84.00
	Naive RAG Lewis et al. (2020)	81.67
	Naive RAG+adaptive Jeong et al. (2024)	85.00
	IRCoT Trivedi et al. (2023)	82.33
	ReAct Yao et al. (2023)	86.20
	KAG-Med-Thinker	87.00
DeepSeek-R1-Distill-Qwen-14B	Naive Generation DeepSeek-AI et al. (2025)	79.67
	Naive RAG Lewis et al. (2020)	79.00
	Naive RAG+adaptive Jeong et al. (2024)	81.48
	IRCoT Trivedi et al. (2023)	81.33
	ReAct Yao et al. (2023)	80.87
	KAG-Med-Thinker	85.28

Table 10: Accuracy of different models on the MedQA dataset. The best performance is set in bold.

基于自建医疗检索库的MedQA数据集



Thinker在维基百科的QA数据集上模型过度自信曲线 [不知道自己不知道]

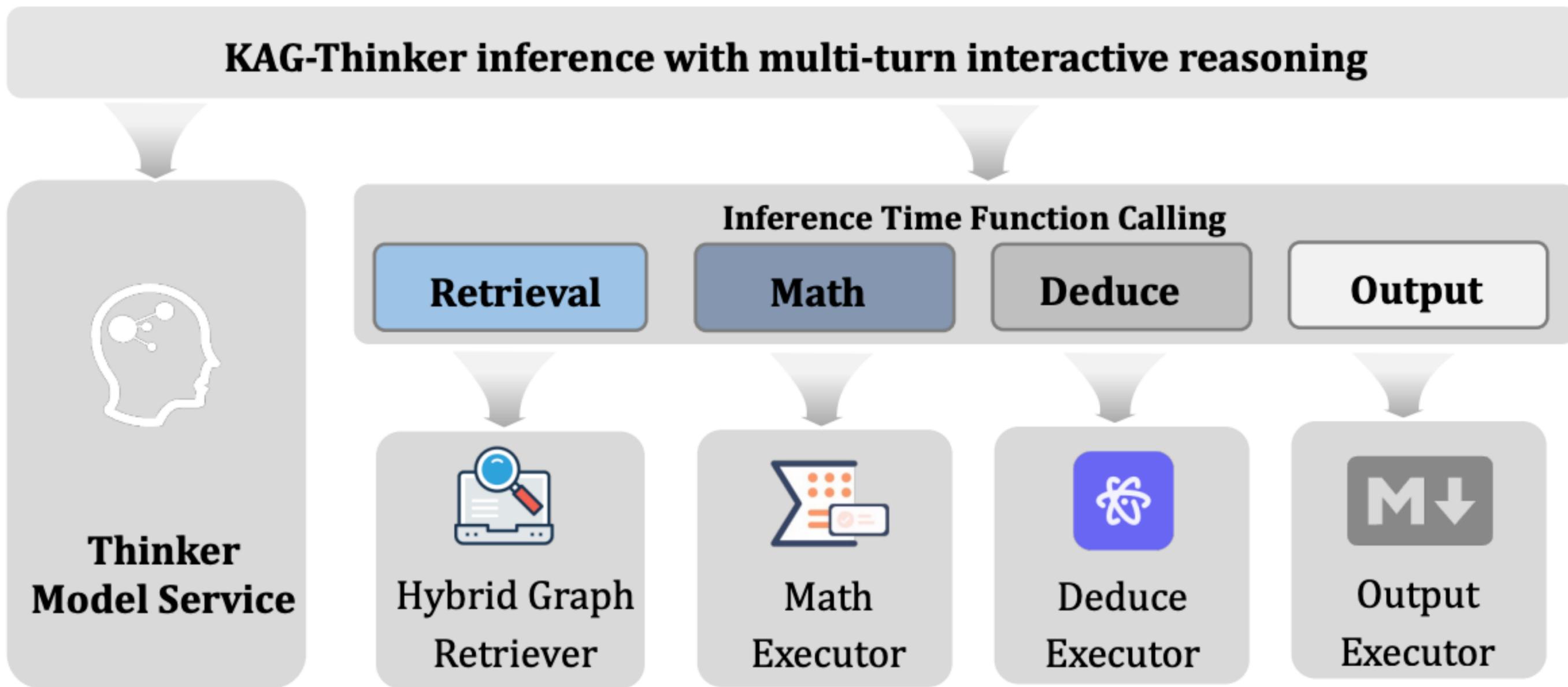


Figure 4: Overview of KAG-Thinker multi-turn interactive thinking and deep reasoning.



疾病诊断推理

+ 新建对话

患者，女，25岁，5天前出现噪...

患者男性，30岁，最近2天大便...

男孩，3岁。体重16kg，双下肢...

男孩，3岁。体重16kg，双下肢...

患者男性，30岁，最近2天大便...

患者男性，30岁，最近2天大便...

患儿男，3岁，自昨晚七点开始...

糖尿病早期有什么症状？

患者，女，39岁，自述气血两虚...

我有点喉咙疼，可能是什么病？



患者，女，25岁，5天前出现嗓子有痰堵塞的症状，伴有轻微咳嗽，现已不咳。期间服用“蒲地蓝”和“急支糖浆”进行治疗。患者自觉咽部有灼痛感，吞咽时加重，体温最高达38.2°C，咽部视诊见明显充血和肿胀。根据患者的症状和检查结果，最有可能的诊断是什么？

开始时间：2025-06-12 14:36:44

生成回答中...

请输入~

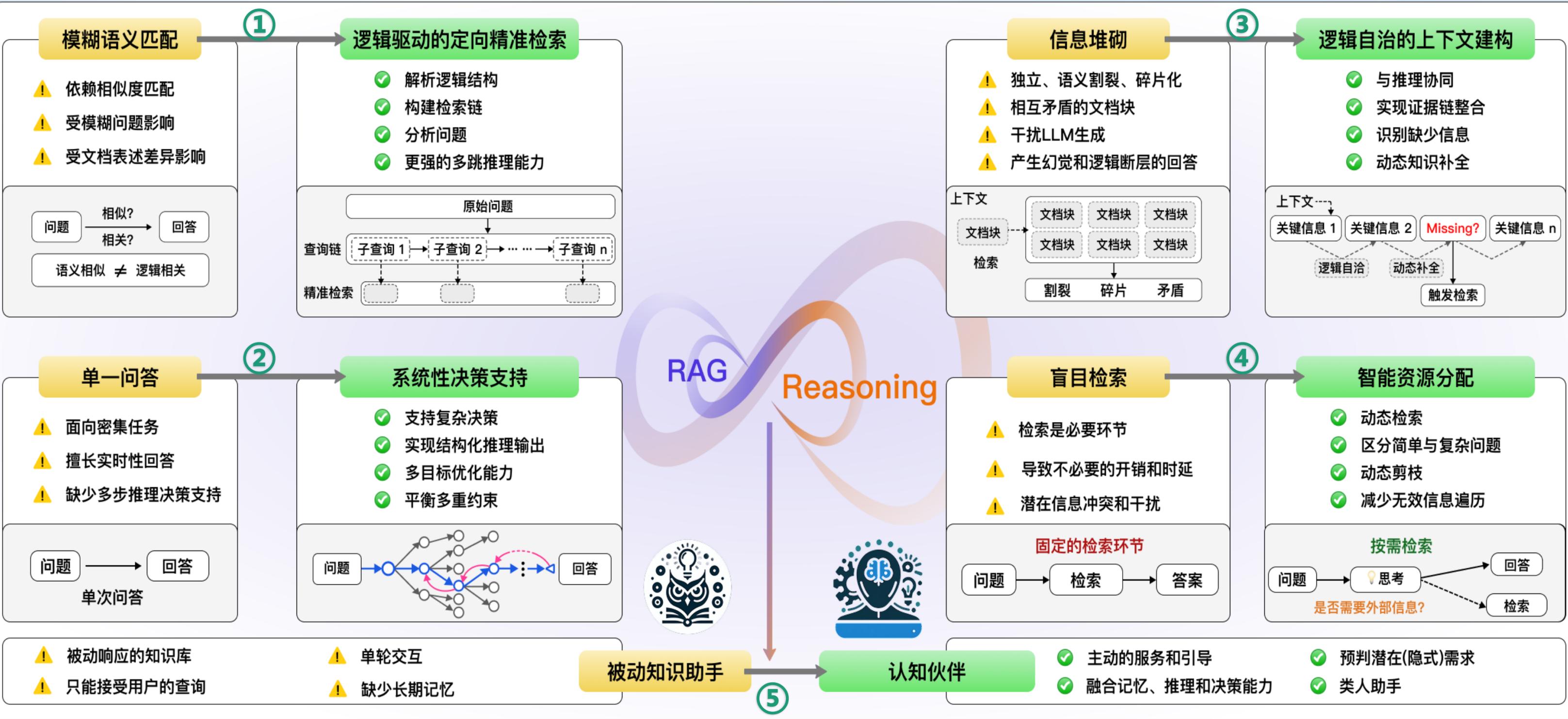


深度推理



联网搜索





THANKS

谢谢观看

欢迎关注、社区共建

Github: <https://github.com/OpenSPG/KAG-Thinker>

Github: <https://github.com/OpenSPG/KAG>

Github: <https://github.com/OpenSPG/openspg>

