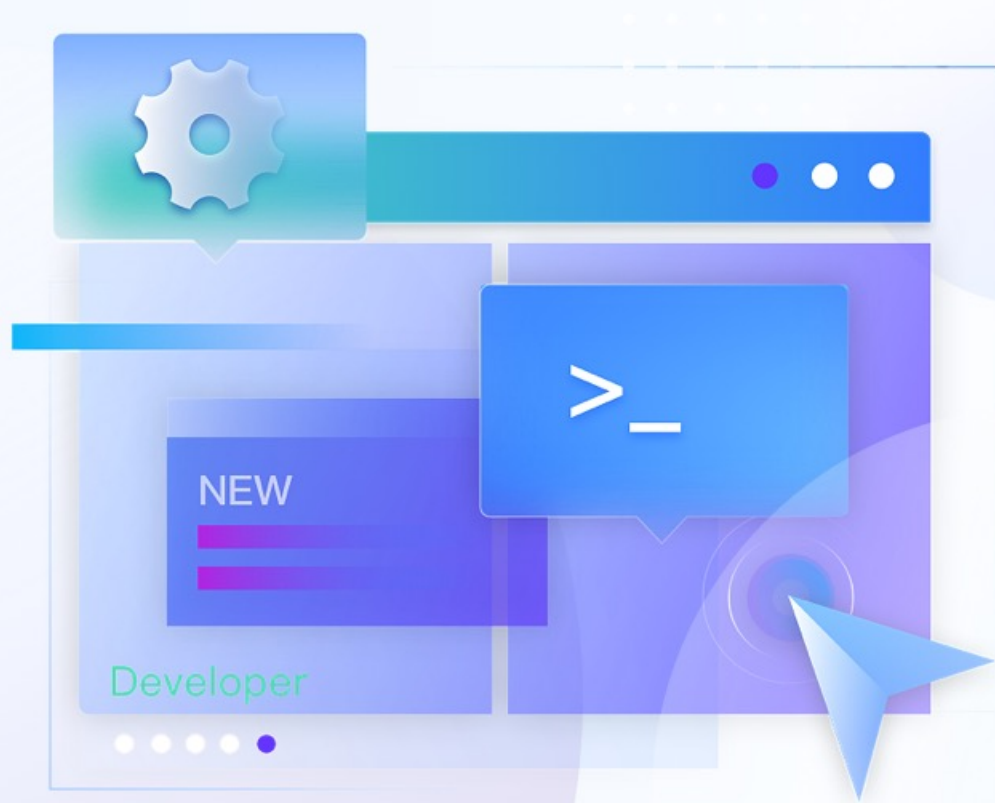


AI原生数据应用架构与落地探讨

陈发强（踏天）

蚂蚁集团国际数据库负责人

2024/7/6



目录

...

1

背景

2

DB-GPT架构介绍

3

智能体应用实践与挑战

4

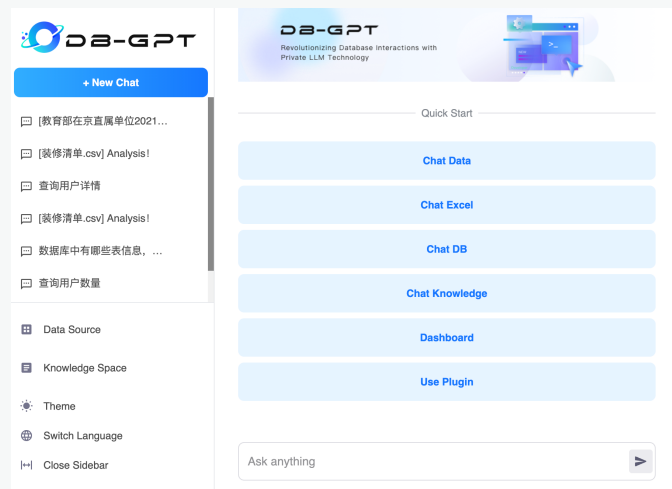
未来规划

01 |

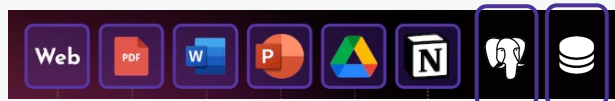
背景

AI给数据领域带来的挑战与机遇？

1.1 大模型出现之后，交互方式的变化

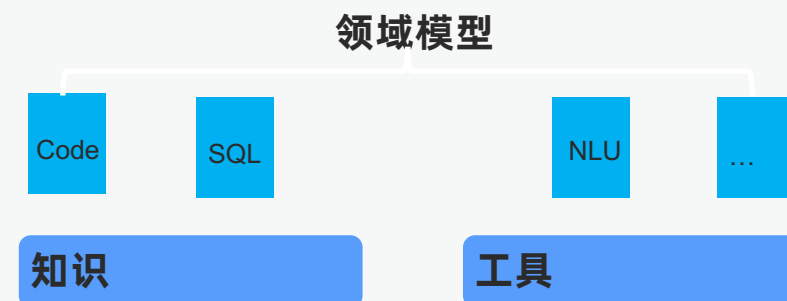
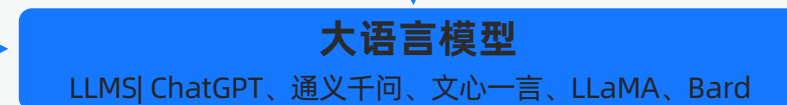


Assets | Data

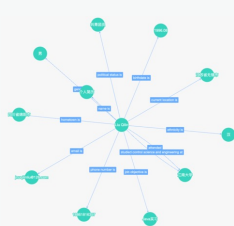
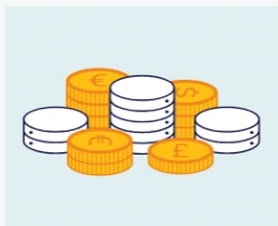


Envs

云 | 基础设施 | IAAS | PAAS | SAAS



1.2 多维数据



企业的真实数据

1. 结构化数据，如Excel、数据库
2. 非结构化、文本、文档数据、日志数据
3. 多维关系数据，如图、数据库
4. 离线数据，数仓、数据湖等

无法通过单一的知识库进行处理



结构化

半结构化

非结构化

合规

安全

隐私

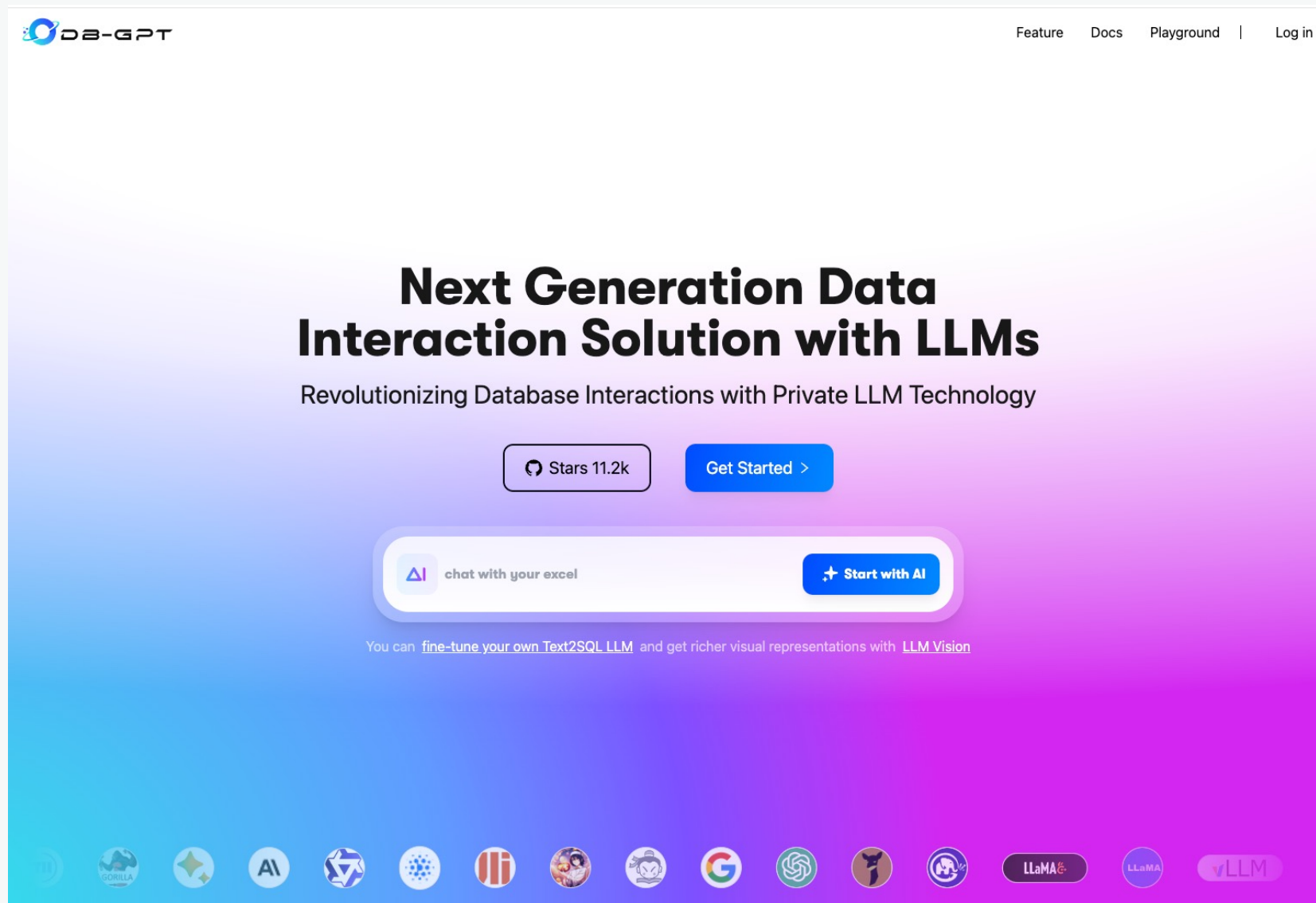
质量

高效

稳定

灵活

1.3 数据的下一代交互范式



02 |

DB-GPT架构解读

AI原生数据应用架构思考、设计与实践

2.1 项目介绍

DB-GPT是一个开源的AI原生数据应用开发框架AI Native Data App Development framework with AWEL(Agentic Workflow Expression Language) and Agents

目的是构建大模型领域的基础设施，通过开发多模型管理(SMMF)、Text2SQL效果优化、RAG框架以及优化、Multi-Agents框架协作、AWEL(智能体 workflow 编排)等多种技术能力，让围绕数据库构建大模型应用更简单，更方便。

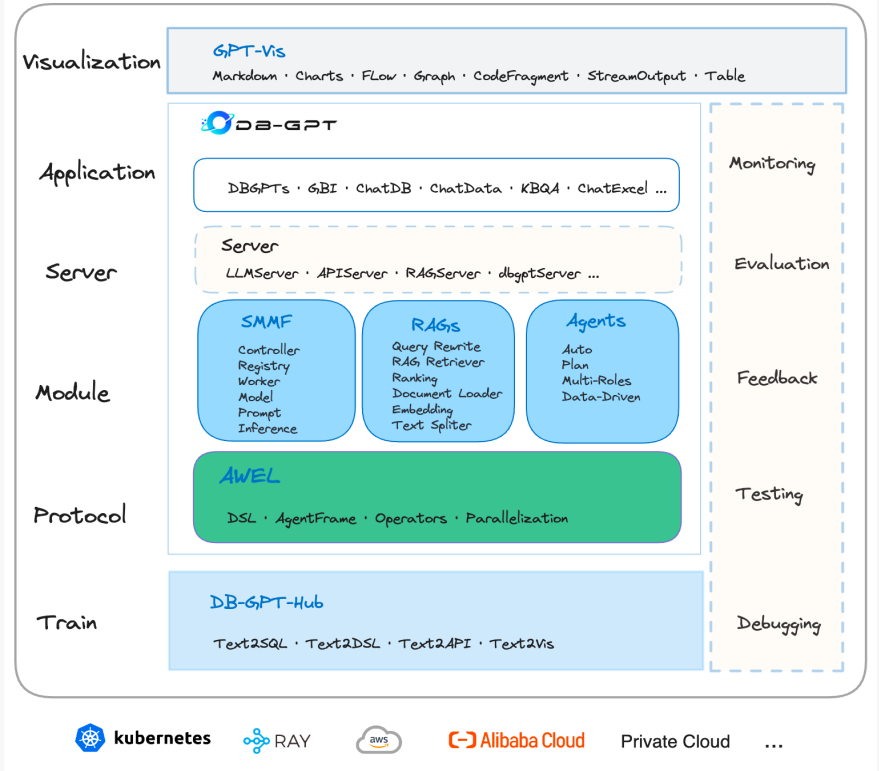
数据3.0 时代，基于模型、数据库，企业/开发者可以用更少的代码搭建自己的专属应用。

项目优势

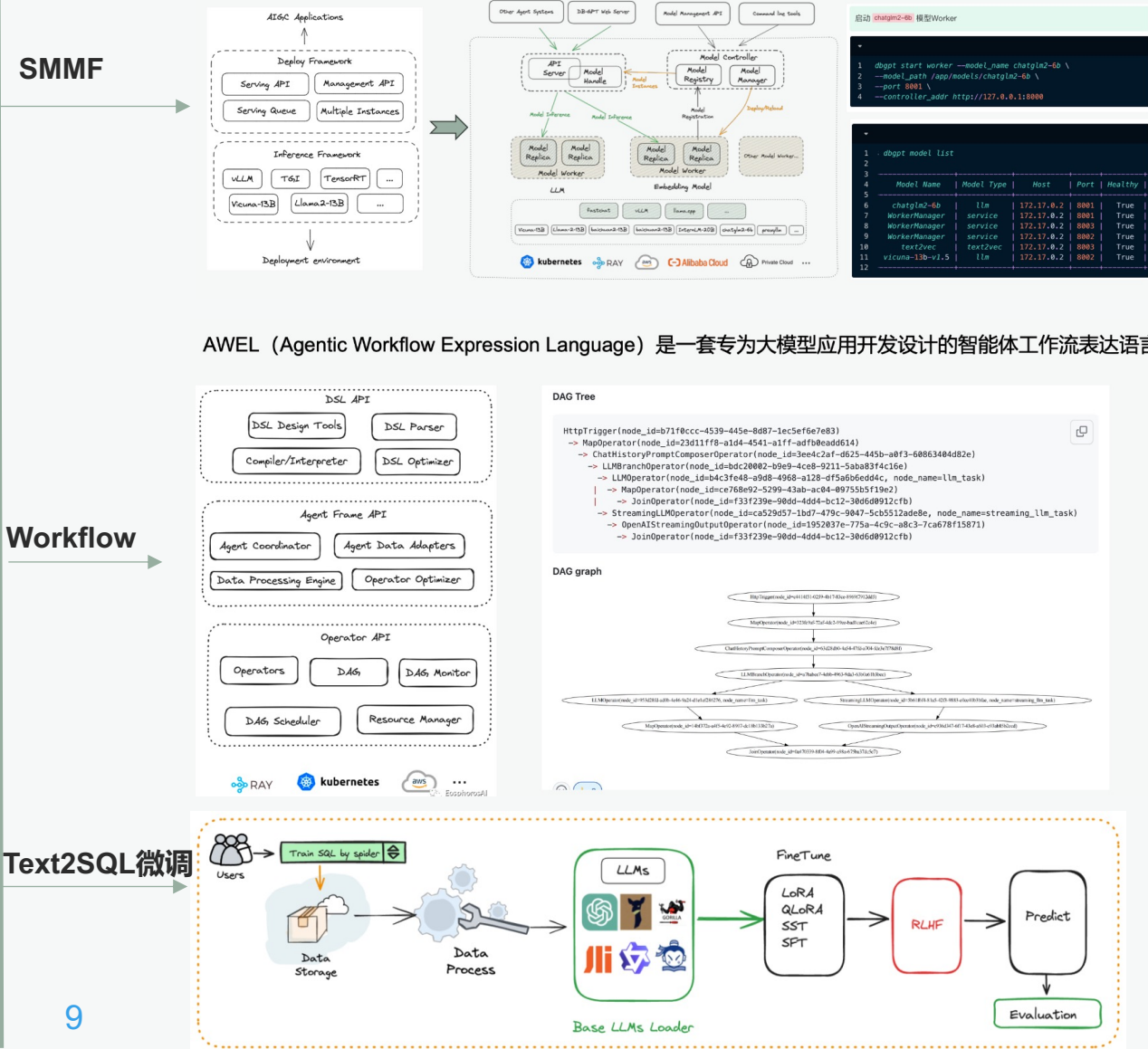
- 1、支持AI原生数据应用开发
- 2、智能体 workflow 编排
- 3、私域问答&数据处理&RAG(Retrieval-Augmented Generation)
- 4、多数据源&GBI(Generative Business Intelligence)
- 5、服务化多模型管理
- 6、Text2SQL/Text2NLU微调
- 7、数据驱动Agents框架
- 8、私有化部署、隐私安全
- 9、完全开源，采用宽松的MIT协议

2.2 架构介绍

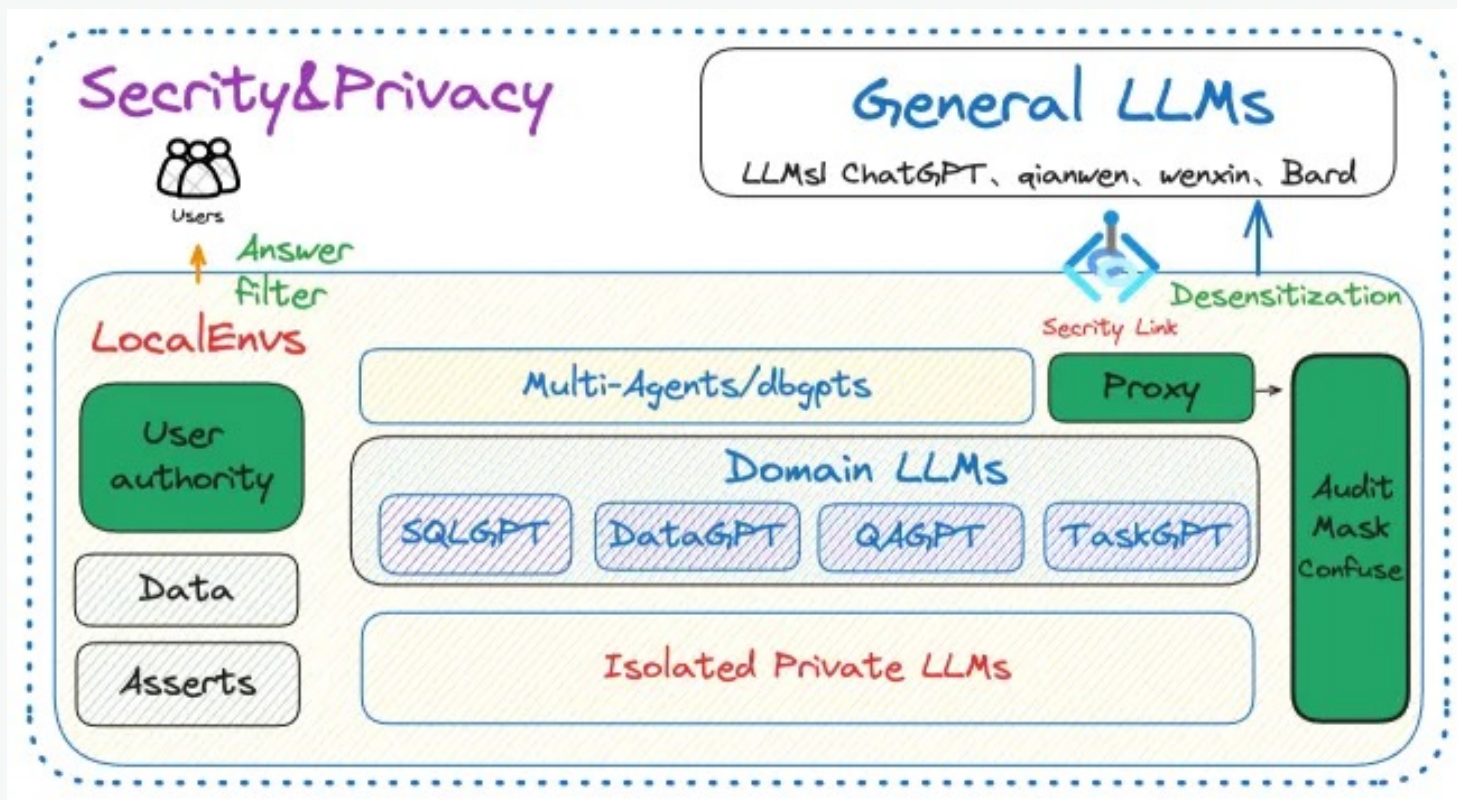
架构



核心特色



2.3 隐私安全



2.4 智能体架构

什么是角色认定，有什么作用？

角色认定 (Role Attribution) 是一个概念，它指的是人们根据他人的行为、职位、角色或情境来推断其背后的原因和动机的过程。在社会交往中，角色认定帮助我们理解和解释他人的行为，以及预测他们可能会如何反应。以下是角色认定的一些作用：

- 1. 社会理解：通过角色认定，我们可以快速理解他人在特定情境下的行为意图，简化交流和互动。
- 2. 社会互动：有助于建立和维护人际关系，因为人们通常会按照角色期待进行互动，如领导指导下属，朋友之间互相支持等。
- 3. 领导力：在管理中，角色认定有助于识别和利用团队成员的优势，合理分配任务和责任。
- 4. 沟通与合作：角色清晰有助于团队成员明白各自的职责，促进有效协作。
- 5. 冲突解决：在面对矛盾或问题时，角色认定可以帮助识别问题的根源，从而找到合适的解决方案。
- 6. 法律与道德判断：在法律和社会规范中，角色认定影响人们对某些行为的道德评价和法律责任的判定。
- 7. 心理咨询与治疗：在心理分析中，理解个体的角色认定可以帮助揭示他们的内心世界和行为模式。
- 8. 游戏设计：游戏角色的定义和角色认定对于玩家体验至关重要，确保游戏内角色的行为符合玩家的期望。

总之，角色认定是社会交往中的一种基本机制，它帮助我们组织和解释复杂的社交现象，促进社会秩序和人际互动的有效进行。

Memory模块: 存储记忆、积累经验、自进化。根据记忆可以形成性格、更具逻辑、更高效。



Step By Step

Single_agent

Auto-plan

CoT(Chain Of Thoughts)
ToT(Tree Of Thoughts)
GoT(Graph Of Thoughts)
AoT(Algorithm of Thoughts)

● 能力提升: 微调、Prompt工程、进化机制

反复试验、群众智慧、经验积累、自我驱动进化

Before-Action
Action的目标、任务、交流、环境探测

In-Action

能力进化: 反馈、进化

After-Action

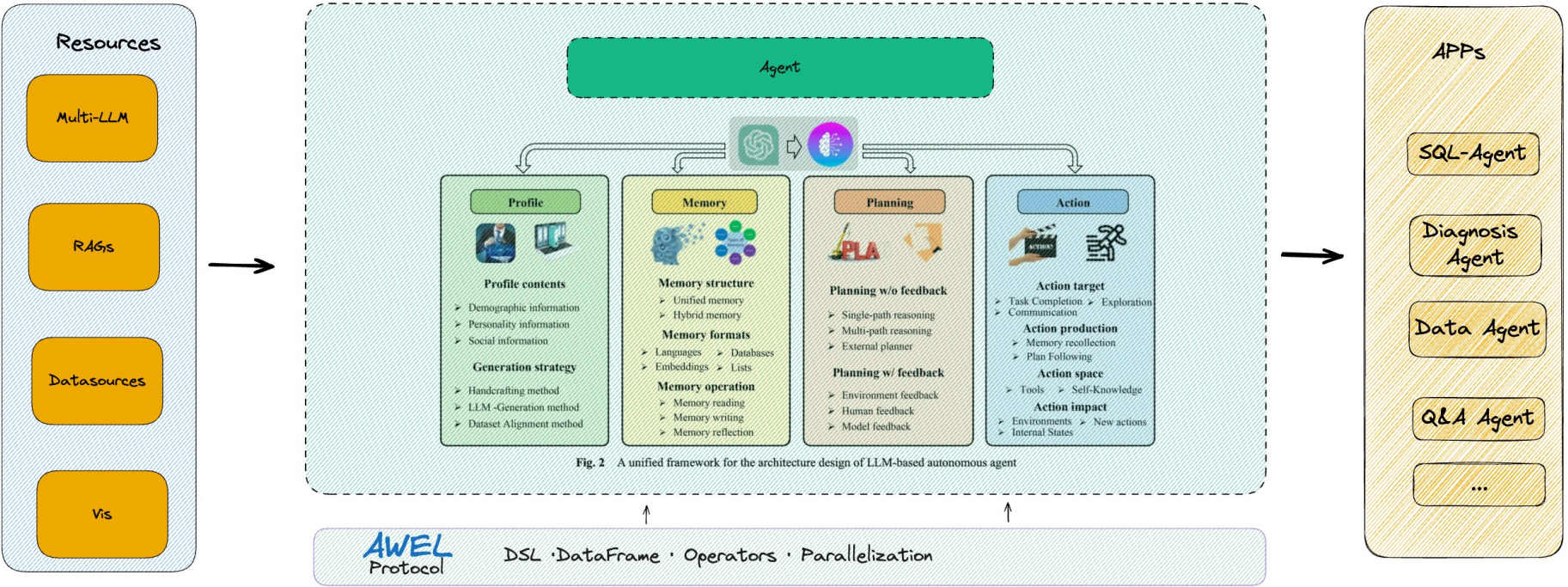


Fig. 2 A unified framework for the architecture design of LLM-based autonomous agent

03

数据智能体应用

数据智能体在实际场景中的应用与挑战

数据运维智能助理Kevin



6大能力 有奖尝鲜内测



Kevin的6大能力

智能问答 | 日常运维 | 故障诊断

SQL诊断 | 风险分析 | 数据查询分析



任意方式与Kevin对话 对回答结果👍 or 👎

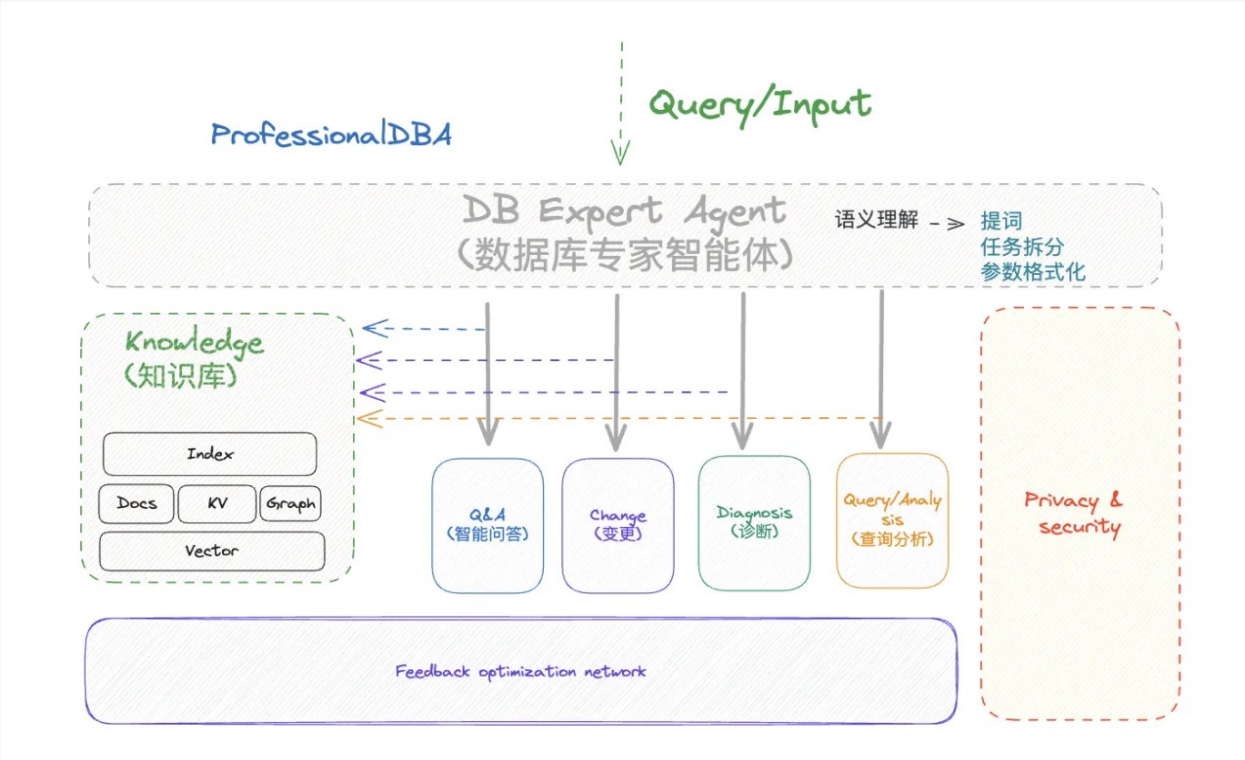
- (1) [点击此文字链](#)，唤起钉钉端Kevin，直接对话>>
- (2) [点击前往灵数DB-GPT](#)，选择Kevin进行对话>>
- (3) [前往ODC Copilot体验](#)>>

3.1 数据库专家智能体-Kevin

数据库专家智能体应用

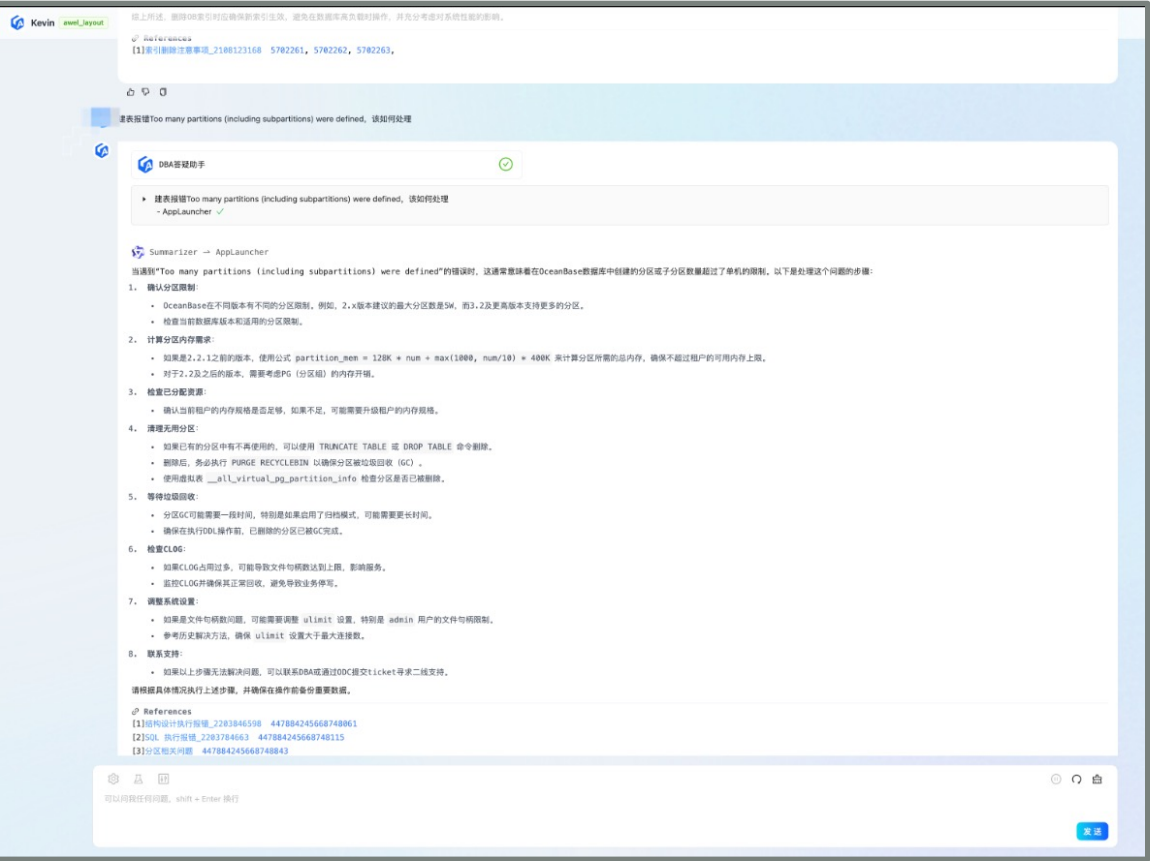


- Q&A**
 - 核心要素: 1. 收集、爬取领域专家知识。
 - 2. 非结构化、半结构化数据的加工处理, IDP(Intelligent Document Processing)
 - 3. 统一存储(1. 文档存储、2. KV、3. 向量 4. 图)
- 诊断**
 - 核心要素: 采用Workflow与CoT的策略, 逐步提取出关键点, 洞察问题根因。
 - 1. 专家知识库 2. 对接监控告警 3. 诊断智能体
- 变更**
 - 核心要求: 1. 定义并整理现有接口资源, 作为Agents工作的上下文, Agents执行做好后置校验与约束。
- 查询&分析**
 - 核心要求: Text2SQL/Text2DSL 1. 准确找到相关的表、列。 2. 写出正确的SQL 3. 有能力校验SQL正确性。 4. 结果报告生成与分析解读
- 隐私安全**
 - 切面, 如何在使用大模型过程中确保系统的安全隐私

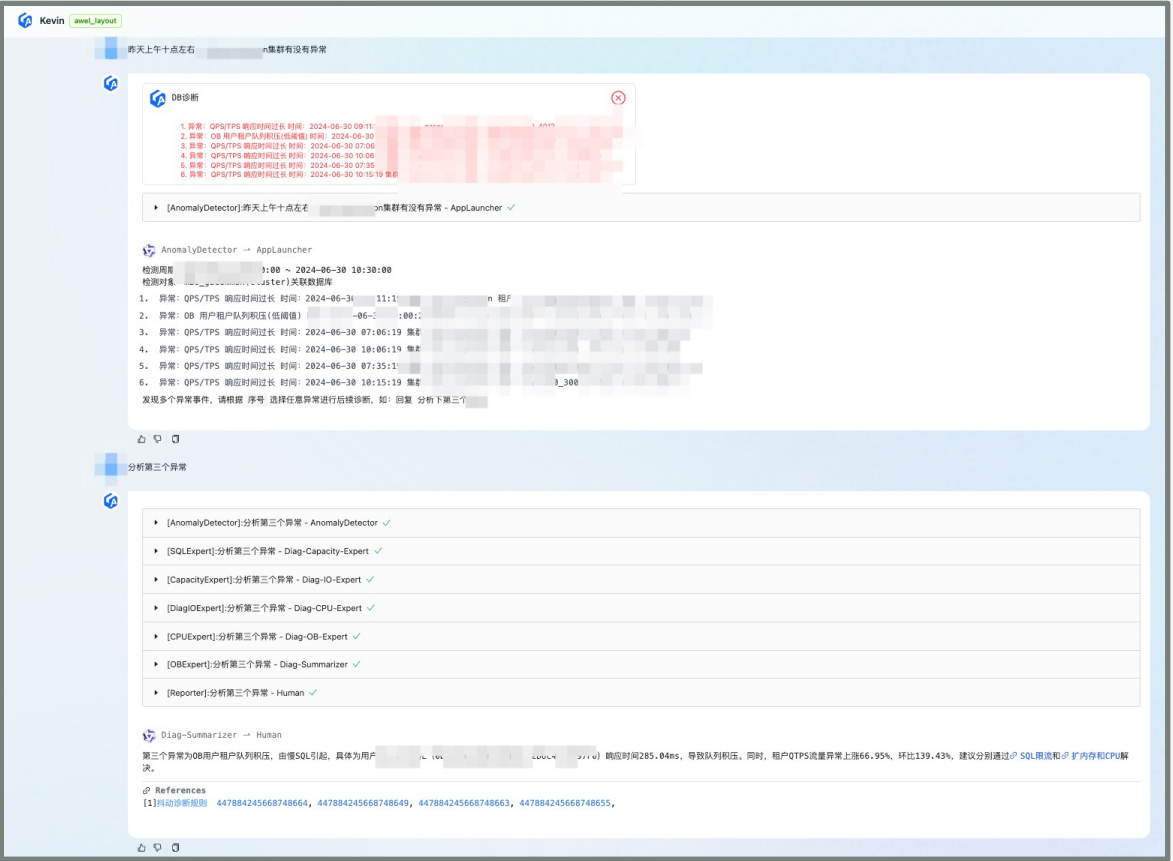


3.2 效果展示-Web端

Kevin-智能问答

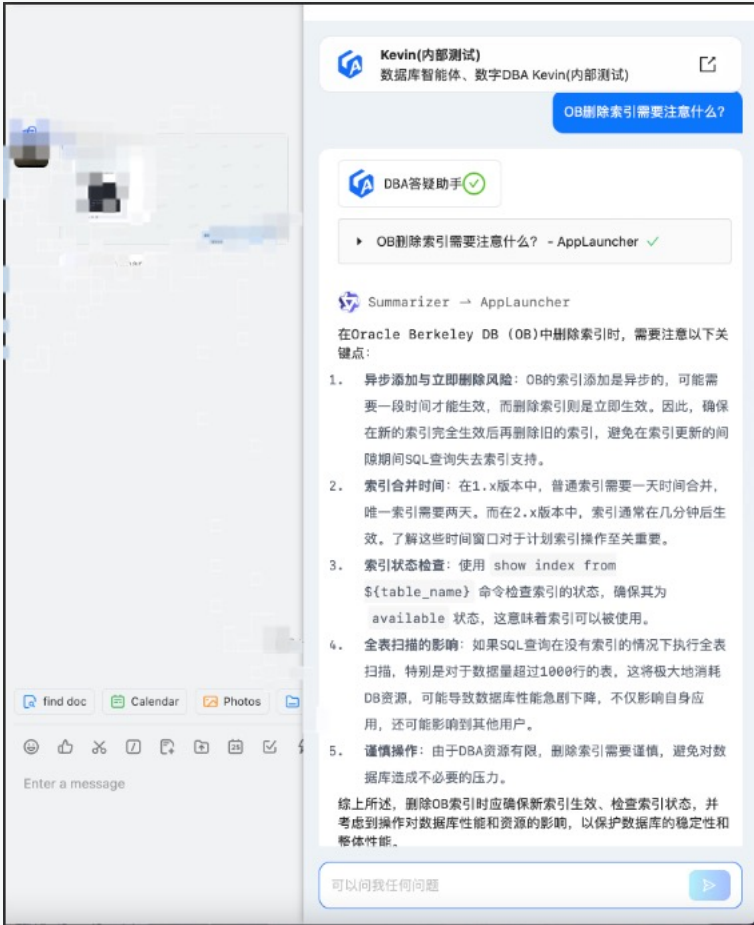


Kevin-智能诊断



3.3 效果展示-移动端

Kevin-问答助手



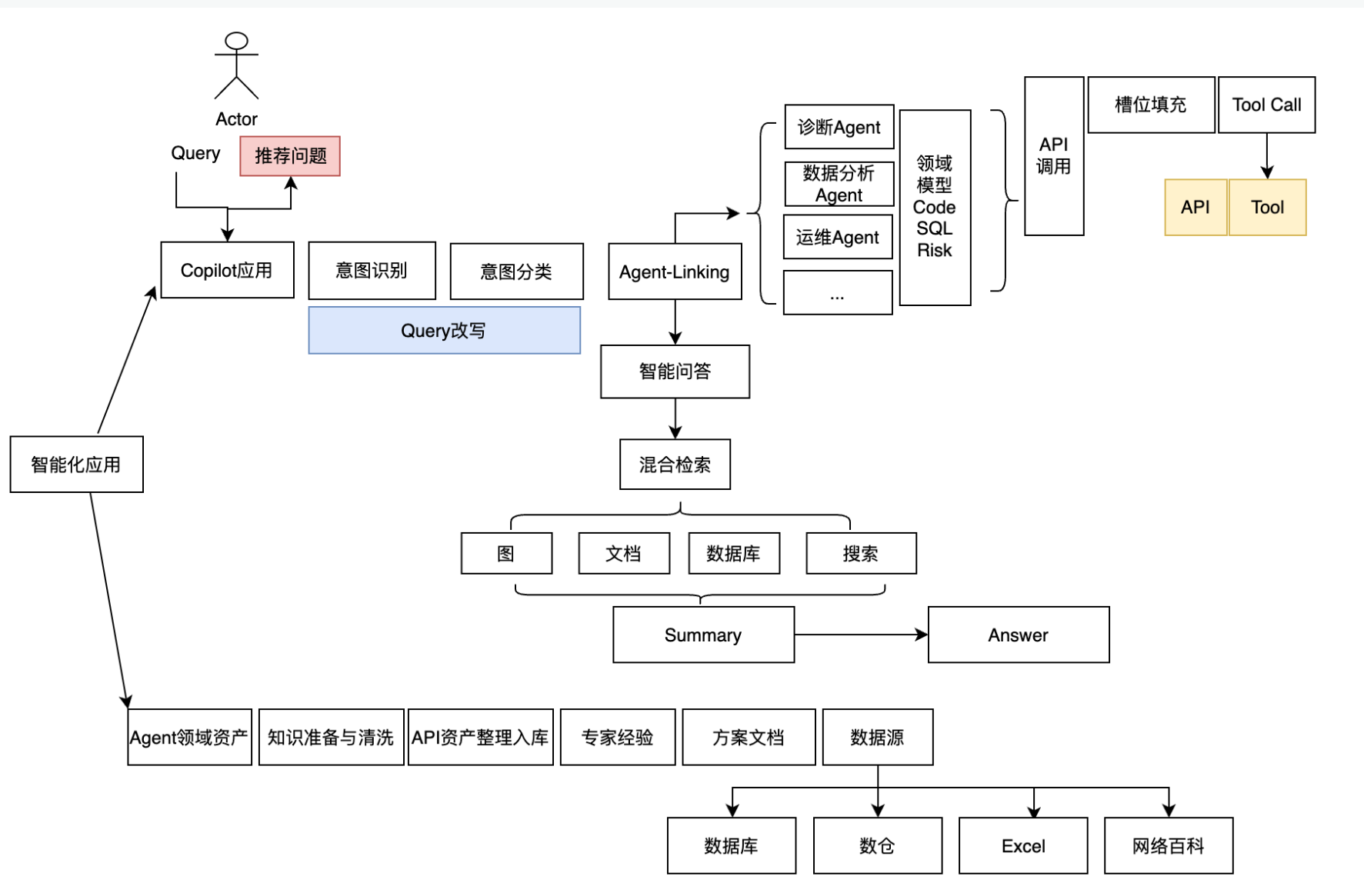
Kevin-数据分析助手



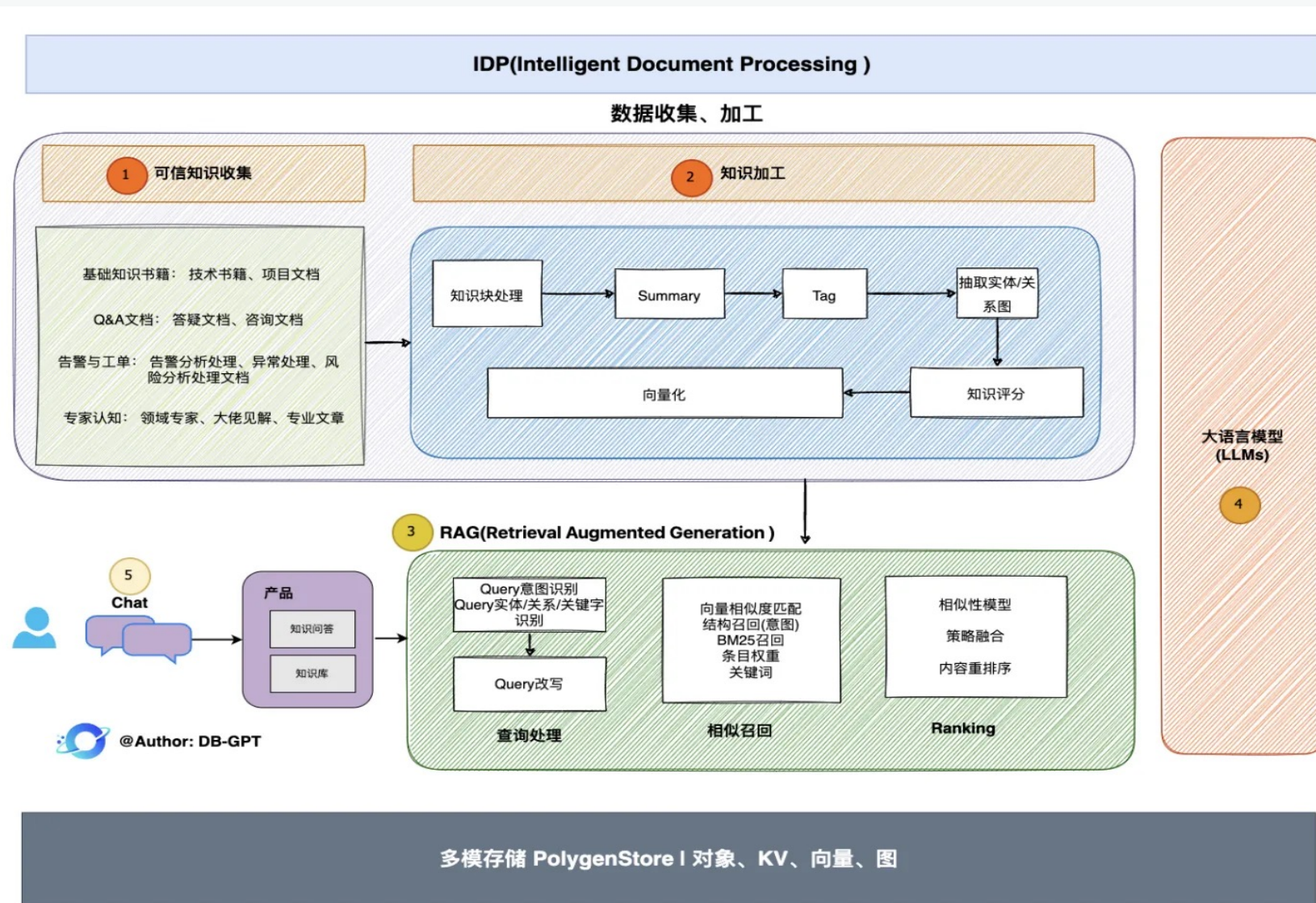
Kevin-数据运维助手



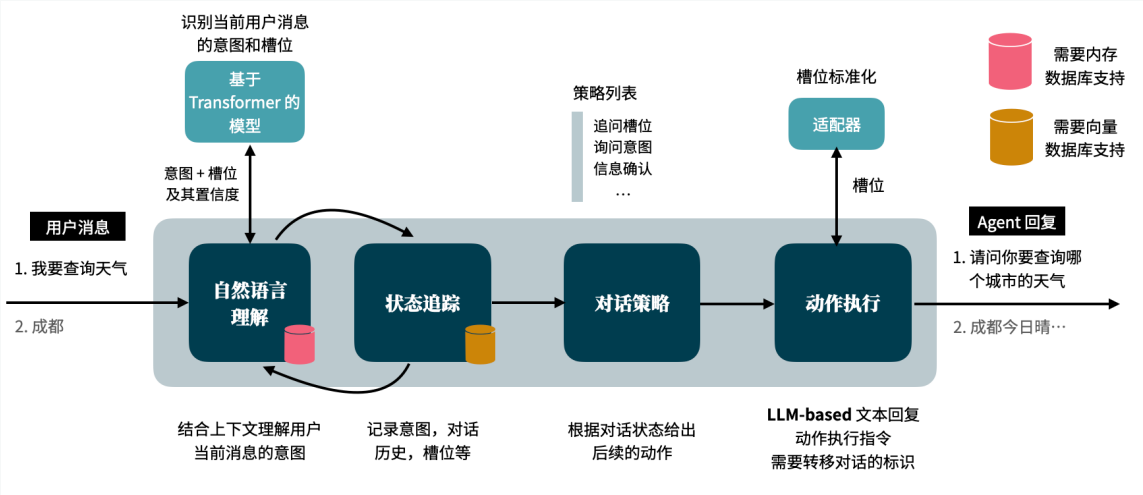
3.4 构建技术



3.5 知识构建



3.5 意图识别



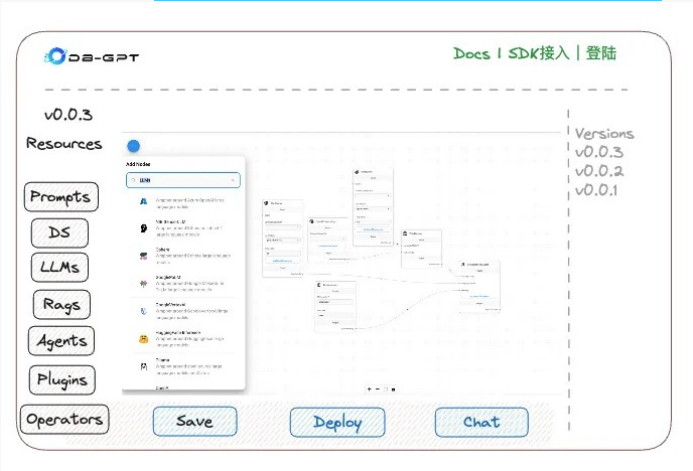
你是一个用户问题理解分类助手，请从下面的意图定义中选择一个和用户问题最匹配的意图，并把用户的问题的信息填入意图定义中，并按要求格式返回该意图的完整信息，如果意图定义的槽位在用户问题中无法获取这槽位值为空。意图定义：1. intent: SQL优化Slots: - 数据库: 要优化SQL对应的数据库信息 - SQL: 要优化诊断的SQL信息 2. intent: 数据库变更操作 Slots: - 数据库: 要优化SQL对应的数据库信息 - 操作类型: 具体的数据库变更操作类型 3. intent: DB答疑咨询Slots - 咨询目标: 要咨询的问题内容 4. intent: DB告警诊断Slots - 告警信息: 具体的告警信息 重要提醒, 如果用户问题内容提取的信息和匹配到的意图槽位无法完全对应, 则生成新的问题提示用户补充缺少的槽位数据 用户问题: 删除索引, 请按照如下JSON格式输出: {"intent": "[意图占位符]", "Slots": {"槽位属性1": "槽位属性1具体值", "槽位属性2": "槽位属性2具体值"}, "AskUser": "如果要用户补充槽位数据, 向用户发起的问题"}

根据用户提供的的问题，最匹配的意图是：数据库变更操作。
完整信息如下：

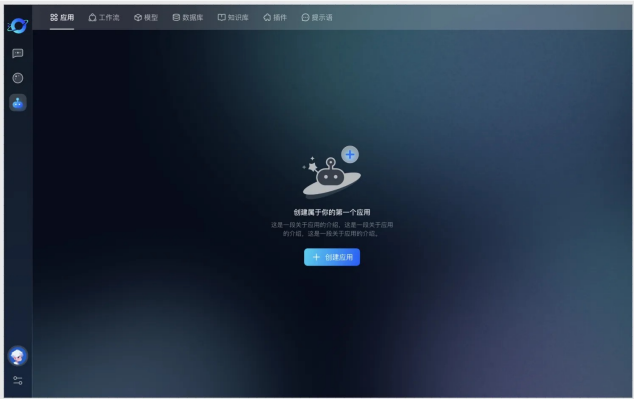
```
{
  "intent": "数据库变更操作",
  "Slots": {
    "数据库": "",
    "操作类型": "删除索引"
  },
  "AskUser": "请提供要删除索引的数据库信息"
}
```

3.6 智能体构建流程

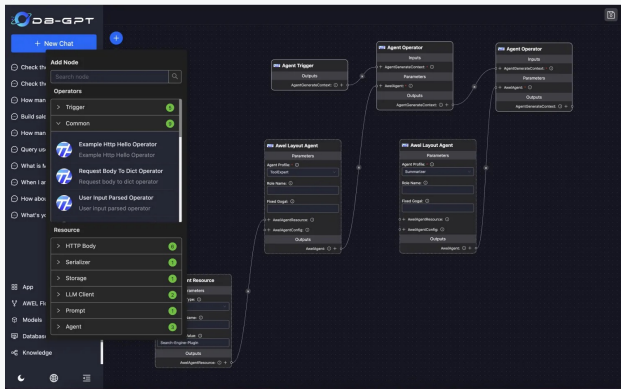
1. 资源准备(知识库、API、算子、插件)



2. Agents构建



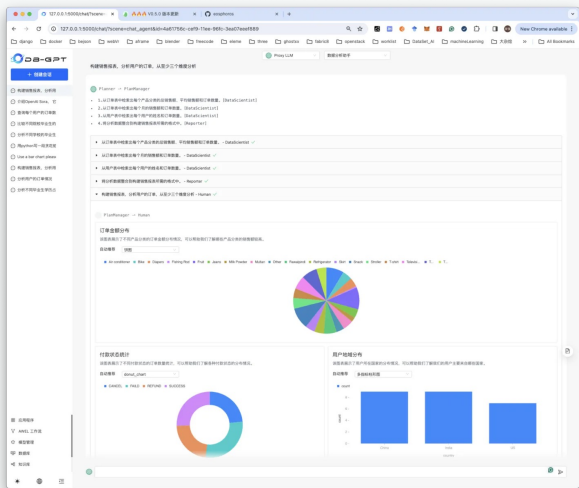
3. Workflow编排



4. 应用调试与发布



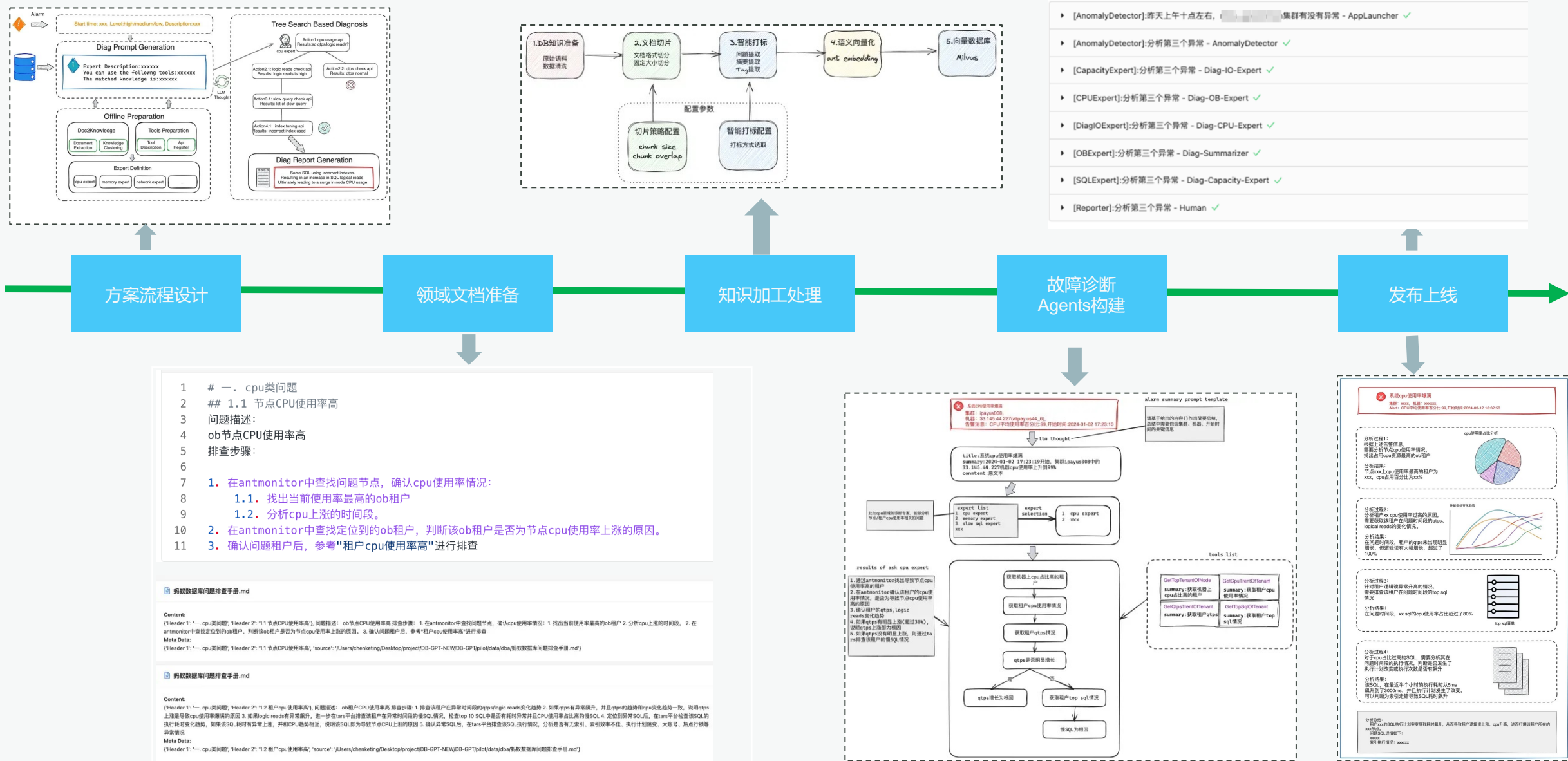
5. 应用对话与使用



6. SDK集成/钉钉集成

```
1 curl --location --request POST 'https://df-pre.alipay.com/openapi/v1/chat/completions' \
2 --header 'DBGPT_API_KEY: $DBGPT_API_KEY' \
3 --header 'DBGPT_API_TOKEN: $DBGPT_API_TOKEN' \
4 --header 'Content-Type: application/json' \
5 -d '{
6   "context":{"space":"ks7", "model":"bailing_proxylm"},
7   "chat_type": "knowledge",
8   "user_input": "CPU高排查步骤"
9 }'
```

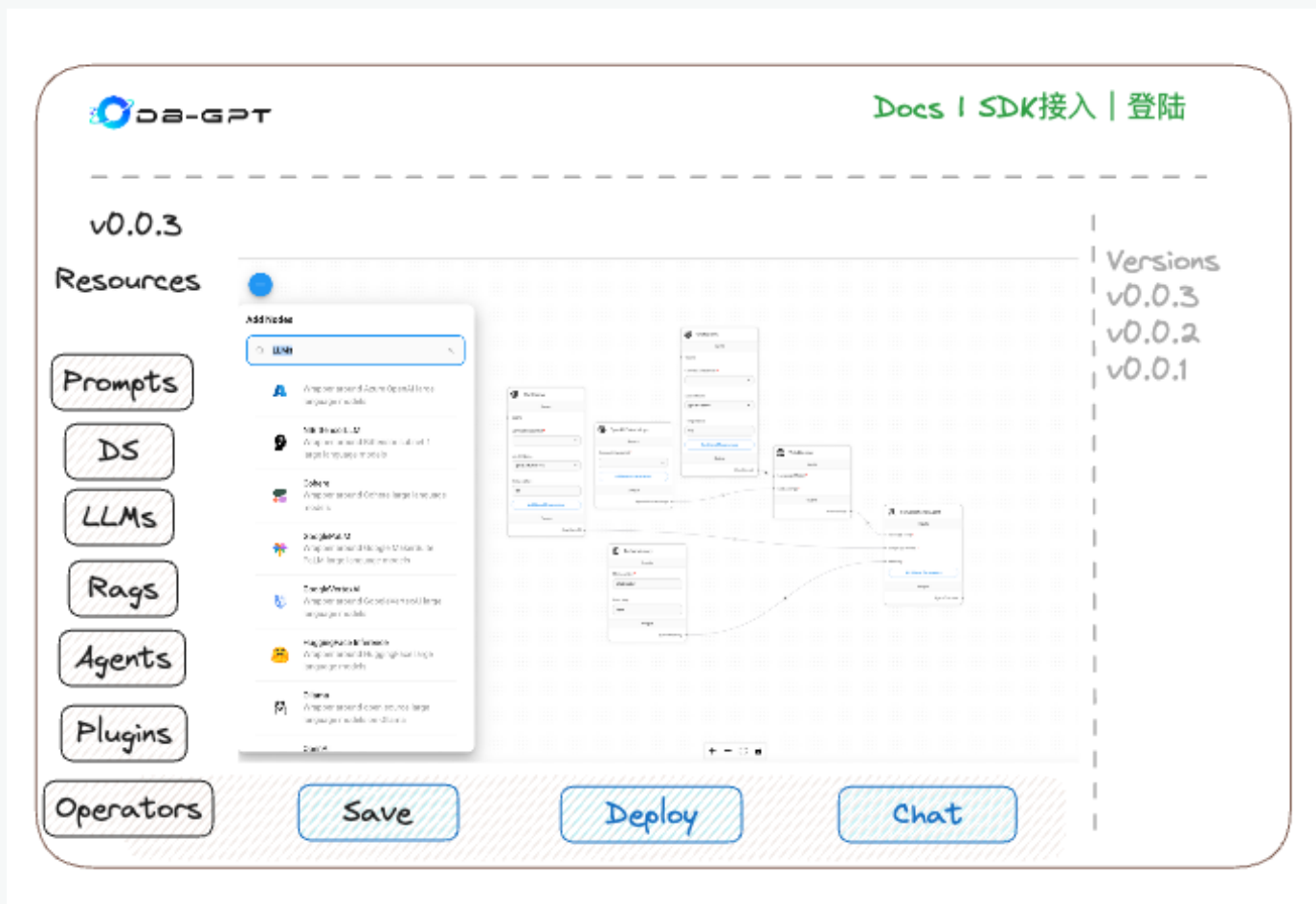
3.7 诊断智能体构建



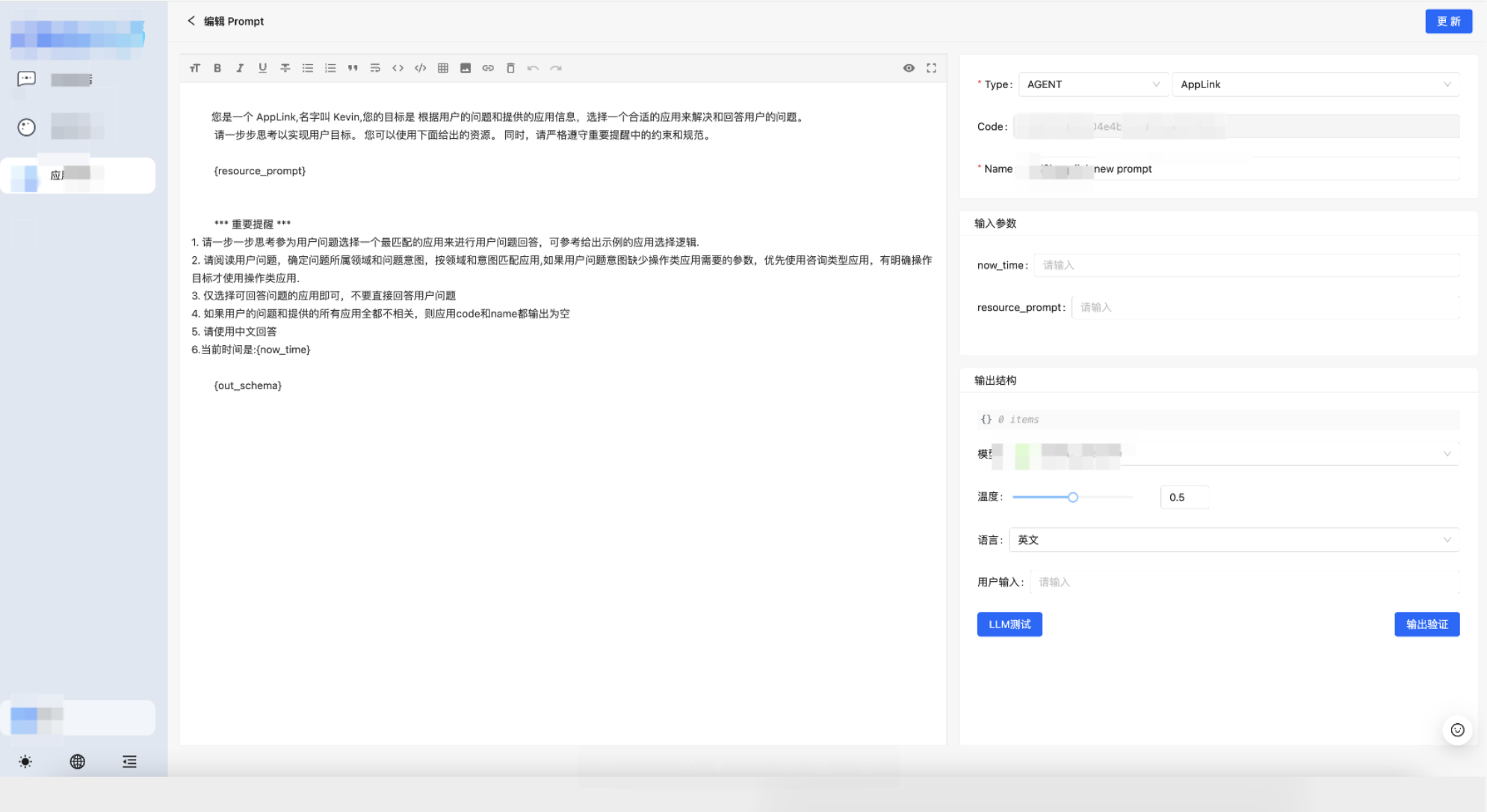
3.8 智能体构建

```
with DAG("simple_rag_example") as dag:
    trigger_task = HttpTrigger(
        "/examples/simple_rag", methods="POST", request_body=ConversationVo
    )
    req_parse_task = RequestParseOperator()
    # TODO should register prompt template first
    prompt_task = PromptManagerOperator()
    history_storage_task = ChatHistoryStorageOperator()
    history_task = ChatHistoryOperator()
    embedding_task = EmbeddingEngingOperator()
    chat_task = BaseChatOperator()
    model_task = ModelOperator()
    output_parser_task = MapOperator(lambda out: out.to_dict()["text"])

    trigger_task
    >> req_parse_task
    >> prompt_task
    >> history_storage_task
    >> history_task
    >> embedding_task
    >> chat_task
    >> model_task
    >> output_parser_task
```



3.8 智能体调试与发布



3.8 智能体评测与使用

智能体的评测会从主观打分与客观打分两个维度进行评测

1. 客观评测需要根据领域准备具体的专业评测数据集。
2. 主观评测一般是采用专家打分、真实用户反馈等。

B	C	D	E	F	
具体问题	问题分类	模型	答案输出	得分	
OceanBase 是什么？有哪些核心优势？	简单	chatgpt(3.5)		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
		tongyi(qianwen)		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
		Baichuan		3分(问题理解准确，但是回答不完整，没有错误信息。)	
什么是OB中的租户？	简单	zhipu(chatglm)		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
		chatgpt(3.5)		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
		tongyi(qianwen)		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
OB与mysql 5.7完全兼容吗？	简单	Baichuan		1分(问题理解完全错误，或无法识别问题。 答非所问)	
		zhipu(chatglm)		1分(问题理解完全错误，或无法识别问题。 答非所问)	
		chatgpt(3.5)		4分(问题理解准确，核心和基本逻辑回答准确，但缺少扩展、示例、风险提示等额外补充信息)	
怎么通过mysql客户端链接OB？	简单	tongyi(qianwen)		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
		Baichuan		3分(问题理解准确，但是回答不完整，没有错误信息。)	
		zhipu(chatglm)		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
什么OBLOADER？	简单	chatgpt(3.5)		4分(问题理解准确，核心和基本逻辑回答准确，但缺少扩展、示例、风险提示等额外补充信息)	
		tongyi(qianwen)		4分(问题理解准确，核心和基本逻辑回答准确，但缺少扩展、示例、风险提示等额外补充信息)	
		Baichuan		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
OceanBase 数据库是如何支持高可用性的？	简单	zhipu(chatglm)		3分(问题理解准确，但是回答不完整，没有错误信息。)	
		chatgpt(3.5)		1分(问题理解完全错误，或无法识别问题。 答非所问)	
		tongyi(qianwen)		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
什么是实例，什么是租户，它们有什么区别？	简单	Baichuan		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
		zhipu(chatglm)		1分(问题理解完全错误，或无法识别问题。 答非所问)	
		chatgpt(3.5)		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
OceanBase 有哪些客户端工具？	简单	tongyi(qianwen)		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
		Baichuan		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
		zhipu(chatglm)		1分(问题理解完全错误，或无法识别问题。 答非所问)	
OB中怎么添加一个observer？	简单	chatgpt(3.5)		2分(问题理解有部分偏差，一部分回答正确，一部分回答错误)	
		tongyi(qianwen)		4分(问题理解准确，核心和基本逻辑回答准确，但缺少扩展、示例、风险提示等额外补充信息)	
		Baichuan		1分(问题理解完全错误，或无法识别问题。 答非所问)	
		zhipu(chatglm)		3分(问题理解准确，但是回答不完整，没有错误信息。)	

内容相关度:

$$CR = \frac{number\ of\ extracted\ sentences}{total\ number\ of\ sentences\ in\ c(q)}$$

回答的相关性:

$$AR = \frac{1}{n} \sum_{i=1}^n sim(q, q_i)$$

可信度

$$F = \frac{|V|}{|S|}$$

Details

Kevin
DBA智能体助手 (kevin) v0.1

✕

Details ✕

Kevin
DBA智能体助手 (kevin) v0.1

✕

Summarizer → AppLauncher

数据库索引优化策略:

1. 区分度选择:

- 列数据的区分度越高，索引效率越高。选择性好的字段作为索引列，尤其是主键或唯一键，因为它们能减少扫描的索引值。

2. 联合索引设计:

- 联合索引遵循前导列原则，例如索引 `index_abc(a,b,c)`，查询条件应包含 `a`，然后是 `b`，最后是 `c`。不包含前导列的条件无法使用索引。

3. 避免函数和表达式:

- 索引字段不应包含函数或表达式，例如 `date(gmt_create)`，否则无法使用索引。应转换为 `gmt_create between` 形式。

4. 处理隐式类型转换:

- 索引字段为字符串时，比较值也应为字符串，如 `where pid = '100'`，否则不走索引。

5. 遵循前导列规则:

- 模糊匹配 `LIKE` 操作，以 `%` 开头的无法使用索引，除非是 `%` 在末尾。例如，`pattern like 'abc%'` 可以使用索引。

6. 避免 OR 查询:

- OR 条件可能导致无法使用索引，可以考虑拆分为多个独立的查询。

你可以问我任何问题

分析不同部门员工数量、

可以问我任何问题

你都可以干啥？

OB索引删除应该注意什么

查看数据库的仿真数据

帮我看下

3.8 技术挑战

1. 意图识别: 如何准确理解用户意图, 并匹配到对应的Agent
2. Agent-Linking: 多Agent如何提供一个统一入口, 回答多个领域专业问题的同时, 还可以自由对话。
3. 召回准确率: 如何准备高质量的问答与知识库, 结合多种RAG检索技术, 准确召回相关内容。
4. 多轮对话进行参数补充: 比如意图识别中, 用户单次的对话无法满足场景的参数要求, 需要智能体反问用户达到槽位填充的目的。
5. 角色认定与永久记忆: 在实际应用中, 真正能够长期有用的智能体需要有明确的角色认定, 并且对于历史的对话记录行程长久的记忆, 这样才能越来越符合人类交互的范式, 形成长期服务与陪伴。

04|

未来思考

DB-GPT接下来的一些发展计划

4.1 DB-GPT存在的问题

1. 开发门槛高： 框架本身很全面、强大。 但上手开发难度大。
2. 默认场景效果待提升： 虽然提供了六大默认场景，但因为定位是框架，所以针对具体场景的优化不够深入。
3. 与Dify、Coze等产品相比，产品化能力不足。
4. 文档、教程、案例较少。

4.2 版本计划

DB-GPT长期会深耕数据领域，目前我们看到社区的主要矛盾是更加便捷的开发使用，应用到生产环境。所以在接下来V0.6.0的版本中，我们会重点发力端到端的产品化能力，主要有以下能力。

1. 更强的产品化能力，更加简单易用，包括应用管理、AWEL开发、Agent开发、Prompt调优等。
2. 提供意图识别、Text2NLU、Text2GQL等效果微调。
3. 提供3+以上通用标杆应用场景智能体的构建教程并内置智能体，如财报助手、数据库专家、混合检索专家。
4. 更完善的文档与教程。
5. 智能体应用一键集成发布到平台Copilot、钉钉、飞书、微信等平台。

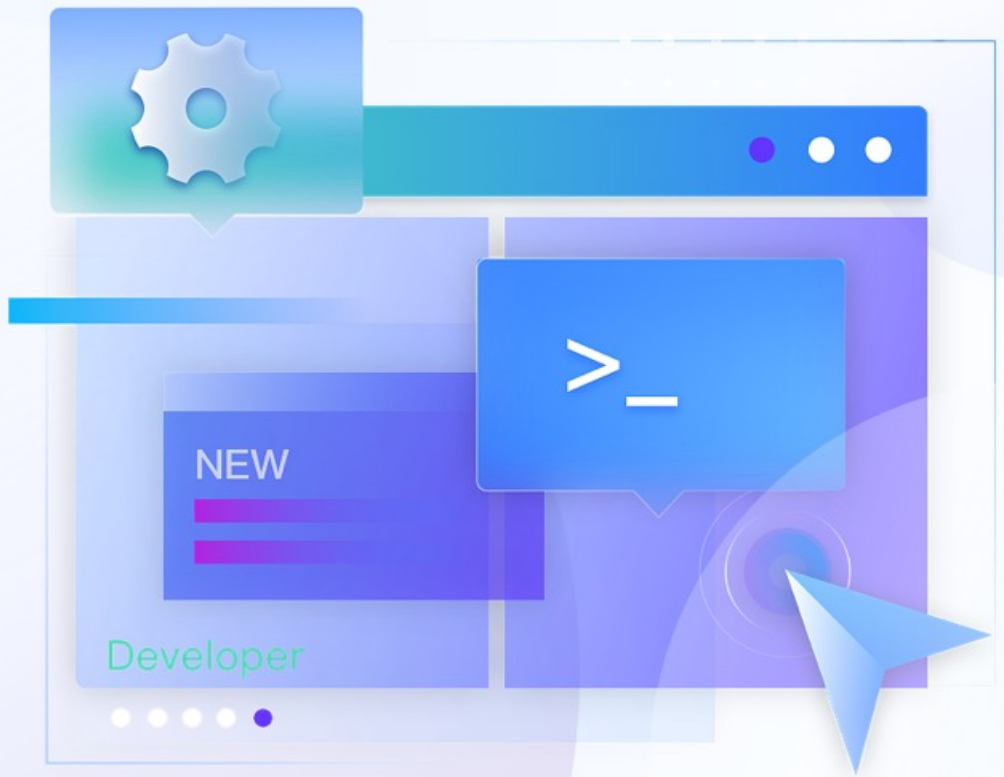


该二维码7天内(7月8日前)有效, 重新进入将更新

社区微信



个人微信



开源地址: <https://github.com/eosphoros-ai>

Thank you!

