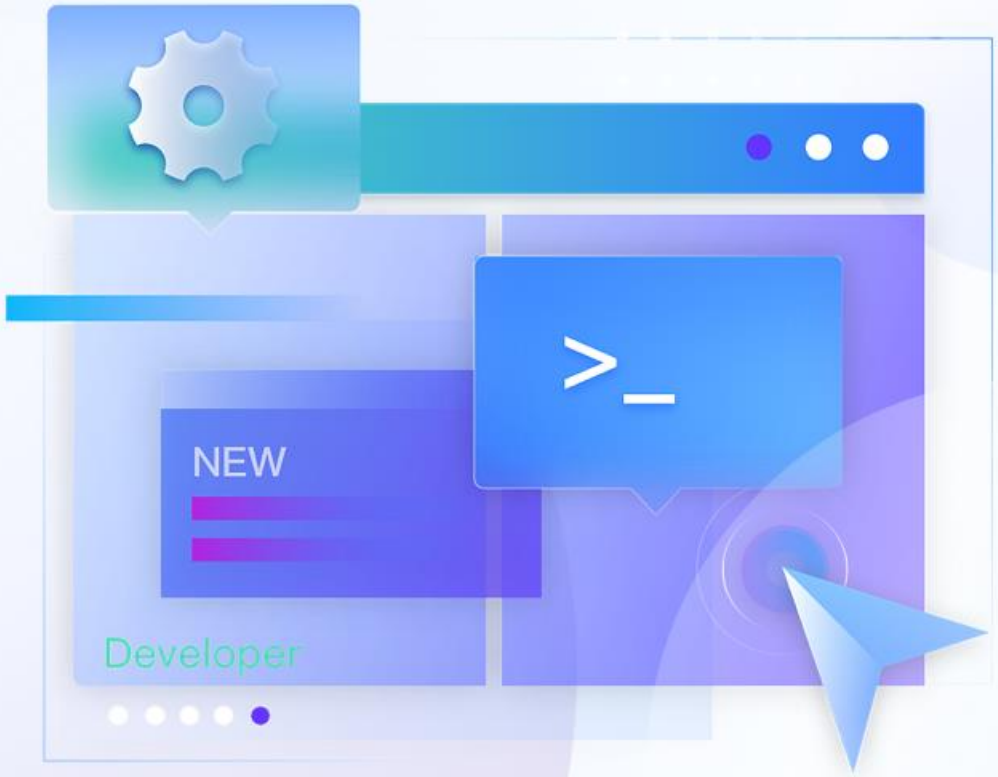
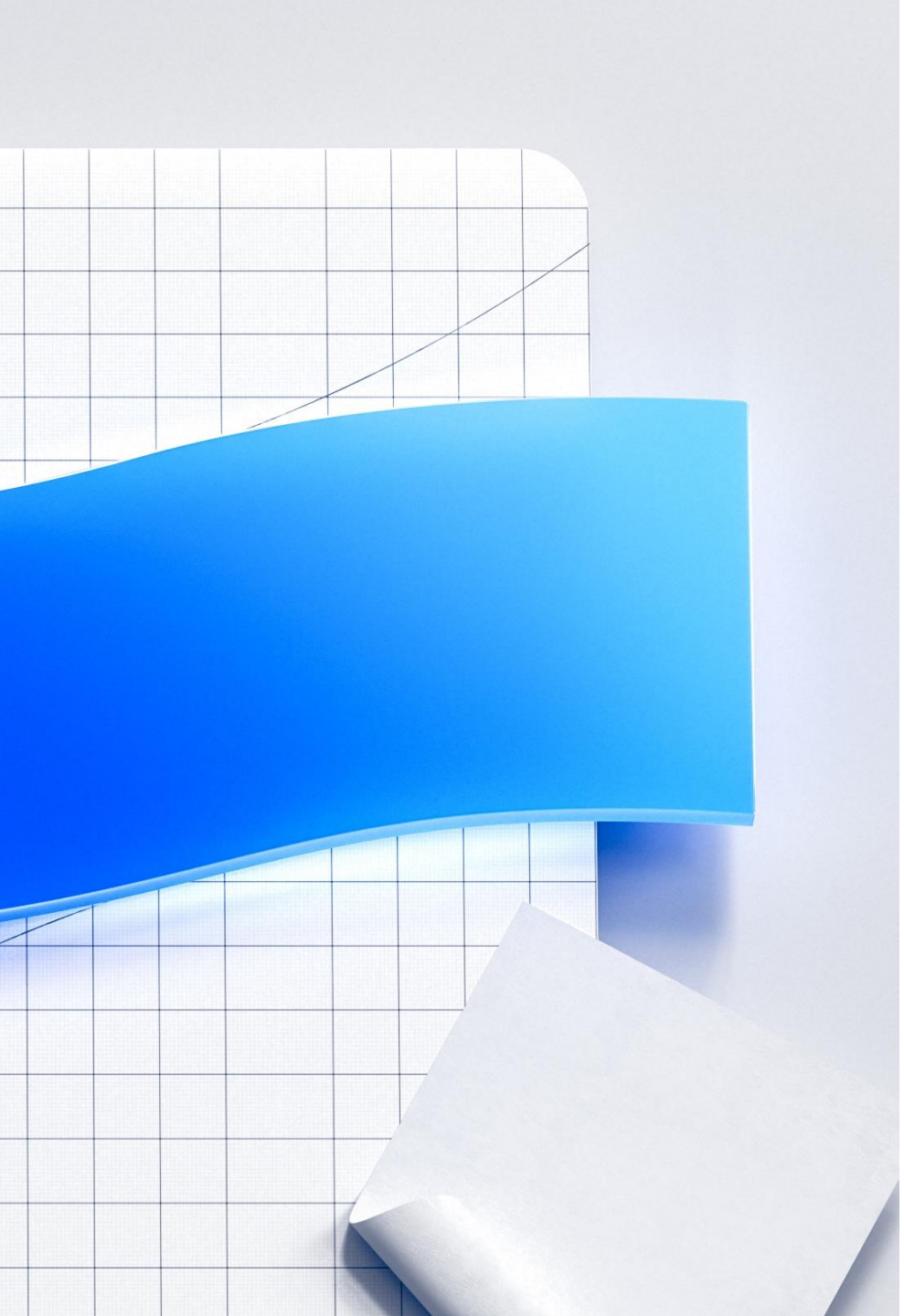
DB-GPT 结合 Graph RAG 落地探索与实践

蚂蚁数据部-数据基础设施-柯廷



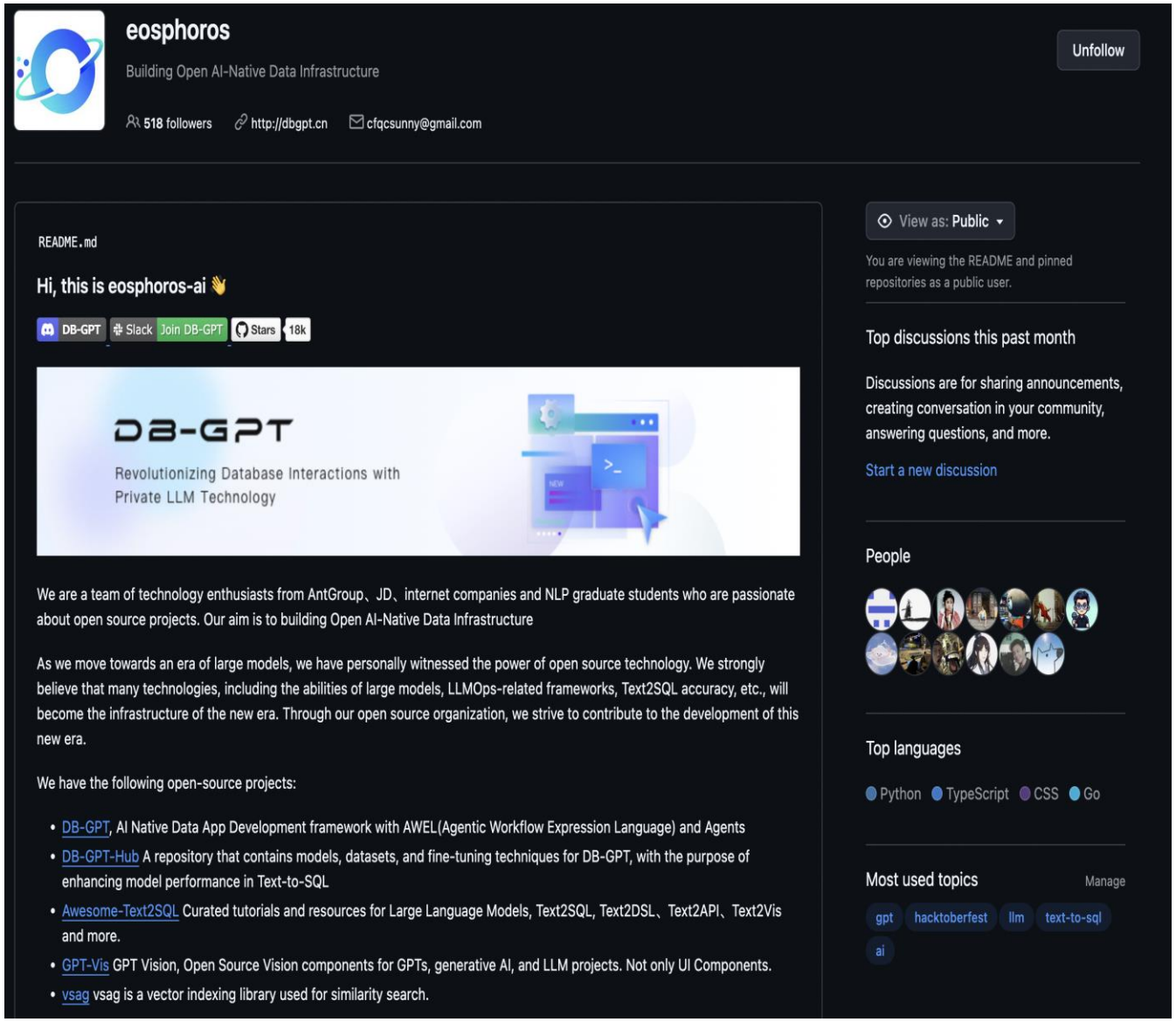
Contents
目录



01 | 社区介绍

社区整体介绍

1.1 DB-GPT 社区介绍(https://github.com/eosphoros-ai/DB-GPT)



eosphoros

Building Open AI-Native Data Infrastructure

Unfollow

518 followers http://dbgpt.cn cfqcsunny@gmail.com

README.md

Hi, this is eosphoros-ai 👋

DB-GPT Slack Join DB-GPT Stars 18k

DB-GPT

Revolutionizing Database Interactions with Private LLM Technology

We are a team of technology enthusiasts from AntGroup、JD、internet companies and NLP graduate students who are passionate about open source projects. Our aim is to building Open AI-Native Data Infrastructure

As we move towards an era of large models, we have personally witnessed the power of open source technology. We strongly believe that many technologies, including the abilities of large models, LLMOps-related frameworks, Text2SQL accuracy, etc., will become the infrastructure of the new era. Through our open source organization, we strive to contribute to the development of this new era.

We have the following open-source projects:

- DB-GPT, AI Native Data App Development framework with AWEL(Agentic Workflow Expression Language) and Agents
- DB-GPT-Hub A repository that contains models, datasets, and fine-tuning techniques for DB-GPT, with the purpose of enhancing model performance in Text-to-SQL
- Awesome-Text2SQL Curated tutorials and resources for Large Language Models, Text2SQL, Text2DSL、Text2API、Text2Vis and more.
- GPT-Vis GPT Vision, Open Source Vision components for GPTs, generative AI, and LLM projects. Not only UI Components.
- vsag vsag is a vector indexing library used for similarity search.

View as: Public

You are viewing the README and pinned repositories as a public user.

Top discussions this past month

Discussions are for sharing announcements, creating conversation in your community, answering questions, and more.

Start a new discussion

People

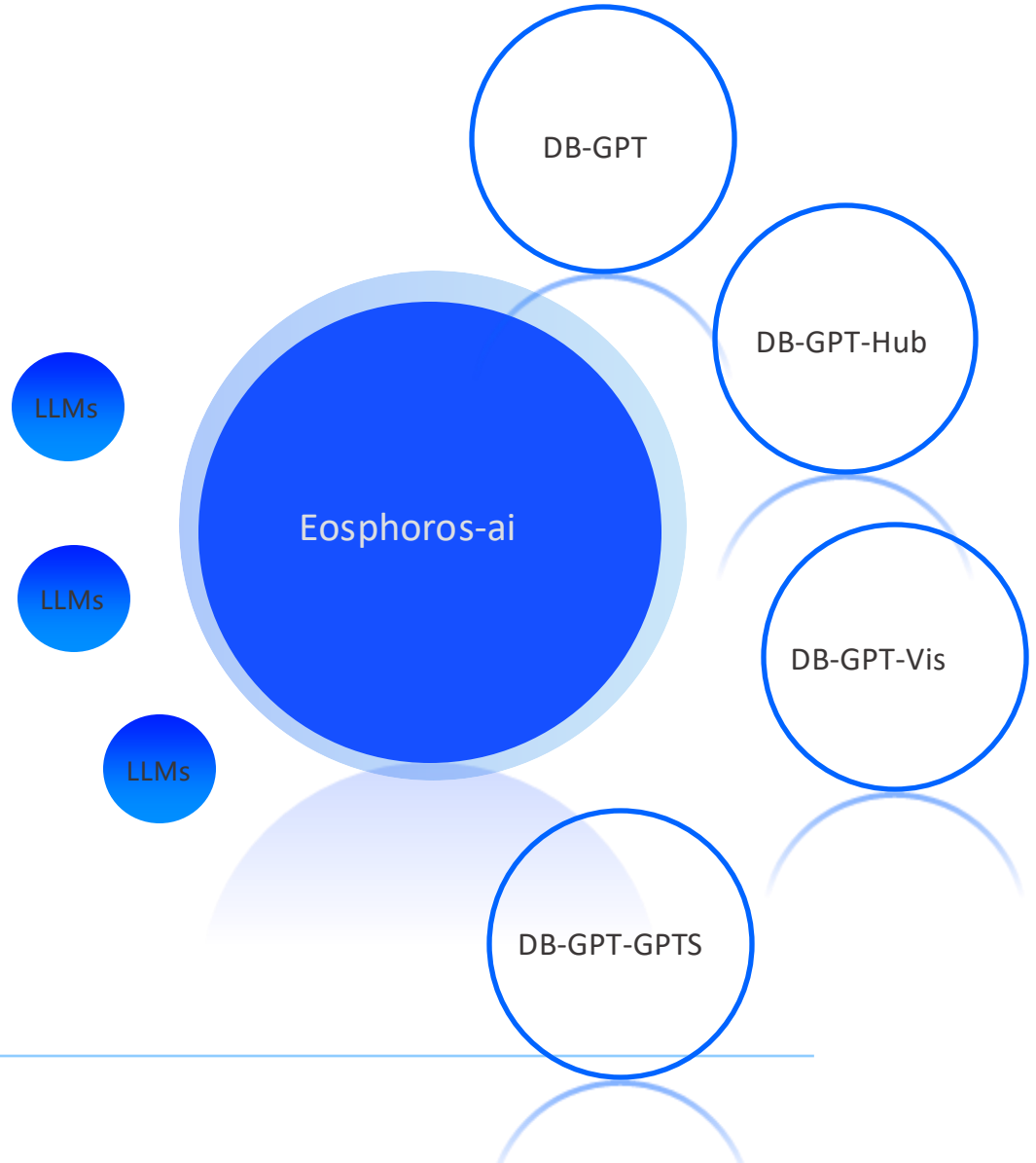
Top languages

Python TypeScript CSS Go

Most used topics Manage

gpt hacktoberfest llm text-to-sql ai

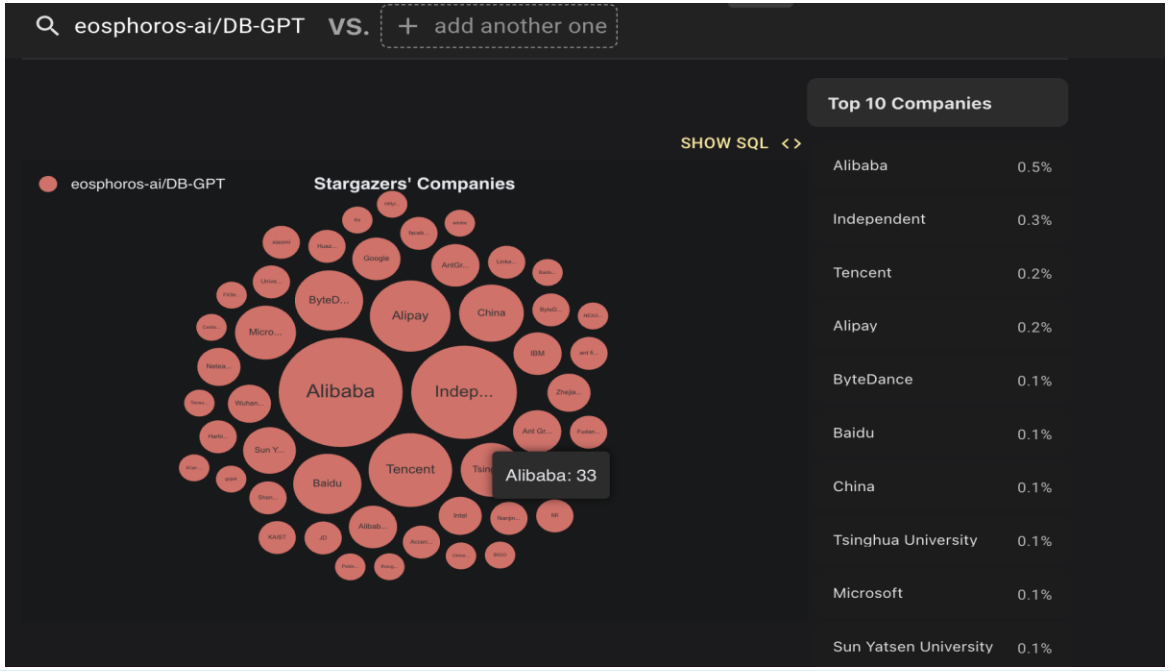
Eosphoros-ai（星辰智能）社区

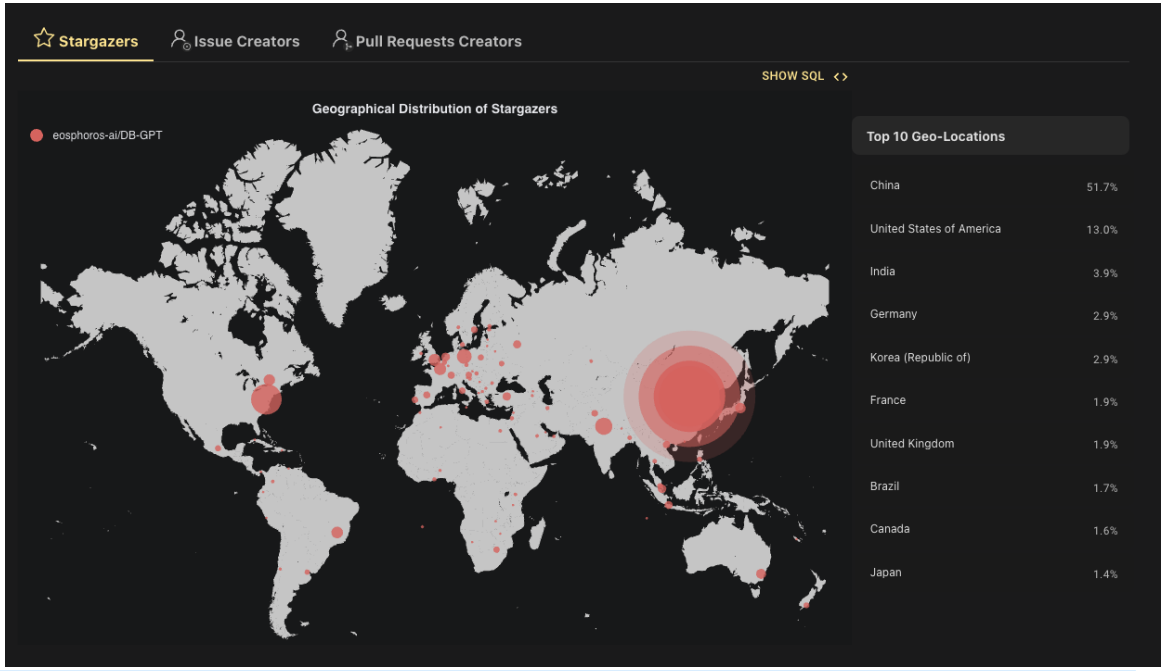


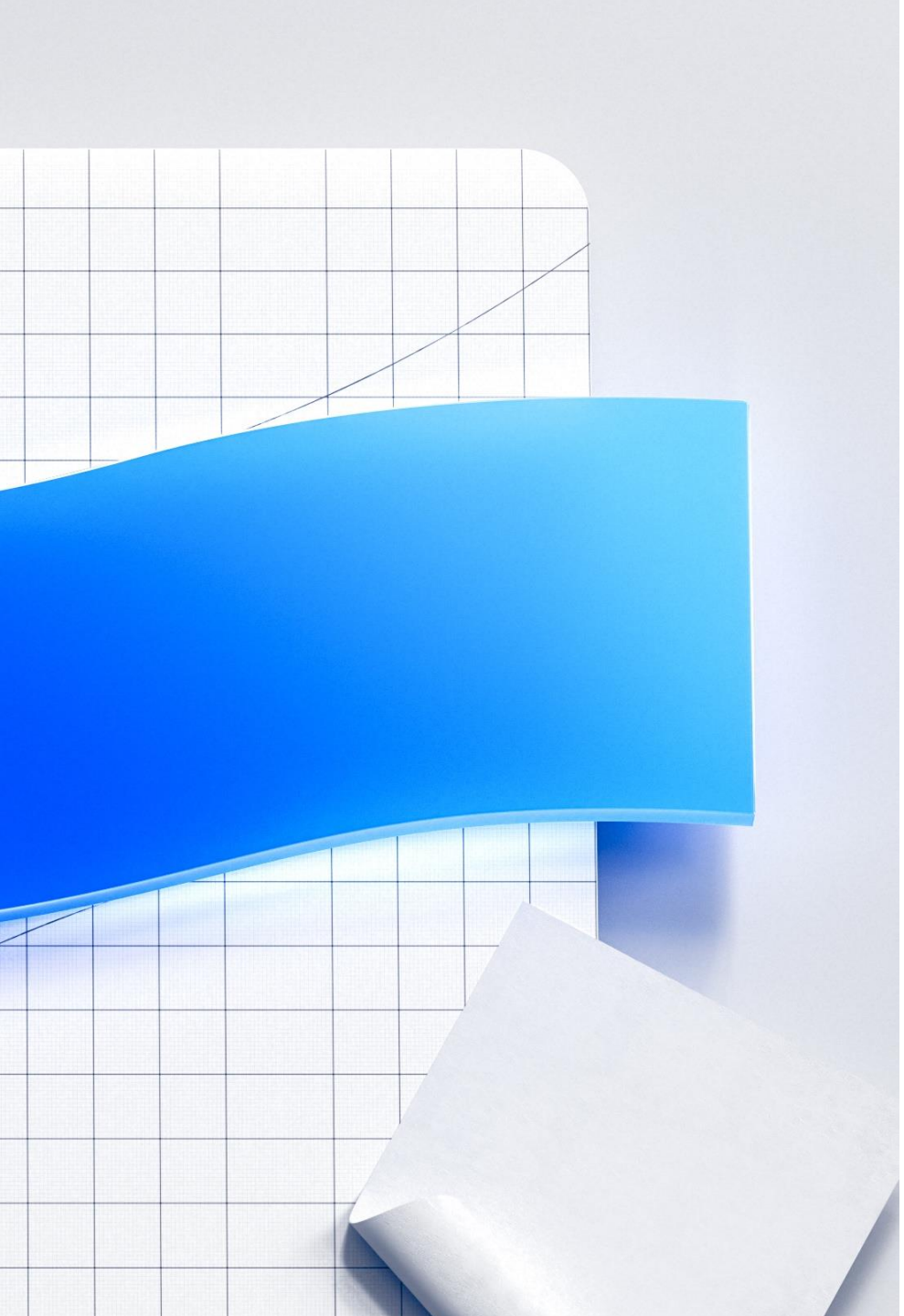
1.2 DB-GPT社区影响力

目前开源社区的现状：**Star数 13.7K+， 社区贡献者100+，3篇Paper**；
社区共建组织有蚂蚁、京东、美团、阿里、唯品会等，社区用户数**5000+**，涵盖了全球范围内的开发者。

社区合作方面已与TuGraph、OceanBase、AntV、ModelScope等社区形成合作。
社区企业用户有: 蚂蚁、京东、美团、唯品会、北京xx能源研究院、北京xx云等公司。







02 | DB-GPT 技术与架构

技术与架构介绍

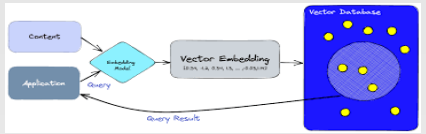
2.1 AI+数据技术变迁之路

整个23年我们一直围绕着AI时代数据的全新交互方式在开源社区做技术探索

向量数据库 → LLM+Data → LLM+SQL → LLM+Agents

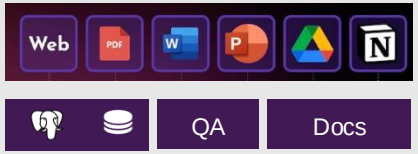
向量存储

向量数据库需求爆发，新老势力开始角逐。 新势力: Pinecone 、Milvus、Weaviate、Qdrant、Chroma 老对手: pgvector、redis、elasticsearch, clickhouse



LLM+Data

框架: DB-GPT、 Langchain、Llama-index、Databrics, **产品:** ChatPDF、ChatExcel、ChatDocs、DB-GPT、ChatDB



LLM+SQL

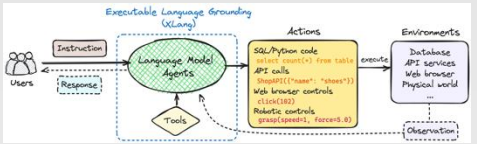
主要产品:
DB-GPT、Chat2SQL、SQLChat、Text2SQL、NL2SQL、SQL-Coplilot、

当前Text2SQL这个领域是围绕传统数据库竞争最激烈的领域。

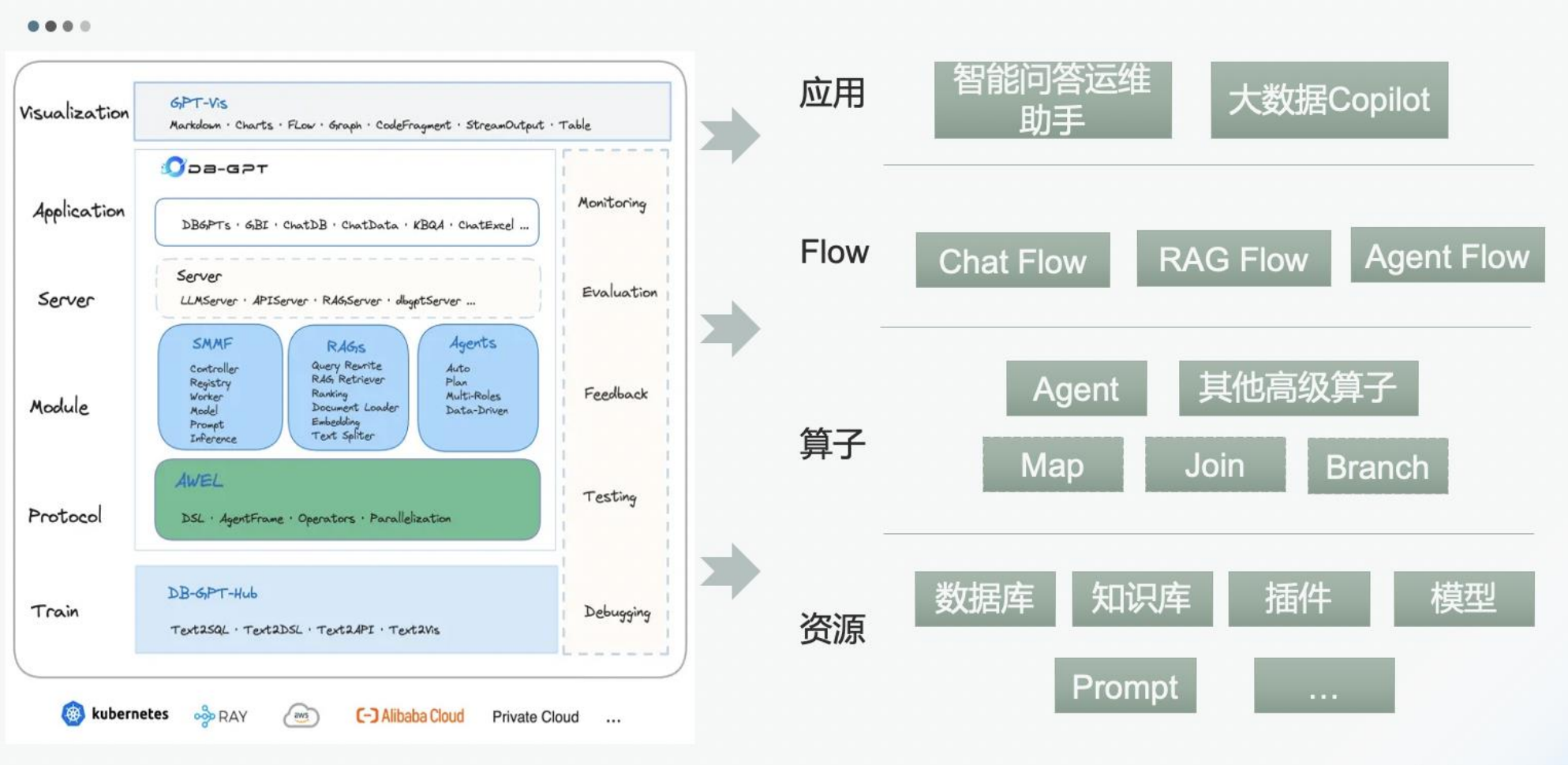
```
SELECT COUNT(*) AS order_count
FROM users
JOIN orders ON users.id = orders.user_id
WHERE users.id = 1;
```

LLM+Agents

Multi-Agents与Plugins是当下最火热的研究方向， 未来数据库领域围绕LLM + Tools也会发生重大的变革。 **理论:** toolformer、Gorilla、toolBench、CoT、ToT **框架工具:** Auto-GPT、DB-GPT、Langchain、GPT-4-plugins、Copilot-X



2.2 整体架构

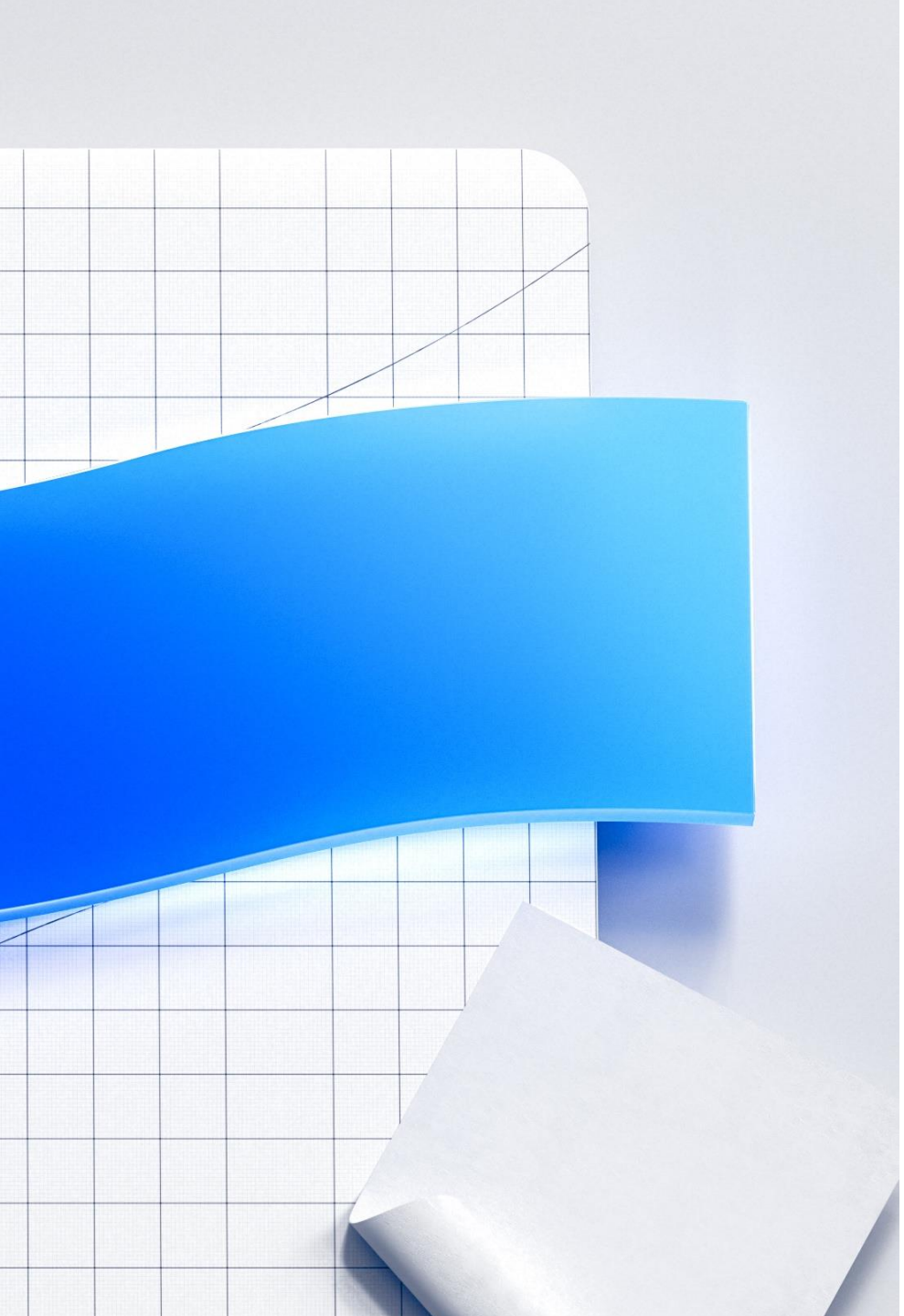

Visualization
GPT-Vis
Markdown · Charts · Flow · Graph · CodeFragment · StreamOutput · Table
Application
DB-GPT
DBGPTs · GBI · ChatDB · ChatData · KBQA · ChatExcel ...
Server
Server
LLMServer · APIServer · RAGServer · dbgptServer ...
Module
SMMF
Controller
Registry
Worker
Model
Prompt
Inference
RAGs
Query Rewrite
RAG Retriever
Ranking
Document Loader
Embedding
Text Spliter
Agents
Auto
Plan
Multi-Roles
Data-Driven
Protocol
AWEL
DSL · AgentFrame · Operators · Parallelization
Train
DB-GPT-Hub
Text2SQL · Text2DSL · Text2API · Text2Vis
Monitoring
Evaluation
Feedback
Testing
Debugging
kubernetes
RAY
Alibaba Cloud
Private Cloud
...
应用
智能问答运维助手
大数据Copilot
Flow
Chat Flow
RAG Flow
Agent Flow
算子
Agent
其他高级算子
Map
Join
Branch
资源
数据库
知识库
插件
模型
Prompt
...

Llama-index VS DB-GPT VS Langgraph

```python
from llama_index.postprocessor import CohereRerank
from llama_index.response_synthesizers import TreeSummarize
from llama_index import ServiceContext

# define modules
prompt_str = "Please generate a question about Paul Graham's life regarding the
prompt_tmpl = PromptTemplate(prompt_str)
llm = OpenAI(model="gpt-3.5-turbo")
retriever = index.as_retriever(similarity_top_k=3)
reranker = CohereRerank()
summarizer = TreeSummarize(
    service_context=ServiceContext.from_defaults(llm=llm)
)

# define query pipeline
p = QueryPipeline(verbose=True)
p.add_modules(
    {
        "llm": llm,
        "prompt_tmpl": prompt_tmpl,
        "retriever": retriever,
        "summarizer": summarizer,
        "reranker": reranker,
    }
)
# add edges
p.add_link("prompt_tmpl", "llm")
p.add_link("llm", "retriever")
p.add_link("retriever", "reranker", dest_key="nodes")
p.add_link("llm", "reranker", dest_key="query_str")
p.add_link("reranker", "summarizer", dest_key="nodes")
p.add_link("llm", "summarizer", dest_key="query_str")
```

```python
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
    )
```

```python
workflow = Graph()

# Add the same nodes as before, plus this "first agent"
workflow.add_node("first_agent", first_agent)
workflow.add_node("agent", agent)
workflow.add_node("tools", execute_tools)

# We now set the entry point to be this first agent
workflow.set_entry_point("first_agent")

# We define the same edges as before
workflow.add_conditional_edges(
    "agent",
    should_continue,
    {
        "continue": "tools",
        "exit": END
    }
)
workflow.add_edge('tools', 'agent')

# We also define a new edge, from the "first agent" to the tools node
# This is so that we can call the tool
workflow.add_edge('first_agent', 'tools')

# We now compile the graph as before
chain = workflow.compile()
```

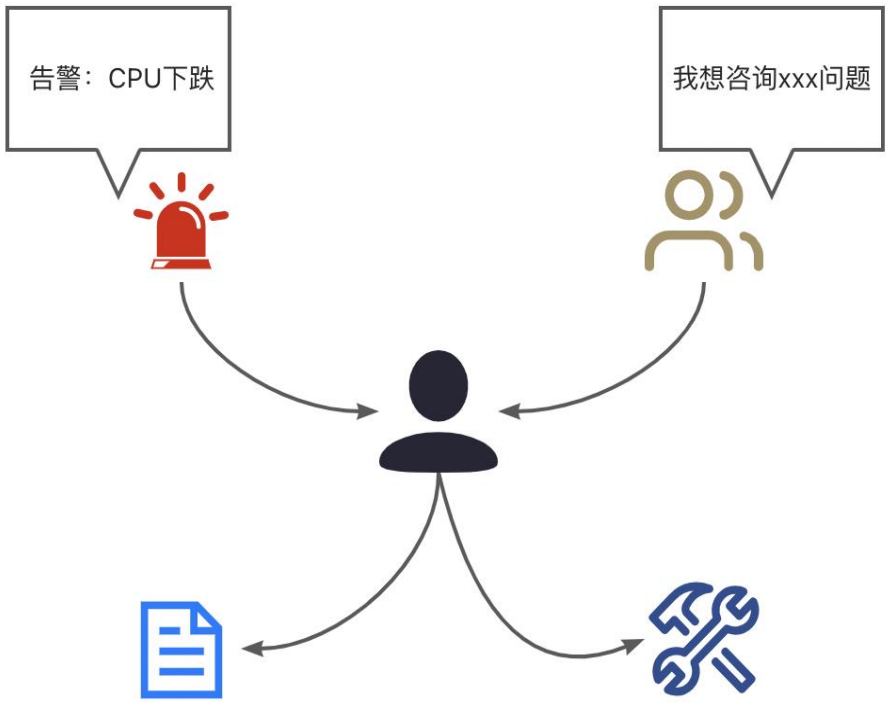


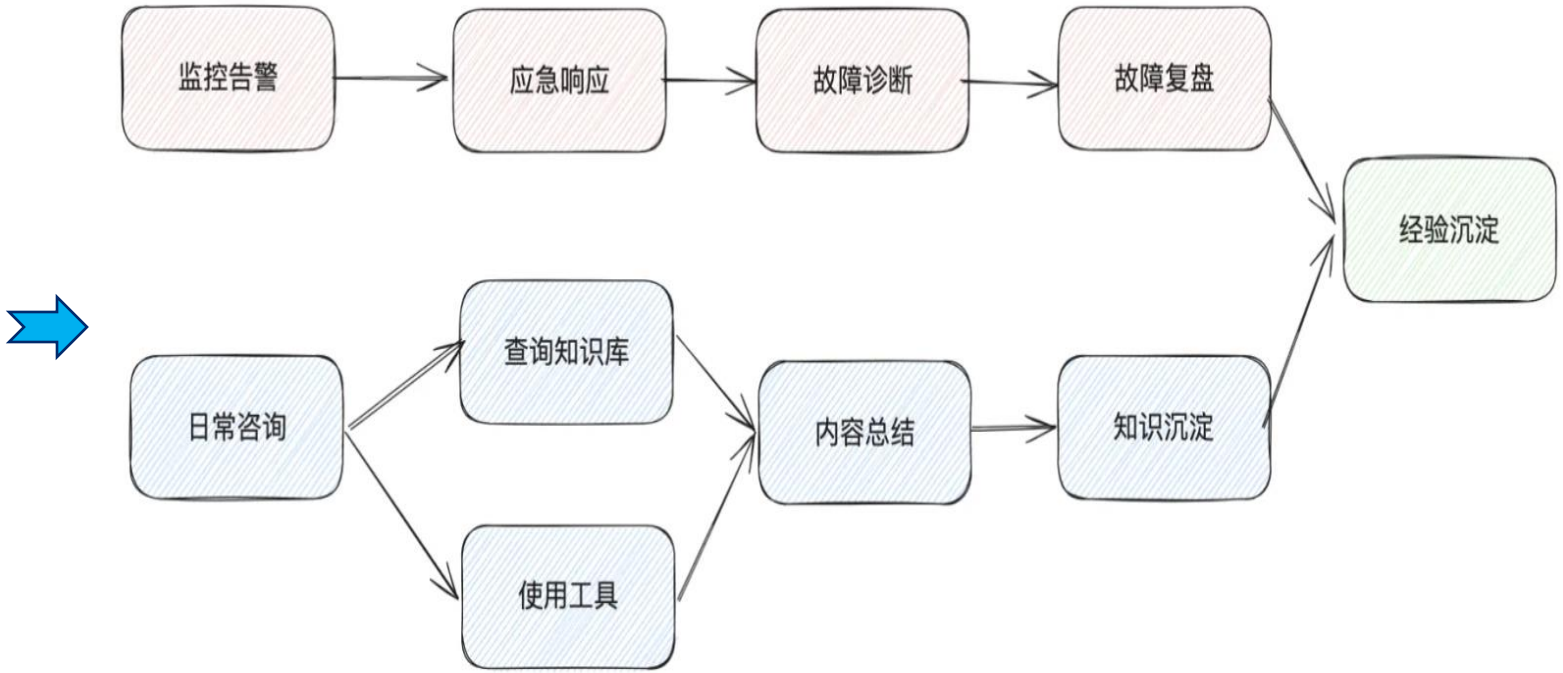
03 | 落地实践分享

DB-GPT结合GraphRAG的企业落地实践

3 实践-数据基础设施智能体

数据基础设施SRE日常工作内容





3 实践-数据基础设施智能体

数据基础设施智能体

传统的 RAG + Agent 技术可以解决通用的，确定性没那么高的，单步任务场景。
但是面对数据基础设施领域的专业场景，整个检索过程必须是确定，专业和真实的，并且是需要一步一步推理的。

记忆类型	普通智能体需要的RAG	数据基础设施智能体需要的RAG
短期记忆	为智能体提供一次性的工作上下文记忆	基础设施智能体需要的专属工作流程(告警->定位->止血->恢复)上下文记忆
长期记忆	为智能体提供长期记忆，可以把相关知识和用户偏好进行存储，通常采用向量数据库	基础设施智能体需要的应急经验 沉淀，为了到达完整，确定，专业，采用知识图谱进行建模

DB抖动排查步骤

数据库抖动（Database Thrashing）通常指的是数据库系统在高负载下，由于频繁的I/O操作和资源争用，导致性能急剧下降的现象。这可能是因为内存不足、磁盘I/O瓶颈、查询效率低下、索引设计不合理等原因造成的。以下是一些排查和解决数据库抖动的步骤：

1. 监控和收集数据

- **性能监控**：使用数据库自带的监控工具或第三方工具（如Prometheus、Grafana、Zabbix等）监控数据库的性能指标，包括CPU使用率、内存使用情况、磁盘I/O、网络延迟等。
- **日志分析**：查看数据库的日志文件，特别是错误日志和慢查询日志，以识别可能的问题。

2. 识别瓶颈

- **资源使用情况**：
  - **CPU**：检查是否有CPU使用率过高的情况，特别是某个进程或线程。
  - **内存**：检查是否有内存不足的情况，特别是缓存命中率低、频繁的交换操作等。
  - **磁盘I/O**：检查磁盘I/O是否成为瓶颈，特别是读写延迟和队列长度。
- **查询性能**：使用 `EXPLAIN` 或类似工具分析慢查询，识别查询计划中的问题。

3. 优化查询

- **索引优化**：确保查询中使用的列上有适当的索引，避免全表扫描。
- **查询重写**：优化SQL查询，减少不必要的子查询、连接操作等。
- **批处理**：将多个小查询合并为一个大查询，减少I/O操作。

3 实践-数据基础设施智能体

知识索引：知识提取 + 图谱构建

问题描述

三两句话让大家了解问题内容及影响

处理过程

详细描述问题发生的时间线

原因分析

详细描述原因

改进优化

介绍哪里做的好、哪里做的不足

ACTION

问题修复action

流程改进

其他（注重举一反三，完全避免问题的发生

复盘沉淀

热点行

事件

事件名：热点行事件，xxx

事件描述

来源

诊断来源

诊断时间

诊断工具

工具平台：

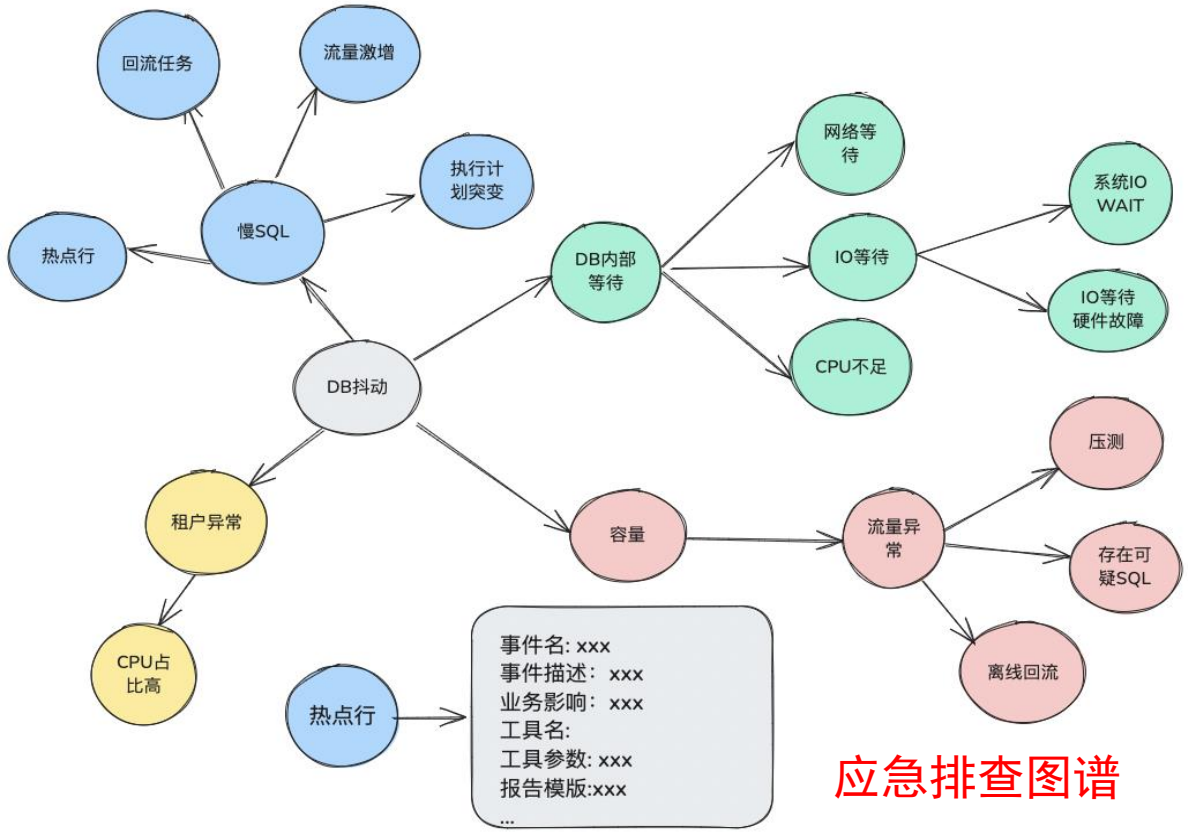
工具

工具参数1

工具参数2

应急预案

知识经验

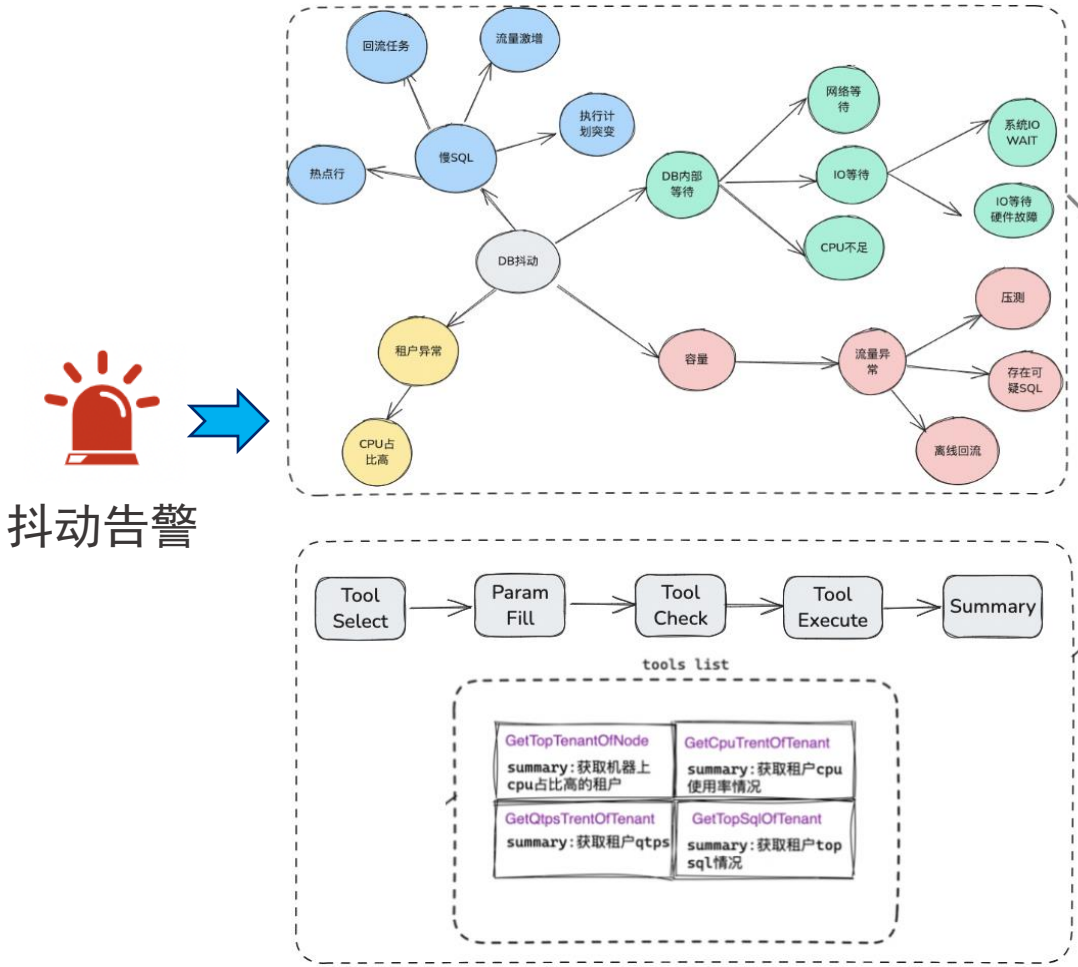


应急排查图谱

通过历史应急事件复盘流程文档 + 专家沉淀的排查故障经验转化为应急排查事件驱动的知识图谱

3 实践-数据基础设施智能体

GraphRAG(静态检索) + 工具执行(动态检索)



1.通过antmonitor找出导致节点cpu使用率高的租户
2.在antmonitor确认该租户的cpu使用率情况，是否为导致节点cpu使用率高的原因
3.确认租户的qtps,logic reads变化趋势
4.如果qtps有明显上涨(超过30%)，说明qtps上涨即为根因
5.如果qtps没有明显上涨，则通过tars排查该租户的慢SQL情况

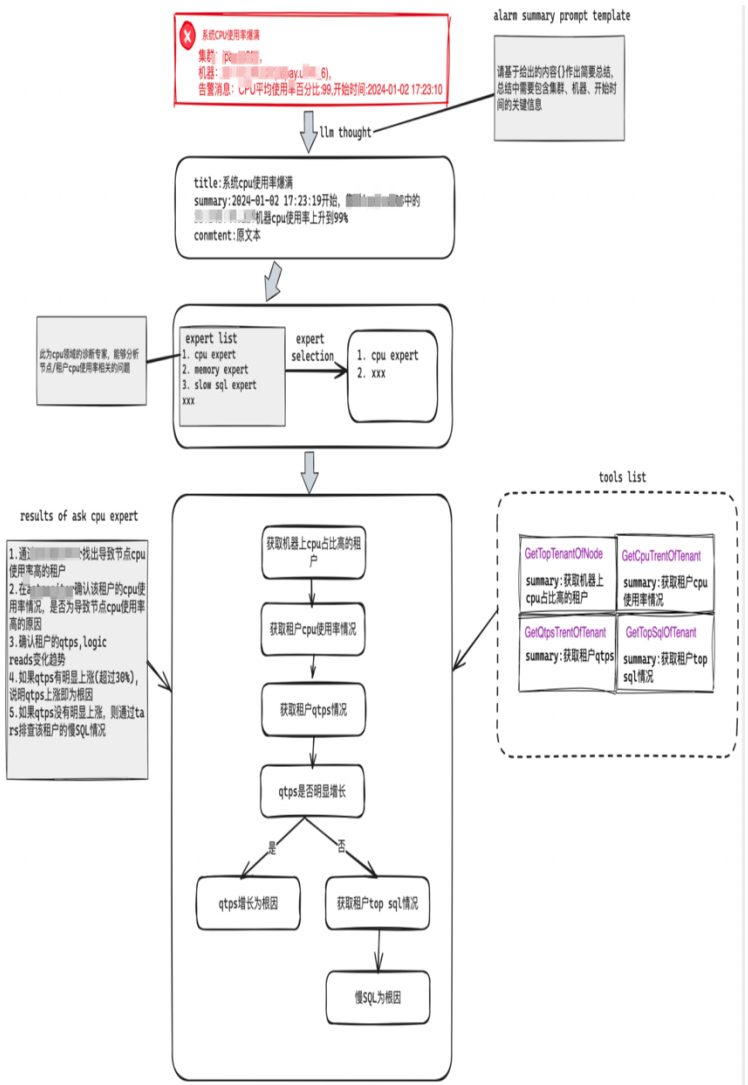
VS

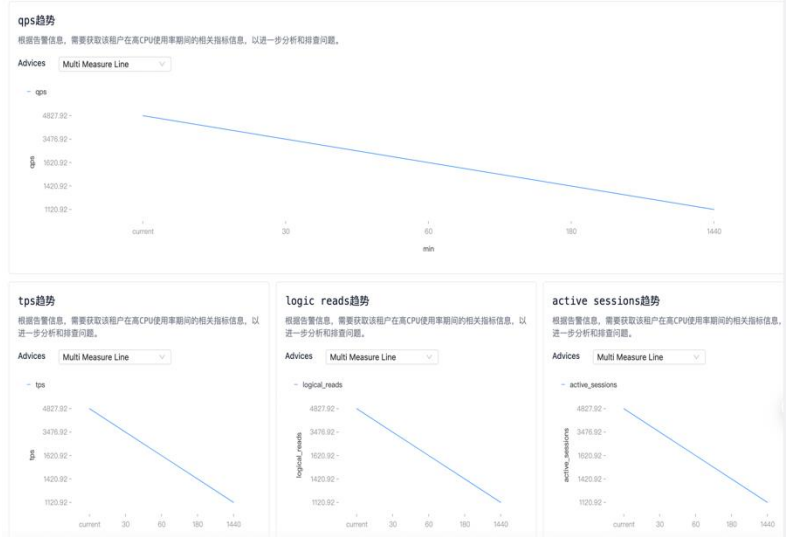
Native RAG

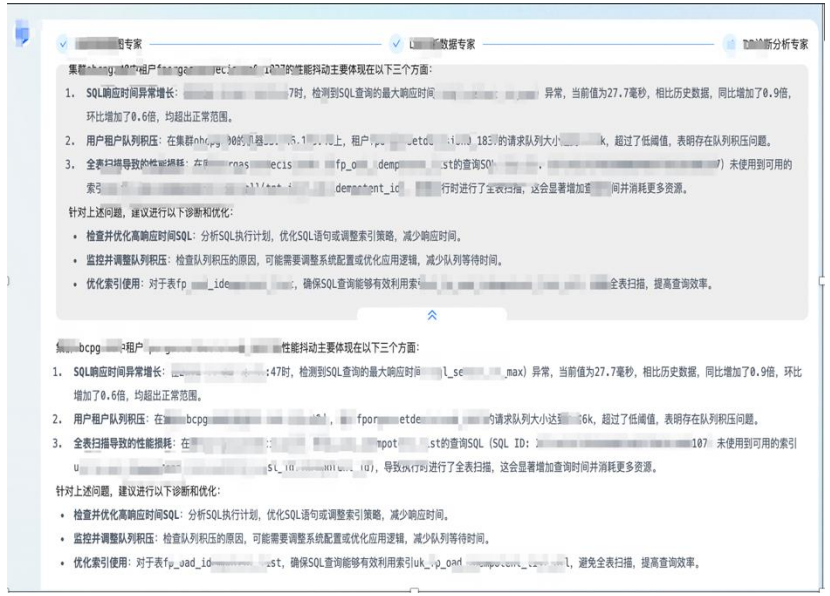
动静结合的混合召回的方式比Native RAG召回，保障了数据基础设施智能体执行的确定性，专业性和严谨性

3 实践-数据基础设施智能体

AWEL + Agent







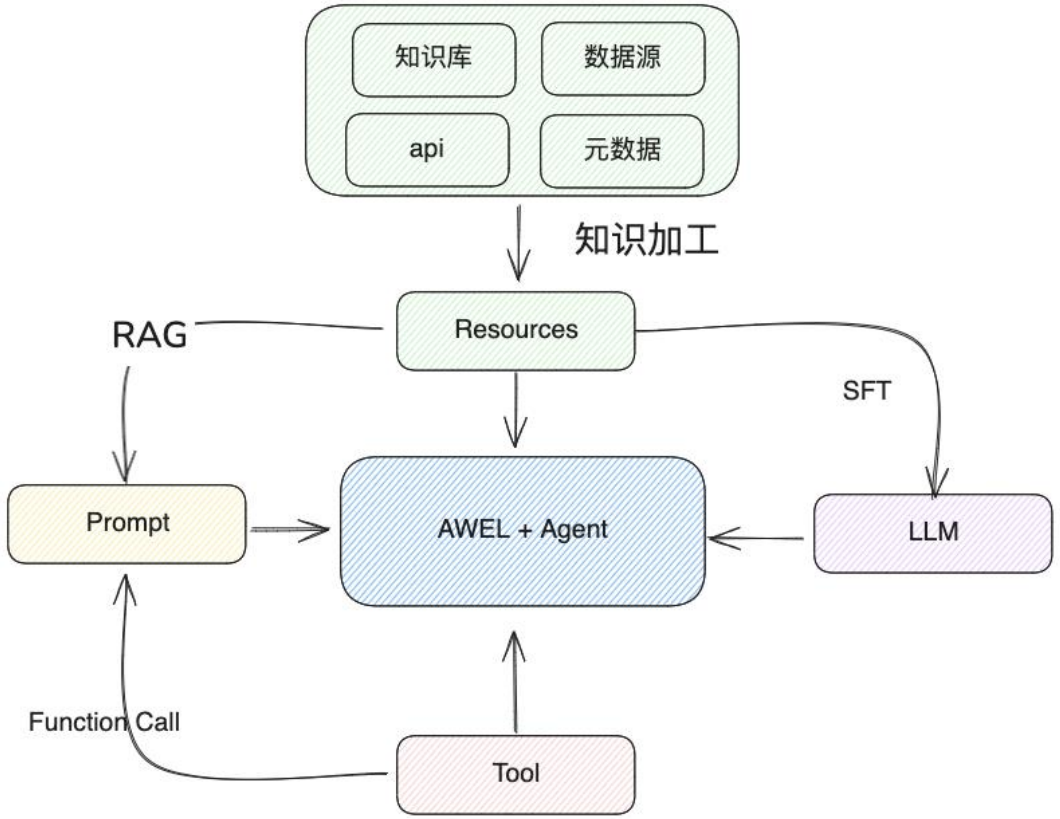
4 总结与展望

1.梳理业务资源/数据资产，文档，数据库，API等

2.知识整理+多元信息结构化抽取(表格/实体关系/语义)

3.添加Agent并绑定需要用到的资源/数据资产

4. 编排AWEL + Agent 形成最终的智能体应用



4 总结与展望

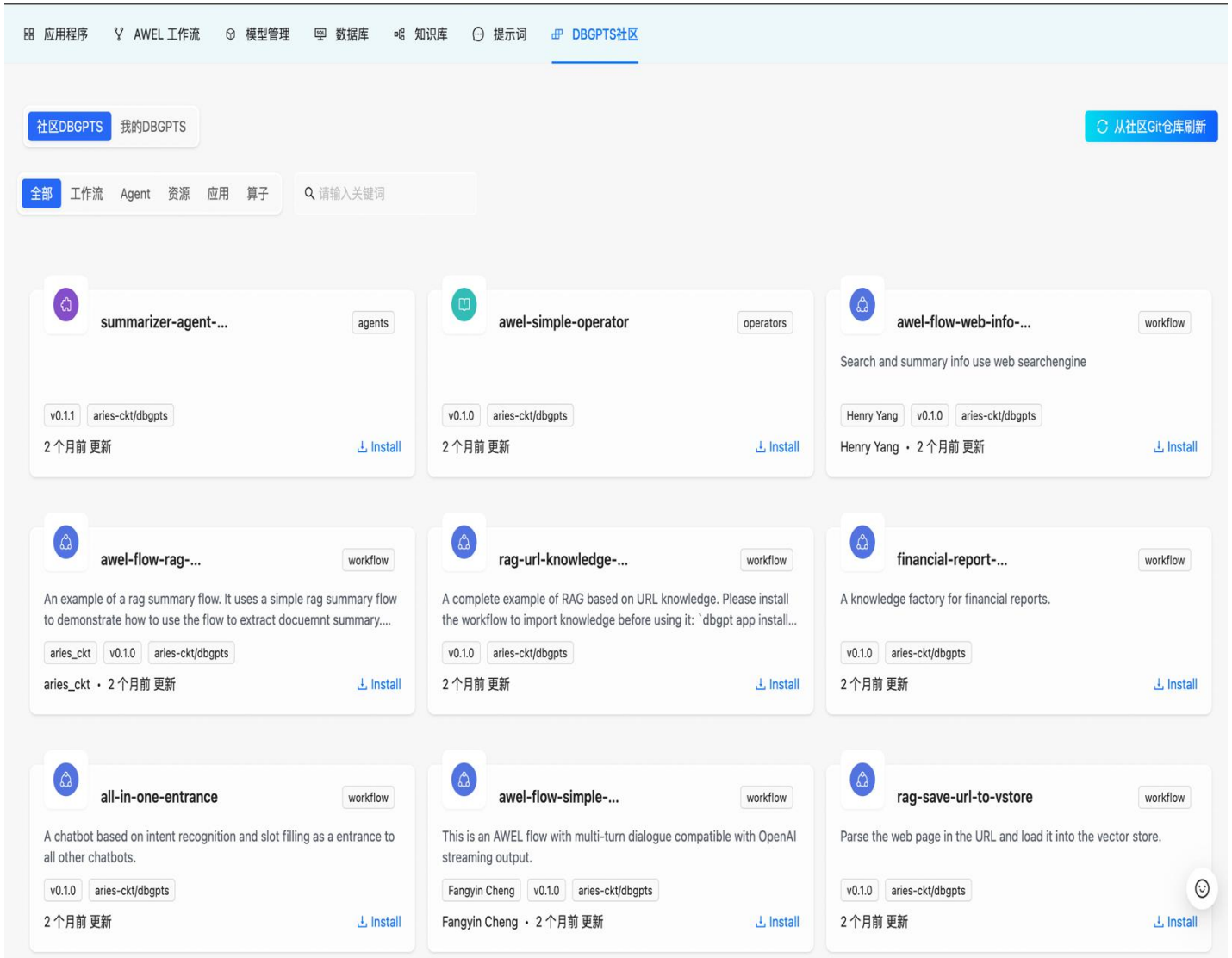
社区展望

1.知识加工工厂，如何多元化，结构化地抽取知识(表格/语义/实体关系)

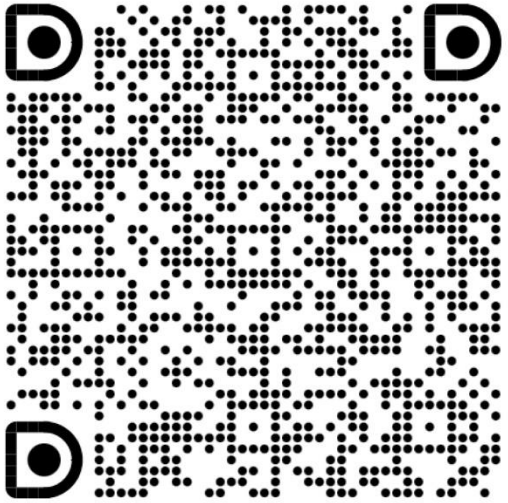
2.知识召回的丰富性和准确性(TableRAG/GraphRAG/ToolRAG/HybridRAG/)

3.评测(RAG & Agent)

4. DB-GPTs



项目地址：https://github.com/eosphoros-ai/DB-GPT



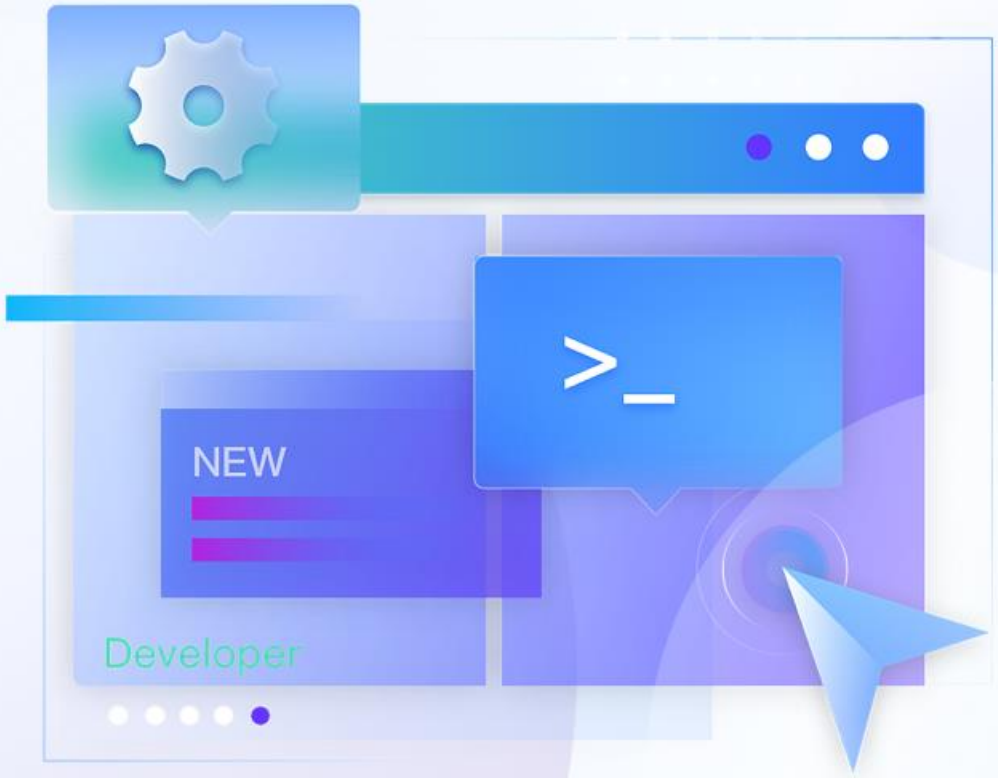
钉钉问答群

微信公众号

Thank you!

DB-GPT实践案例分享

让GenAI更理解数据



Bruce

LLM全栈工程师

2024/11/09

Contents
目录

GenAI为整个高科技价值链增添价值

生成式人工智能可以在整个价值链中发挥重要作用，管理产品开发周期，优化制造流程，并确保有效的资源管理，以协调供需周期

研发 / 工程

加速产品上市以应对剧烈变化的技术格局的需求 - 新兴科技公司的需求

- 用于**加速产品/软件生命周期**的生成式 AI
- 用于**新产品开发（NPD）**的生成式 AI
- 用于**工程测试和分析**的生成式AI
- 用于**识别、构建和验证新商业模式**的生成式AI

财务

通过更好的财务风险管理和成本效益分析来**改善业务决策**

- 用于提供更好的**风险和机会评估**的生成式 AI
- 用于**发票优化**的生成式 AI
- 用于**报告和分析叙述**的生成式 AI
- 用于**场景建模**的生成式 AI

供应链

使供需保持一致从而更好地进行预测，以改善客户体验并降低持有成本

- 用于实现高效的**物料风险管理**的生成式AI
- 用于实现无缝**合同管理**的生成式 AI
- 用于改进**供应商尽职调查**的生成式 AI
- 用于**采购和供应链路线优化**的生成式AI

服务与运营

通过更好的工艺优化**减少浪费并提高产量**

- 基于生成式 AI 的**现场技术助理**
- 生成式 AI 通过**视觉检测**提高质量
- 生成式 AI 生成**多模态工作指令**
- 生成式 AI 通过**数字孪生**提高运营效率

销售与客户体验

以更低的成本提供低接触的个性化营销和销售材料，**加快销售完成速度并提高客户满意度**

- 提供**个性化的销售协助和推荐**的生成式AI
- 基于生成式 AI 的自助服务**虚拟助手**
- 用于**有针对性的营销活动**的生成式 AI

人才管理

优化**招聘决策、培训和资源管理**

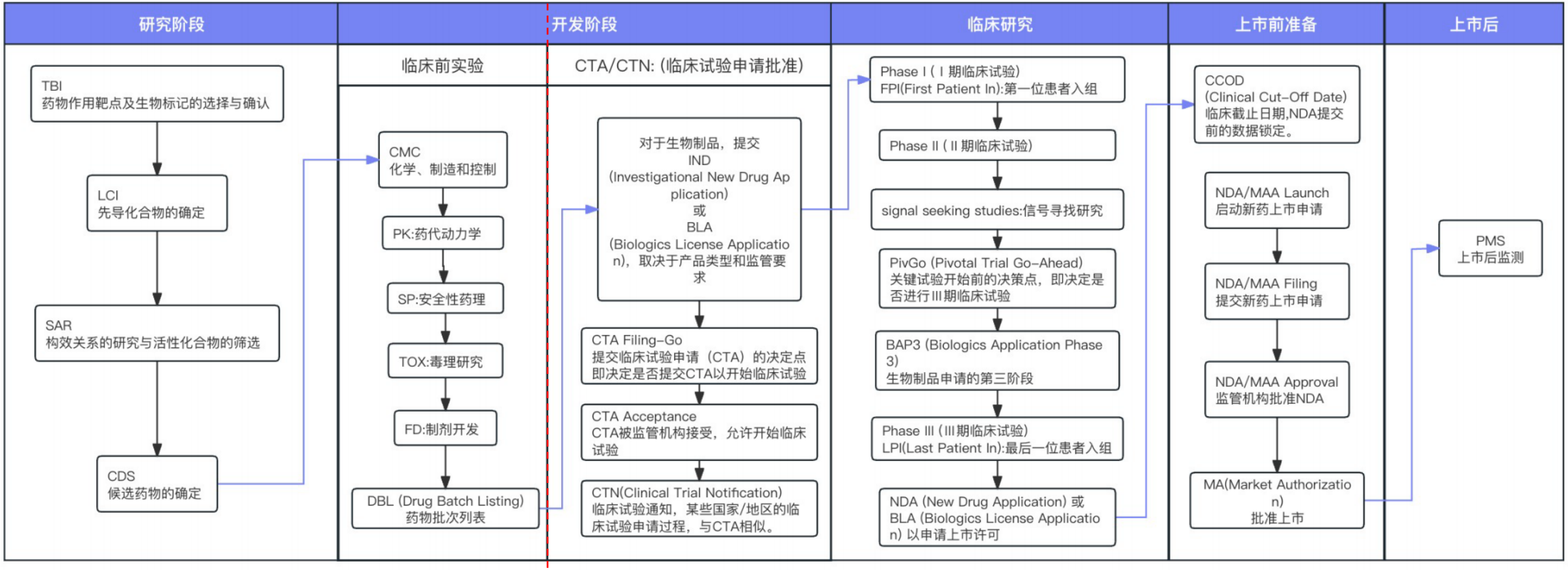
- 用于快速**招聘人才**的生成式 AI
- 用于高效**培训和技能提升的**生成式AI
- 用于**员工绩效管理**的生成式 AI
- 用于**个性化辅导**的生成式 AI

生成式 AI 背景下的现代数据基础

企业需要重新审视数据管理，并为大模型的到来做好知识管理和储备

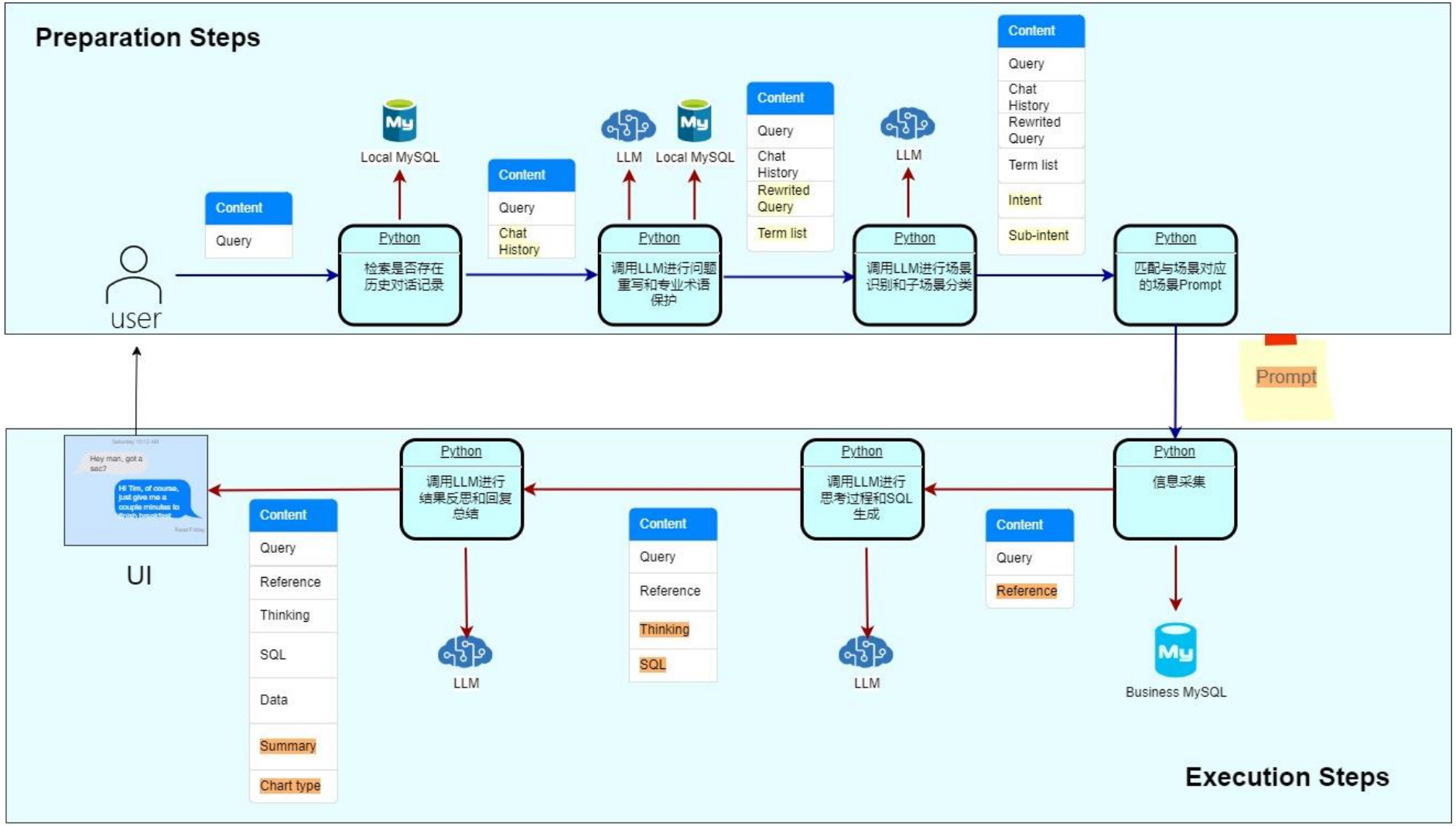
案例-药品研发数据对话式分析

药品研发里程碑示例图



端到端RAG过程

Rag Process = 问题重写 + 意图识别 + 子场景处理 +其他



QA场景划分

让问题从一个空间分散至多个空间，有利于理清思路处理问题

SN	Category	Scenario	Sample Questions
1	简单查询	单属性数据提取	XXX Program的NDA Approval的实际日期
2		多属性数据提取	XXX Program的NDA Approval的计划日期和实际日期
3		跨表查询	所有获得NDA Approval的XXX Program的适应症
4		单program查询	XXX Program的最新状态
5		跨program查询	有多少NDA Launch的Program
6	约束查询	属性约束查询 （注意特殊：时间段，Quarter）	从2020年到2023年，一共获批多少个项目
7		属性排序查询（最好 / 最差 / Top K）	China FPI距离Global LPI时间最长的项目
8	趋势查询	属性值的趋势查询	从2020年到2023年，每一年分别是哪些项目获批
9	推理计算	聚合计算（汇总，平均值）	2023年，XX治疗领域（TA）的项目从CTN到FPI的平均时长
10		判断计算（是/否，有/无）	2023年，是否有DBL-NDA filling超过四个月的项目
11		比较计算	从2020年到2023年，肿瘤学领域获批的项目比眼科学领域获批的项目多几个
12		分组统计	从2020年到2023年，肿瘤学领域和眼科学领域分别获批多少个项目
13		列计算	XXX Program的FPI实际日期与计划日期时间间隔多久
14	KPI计算	单KPI查询	2023年，从CTN-FPI大于5个月的项目是哪个
15	多轮对话	多轮对话能力(上轮内容替换 / 基于上轮对话filter / 下钻)	1.2023年，哪些program进入了CTN阶段 2.这些program中有哪些属于肿瘤学治疗领域
16	术语查询	别名/术语查询（适应症、疾病领域等的别名）	PTS的全称是什么
17		中英文混合查询(适用所有category)	XXX Program获得新药上市批准的actual date是哪天
18	预置模版报告的生成	生成预置模版的报告	PD产品线内容总结报告模板预制与报告生成

截止目前，已成功在国内落地多个生成式 AI场景

01 生成式 AI 起步

通过治理、价值捕获流程和技术支持建立所需的基础，以增强客户的 生成式 AI 能力并加快实现价值的速度：

- 重塑战略
- 生成式 AI COE、POC 和 Pilot
- 人工智能学院
- 负责任的 AI

02 基础模型服务

根据行业和企业定制基础模型，并提供数据、架构和平台服务，以加快客户大规模试验或部署的能力：

- 评估和决策框架
- 数据收集、管理和培训
- 型号微调平台选型
- 模型维护
- 企业知识库管理
- 架构蓝图开发
- 计算基础设施设计
- 监测和控制能力

03 行业、功能和应用特定的 生成式 AI

利用打包的生态系统合作伙伴解决方案和全方位服务，定义、构建和部署定制的行业和功能代 AI 解决方案：

- 生成式AI联络中心
- 营销内容供应链
- IT 转型
- 语言翻译服务
- 企业知识检索
- 行业解决方案
- 数字化产品创新
- 供应链弹性

场景

生命科学

- 中国国内大模型洞察咨询
- 生成式 AI 支持的数据分析和洞察 PoC 实施
- 销售代表智能助手
- 医疗报告的智能解读
- 微信内容生成
- 基于角色的生成式 AI 训练

高科技

- 生成式AI技术和行业洞察咨询，包括算力、大模型加速、生态系统合作伙伴及其客户对生成式 AI的需求等
- 支持 生成式 AI 的测试
- 生成式 AI 税法解读机器人
- MLOps和大模型Ops规划平台设计

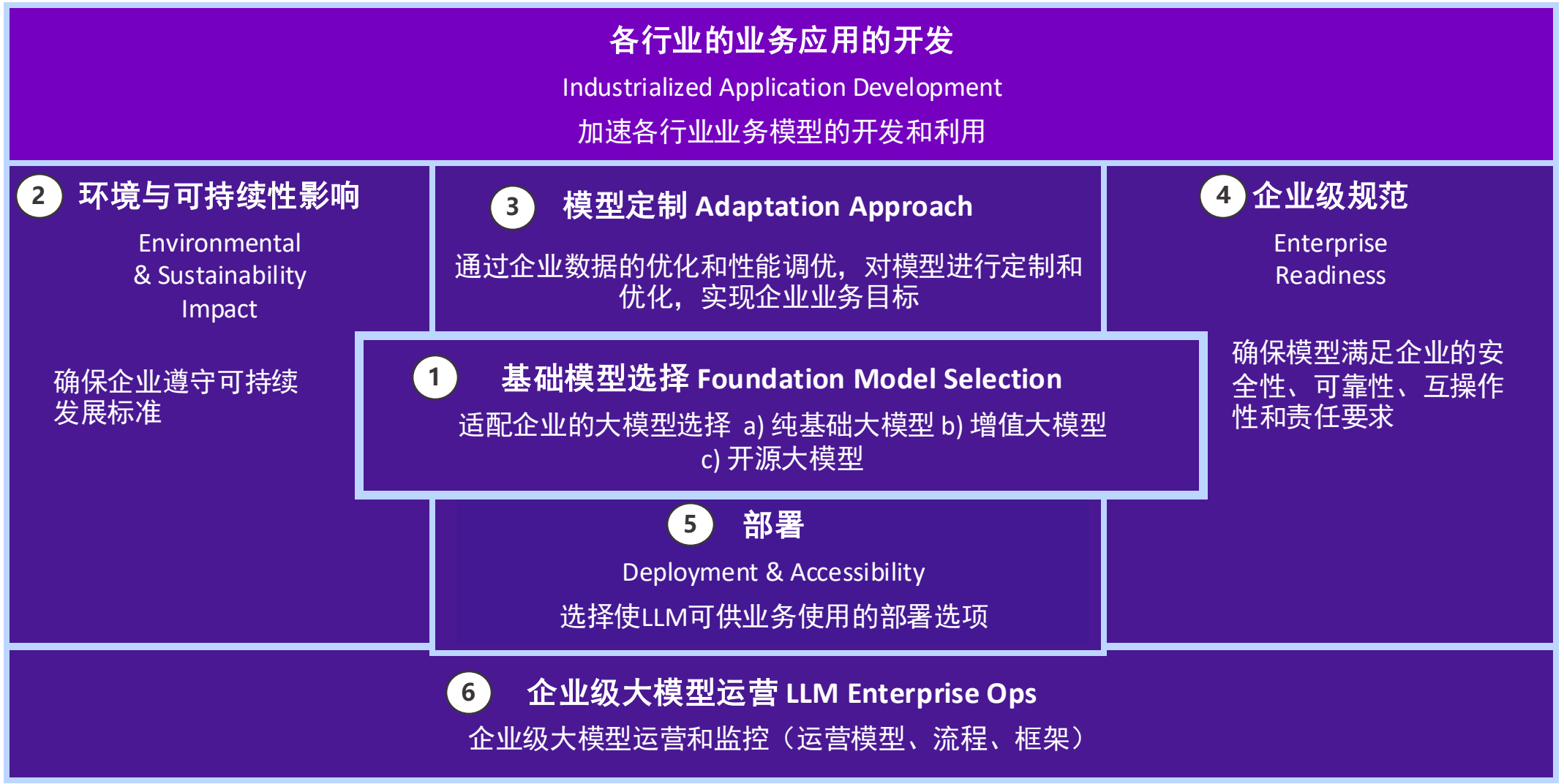
零售

- 生成式 AI 支持代码理解和技术文档生成
- 奢侈品牌的生成式 AI战略规划
- MLOps 和 大模型Ops 规划

工业

- 生成式 AI赋能行业软件产品研发
- 支持生成式 AI 的 SDLC

LLM架构师需要思考的维度

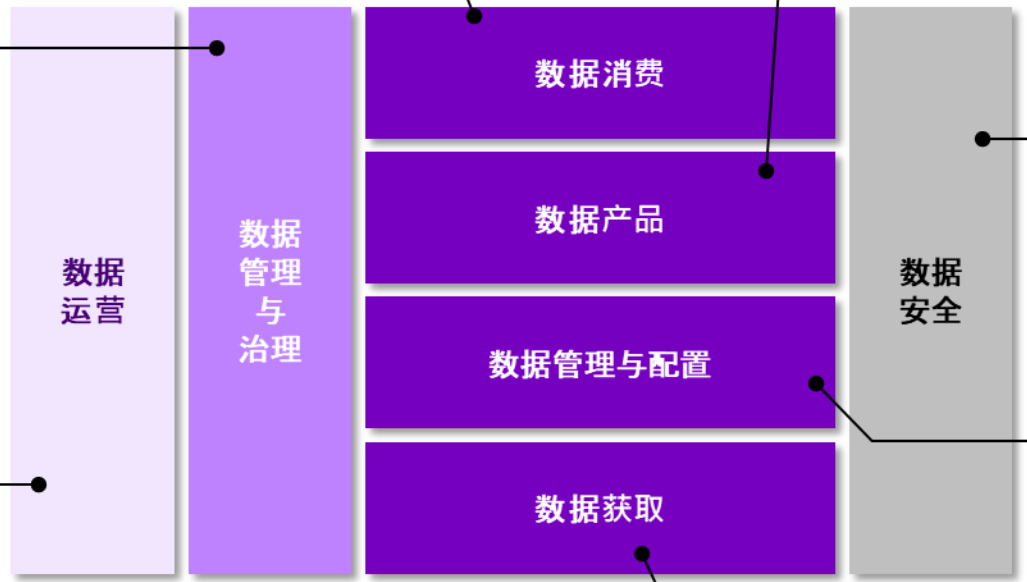


数据基础强化GenAI新能力

- 强化数据治理与数据标准管理
- 严格的数据准入标准
- 通过元数据管理提升生成式AI的透明度与可靠性

- 语义搜索与多模态交互

- 认知型知识图谱
- 面向检索增强生成[1]的矢量化数据产品
- 提示形目录



- 数据伦理与可信赖的人工智能控制
- 动态遮蔽技术
- 基础模型的细粒度访问控制

- 全生命周期的数据运营管理
- 扩展数据的生命周期，涵盖嵌入、提示工程管道、矢量数据存储和知识图谱
- Model ‘gardens’
- LLMOps

- 大规模向量存储与认知知识图谱构建

- 摄取大规模非结构化数据，涵盖多模态

Thank you!