KUBERNETES COMMUNITY DAYS BEIJING 2025

云原生图数据库 NebulaGraph 驱动的 GenAI 技术演进

演讲人：尚卓燃（PsiACE）

KCD
GROWING CLOUD NATIVE TOGETHER
BEIJING

Content

目录

Part 01

背景趋势

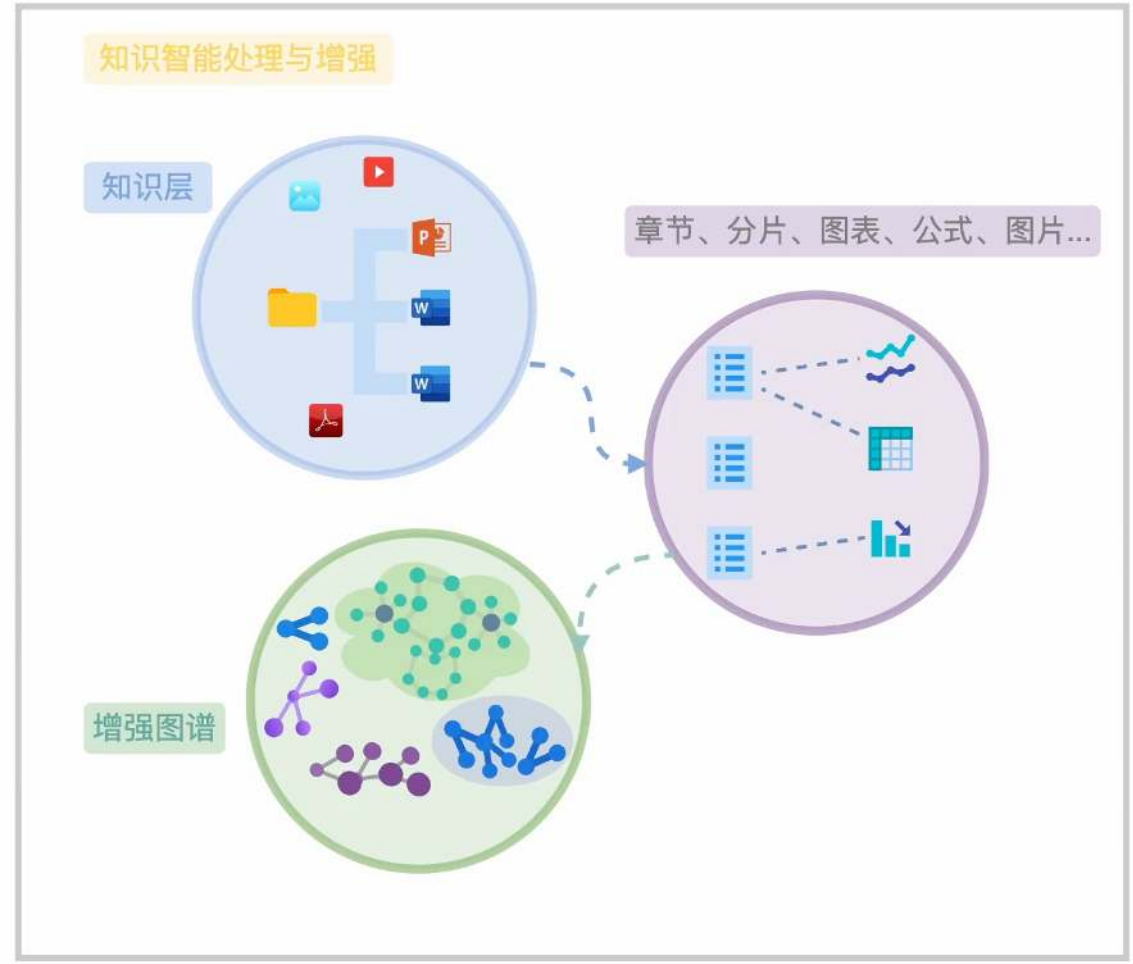
当图数据库遇上 GenAI

传统 RAG 方法的痛点

传统RAG方式面临的挑战：

- **细粒度知识检索能力不足**
- **全局上下文关联缺失**
- **向量相似性与相关性错配**
- **全局性问题及推理型问题回答能力不足**

GraphRAG 的优势



基于图技术的 RAG技术的优势：

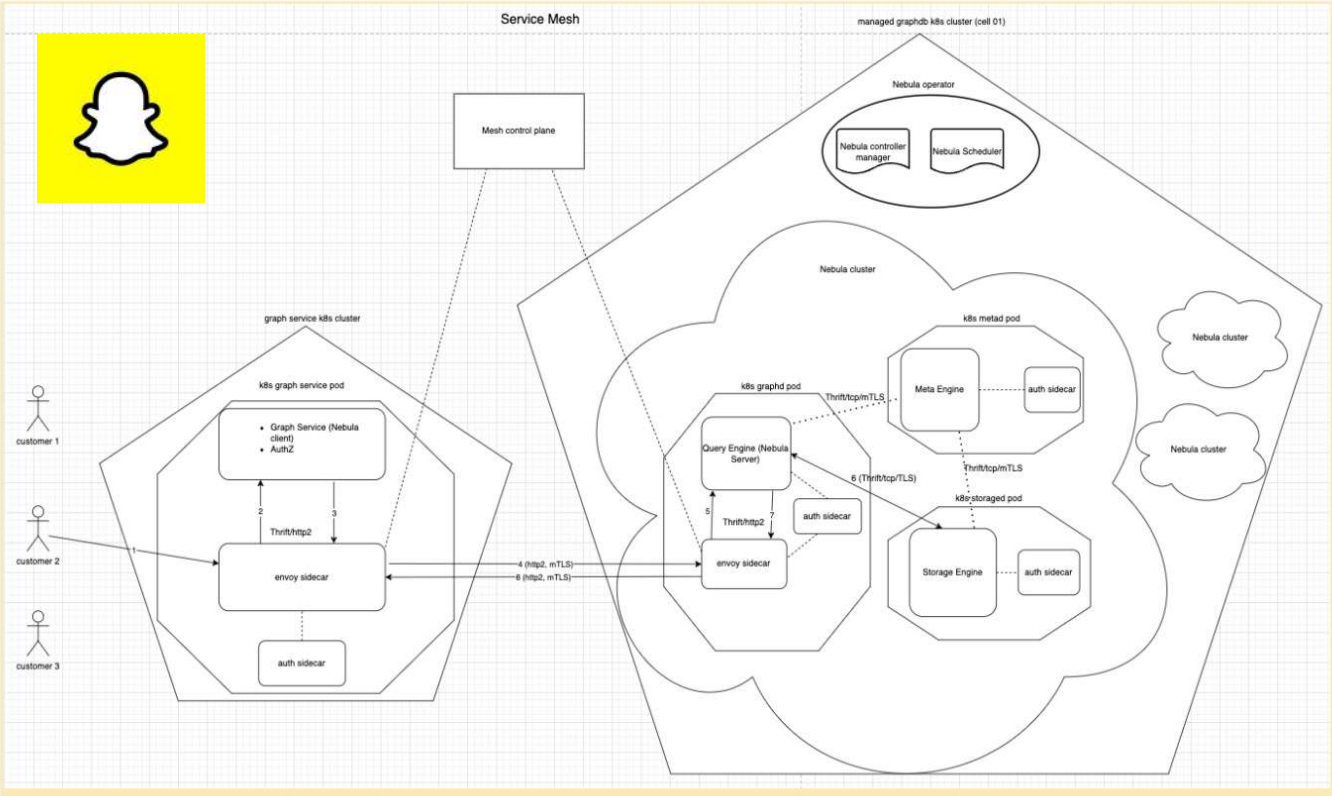
- **细粒度的切分实体和关系，保留了高度凝练的知识细节**
- **保留事物间的关联关系，提升可解释性**
- **图查询和图算法得到相关上下文**

NebulaGraph GenAI 团队的成果：

- **业内首个提出 GraphRAG 方案构想的团队**
- **贡献了 SubGraph RAG 和 Chain of Exploration**

云原生图数据库的价值 – 完整的 Infra 基座

云原生图数据库的价值 – 经过大规模复杂场景检验



解决方案

- 超大规模的的图谱：用户关系超过 100 亿，单个最大规模图谱超过400 亿点，1000 亿边
- 多场景使用：广告推荐、内容推荐、好友推荐、镜头推荐等各类场景
- 超大并发：TPS 超过 150万/秒，QPS 超过 8万/秒

业务成果

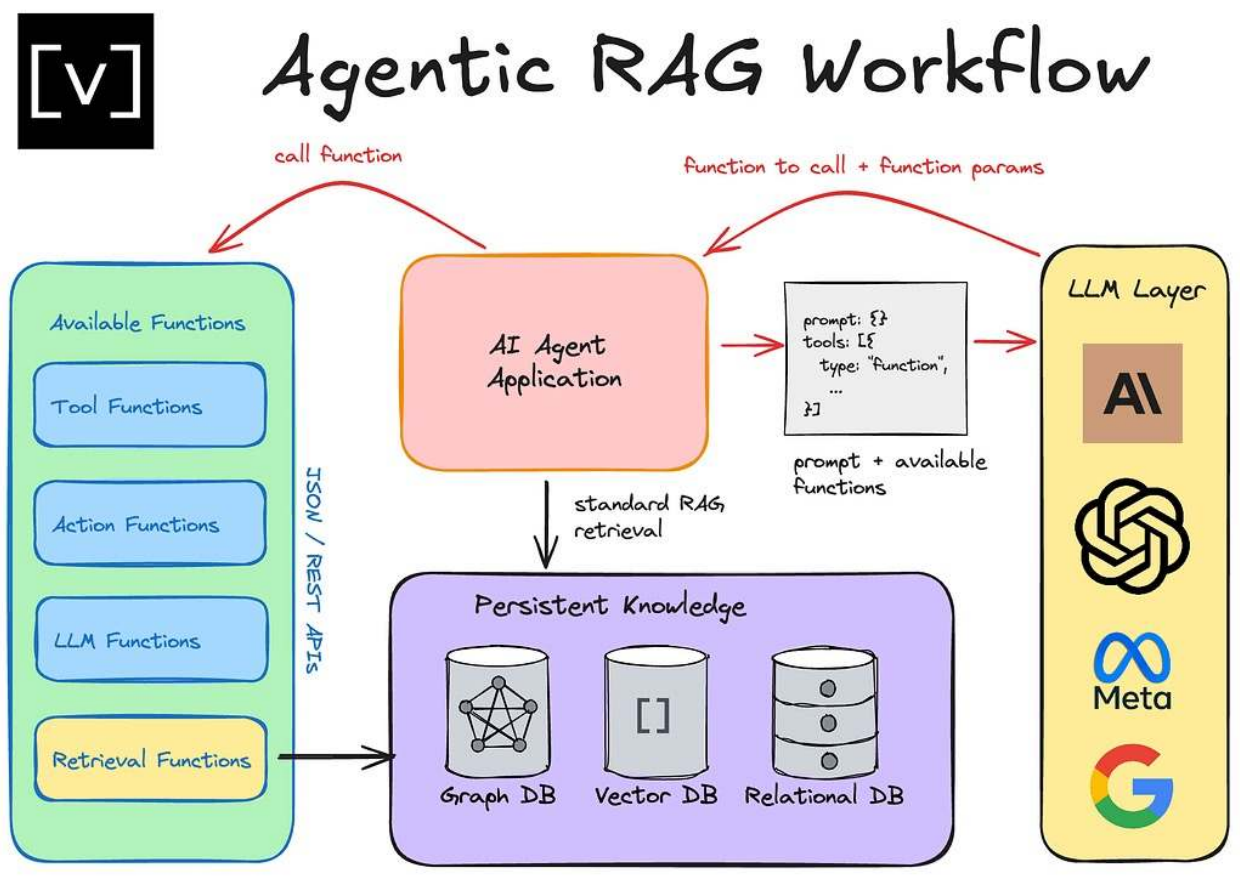
提升用户粘性，促进用户活跃，通过精准的广告推荐提升整体的收入

- 社交网络的场景，对安全的要求非常高，所以我们满足了 mTLS 和 Certificate 的证书每6小时更换一下
- 控制运维成本，目前只需要 2 个工程师即可维护整个 NebulaGraph 的相关产品
- 满足了高 QPS 的需求，包括 90% 以上的延迟可以达到 100ms-150ms 以内，且高可用达到 99.99%

Part 02
技术路线

NebulaGraph 的 GenAI 技术路线和产品

GraphRAG – Agentic Workflow

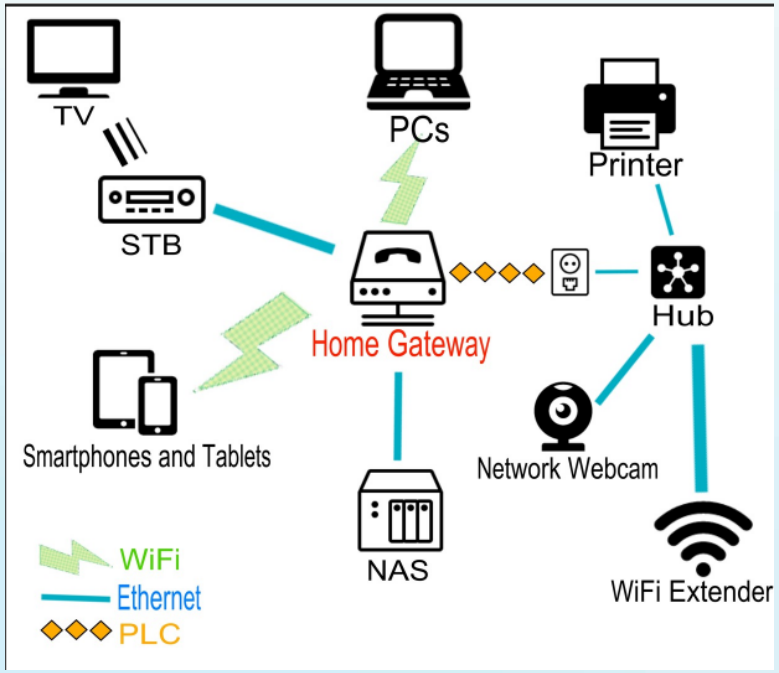


- 图源：https://vectorize.io/how-i-finally-got-agentic-rag-to-work-right/

- Retrieval
- Agent
- Ground Truth

GraphRAG – ParseCraft

- VLM 优先，客户需求驱动
- 专注集成性和可定制性
- One Layer， Any Parser




```python
transform_json_layer = TransformLayer(format=ResultFormat.JSON)
transform_markdown_layer = TransformLayer(format=ResultFormat.MARKDOWN)
metrics_layer = MetricsLayer()
otel_tracing_layer = TracingLayer()

parser = (
    ParseCraft(parser=vlm_parser)
    .layer(otel_tracing_layer)
    .layer(metrics_layer)
    .layer(transform_json_layer)
    .layer(transform_markdown_layer)
)

test_file = test_resources / "sample.fault_tree.pdf"

result = await parser.aparse(
    file=test_file,
)

print(result)

transformed_result = await transform_json_layer.get_transformed_result(result)
print(json.dumps(transformed_result, indent=4, ensure_ascii=False))

transformed_result = await transform_markdown_layer.get_transformed_result(result)
print(transformed_result)
```

接入 VLM 解析工具
提供格式转换、
可观测性等相关能力

GraphRAG – Deep Seach

推理过程

✓ 思考中

✓ 推理中

1. Use "harry_potter_knowledge_search" with query "哈利波特与伏地魔的关系背景、冲突事件和最终结局" to retrieve core narrative information.
2. Analyze retrieved content to identify key relationship aspects: blood connection through Harry's scar, prophecy-driven conflict origins, and Horcrux destruction as resolution mechanism.
3. Structure findings into chronological narrative covering lineage ties (Peverell ancestry), ideological opposition (Order vs Death Eaters), and final duel at Hogwarts Battle with Elder Wand mechanics.

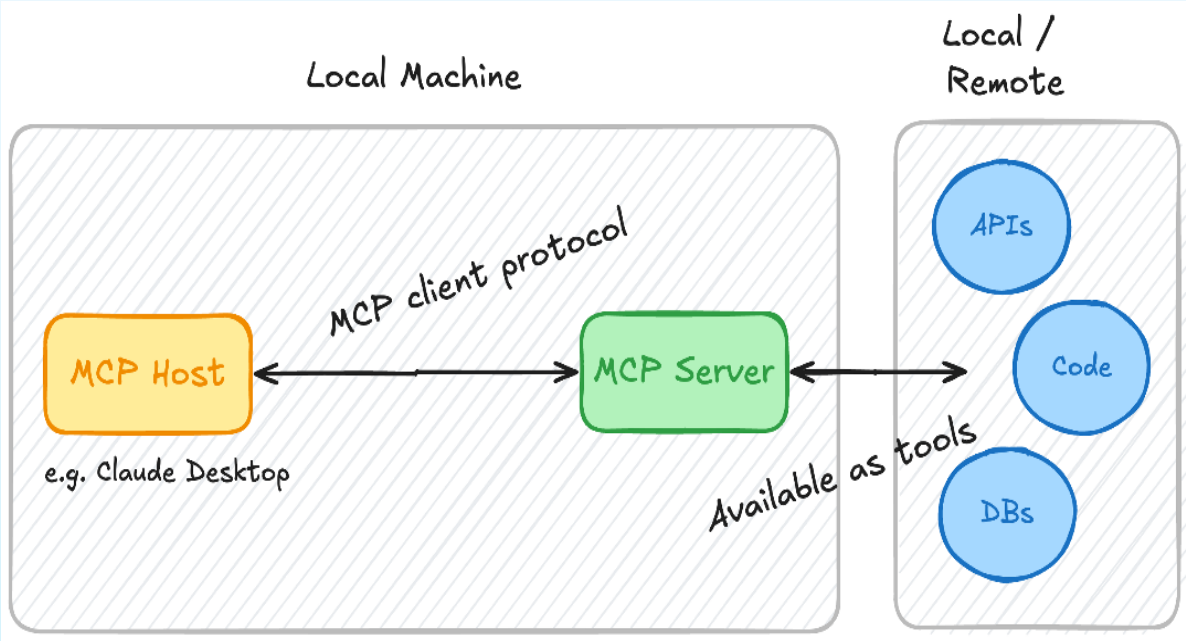
✓ 知识探索

- Harry Potter Knowledge Search

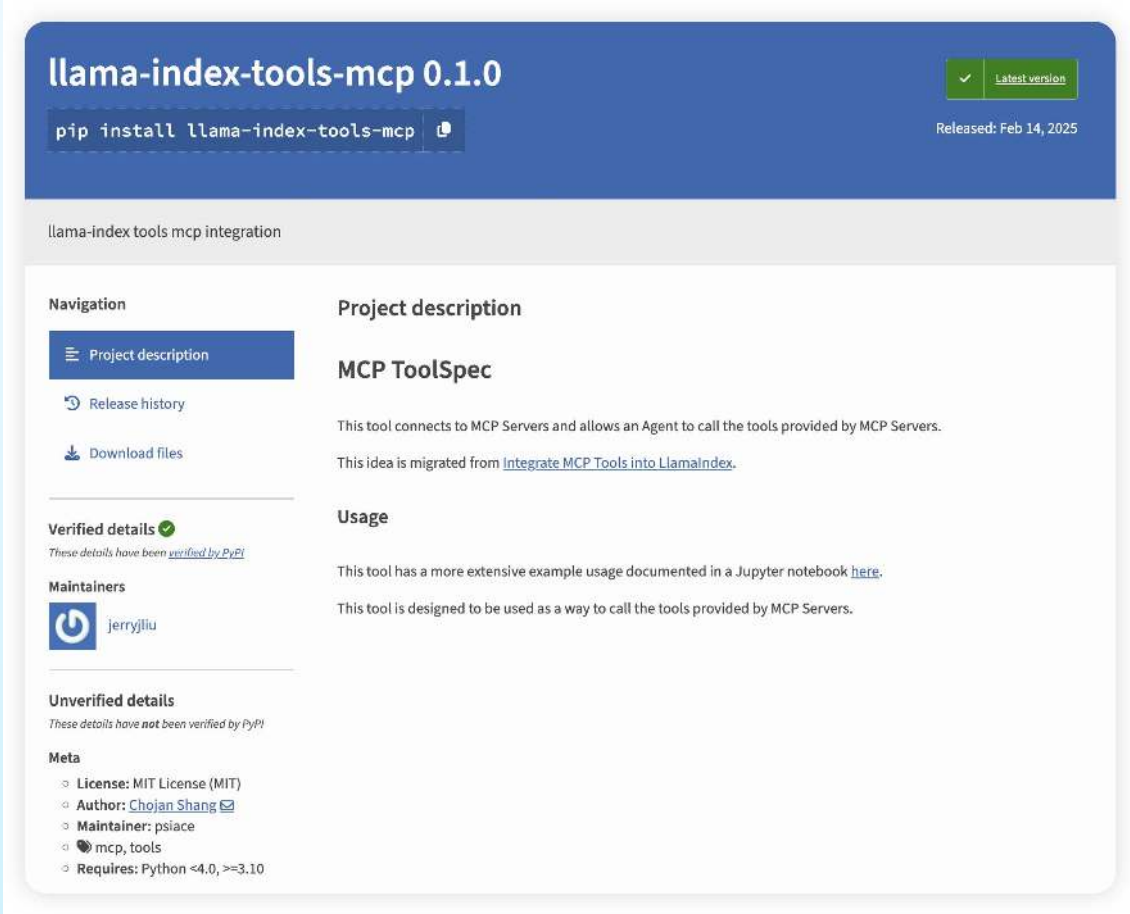
- 利用推理模型规划（DeepSeek R1 / Qwen QwQ）
- 融合 GraphRAG 能力进行充分洞察
- 节约开销、增强效果

GraphRAG – All in MCP

- 上游优先：贡献了 LlamaIndex 社区 McpToolSpec ，官方 MCP 支持层
- 内部改造优化：率先落地 Local MCP 范式

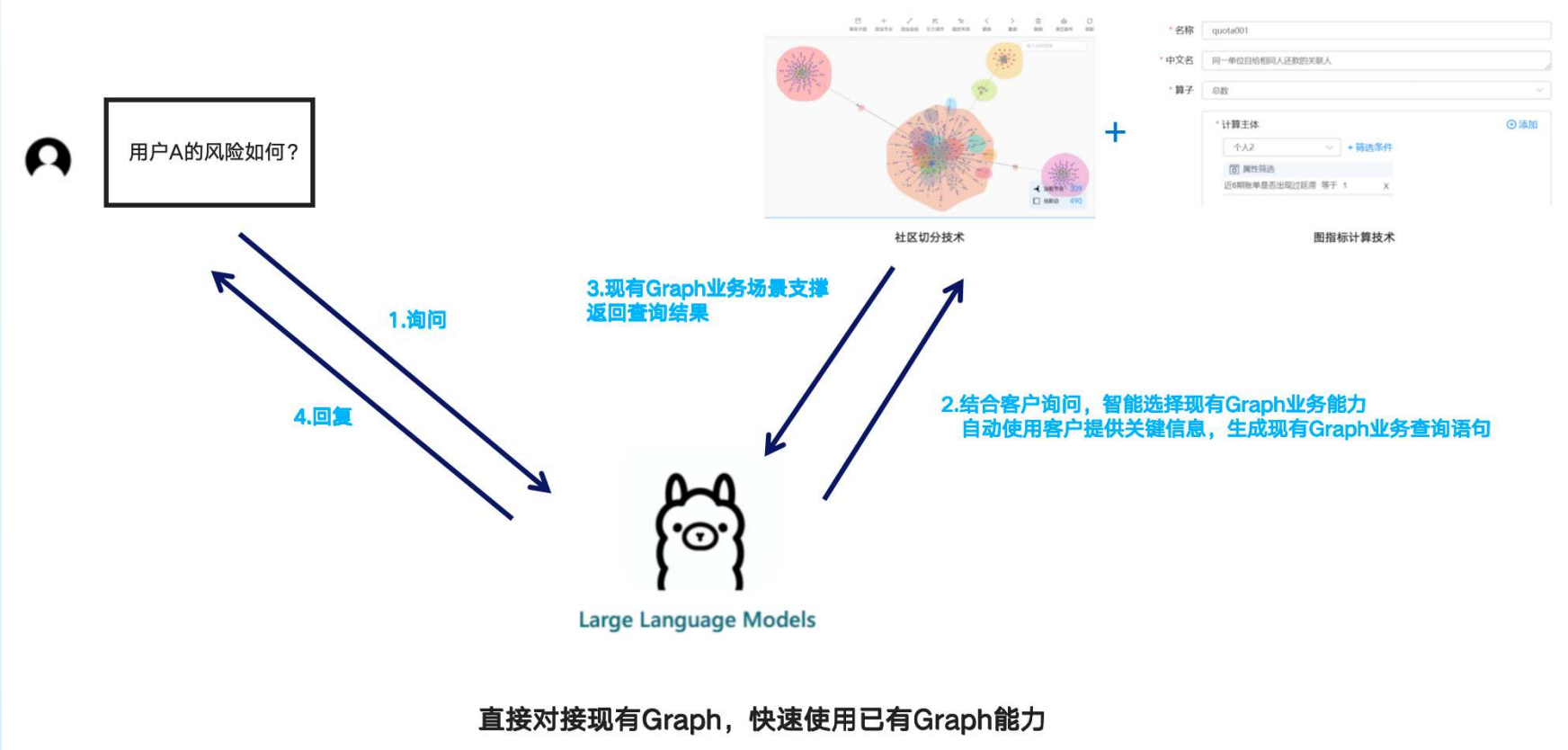


图源：https://hackteam.io/blog/build-your-first-mcp-server-with-typescript-in-under-10-minutes/

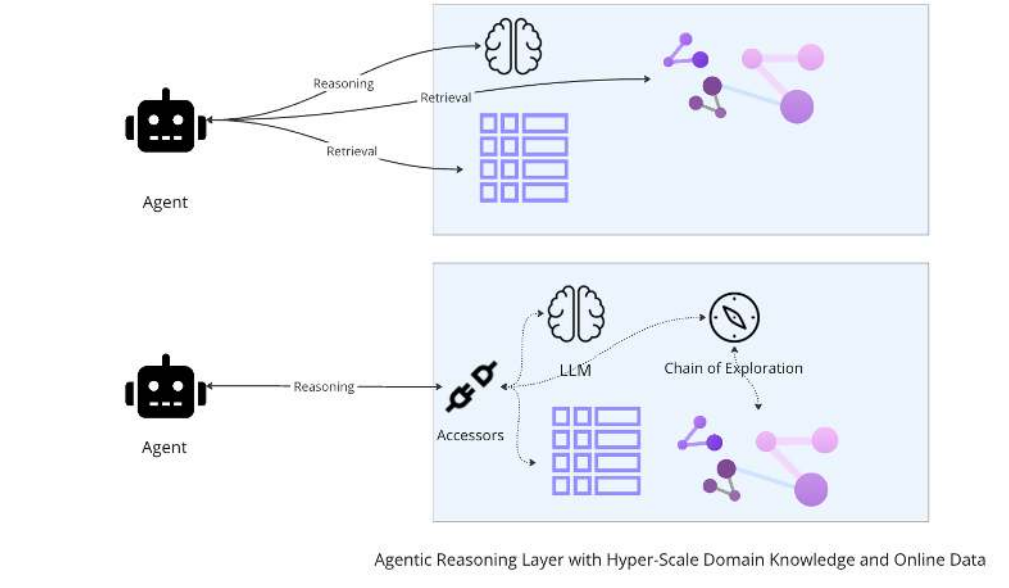


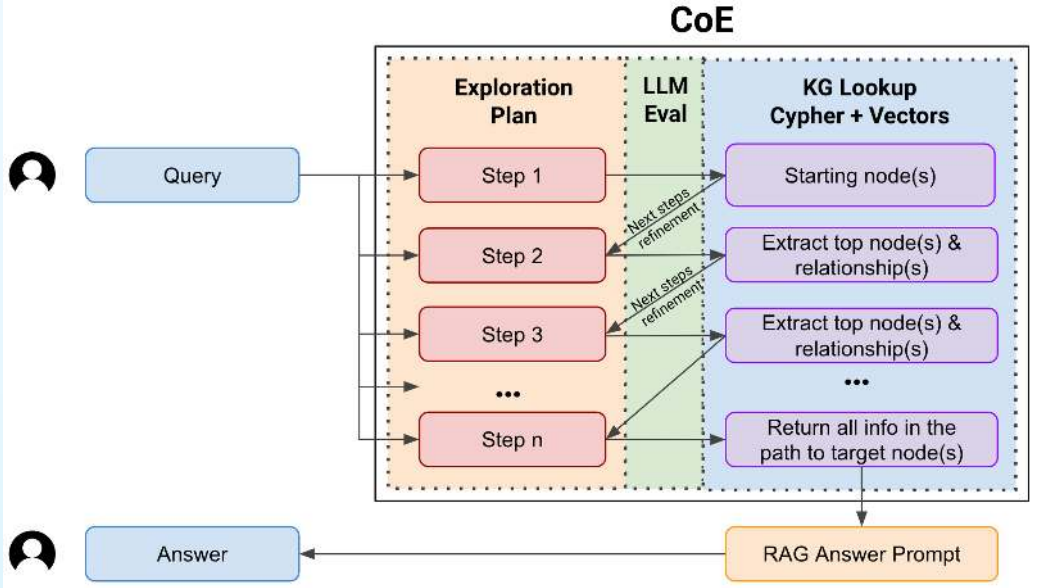
Graph Insight – Text to GQL

- 对接已有图谱，用户可以直接以问答形式进行图查询和图计算
- 基于规则/AST/算子模板的查询生成和校准
- Agent 自省



Graph Insight – Agentic & CoE

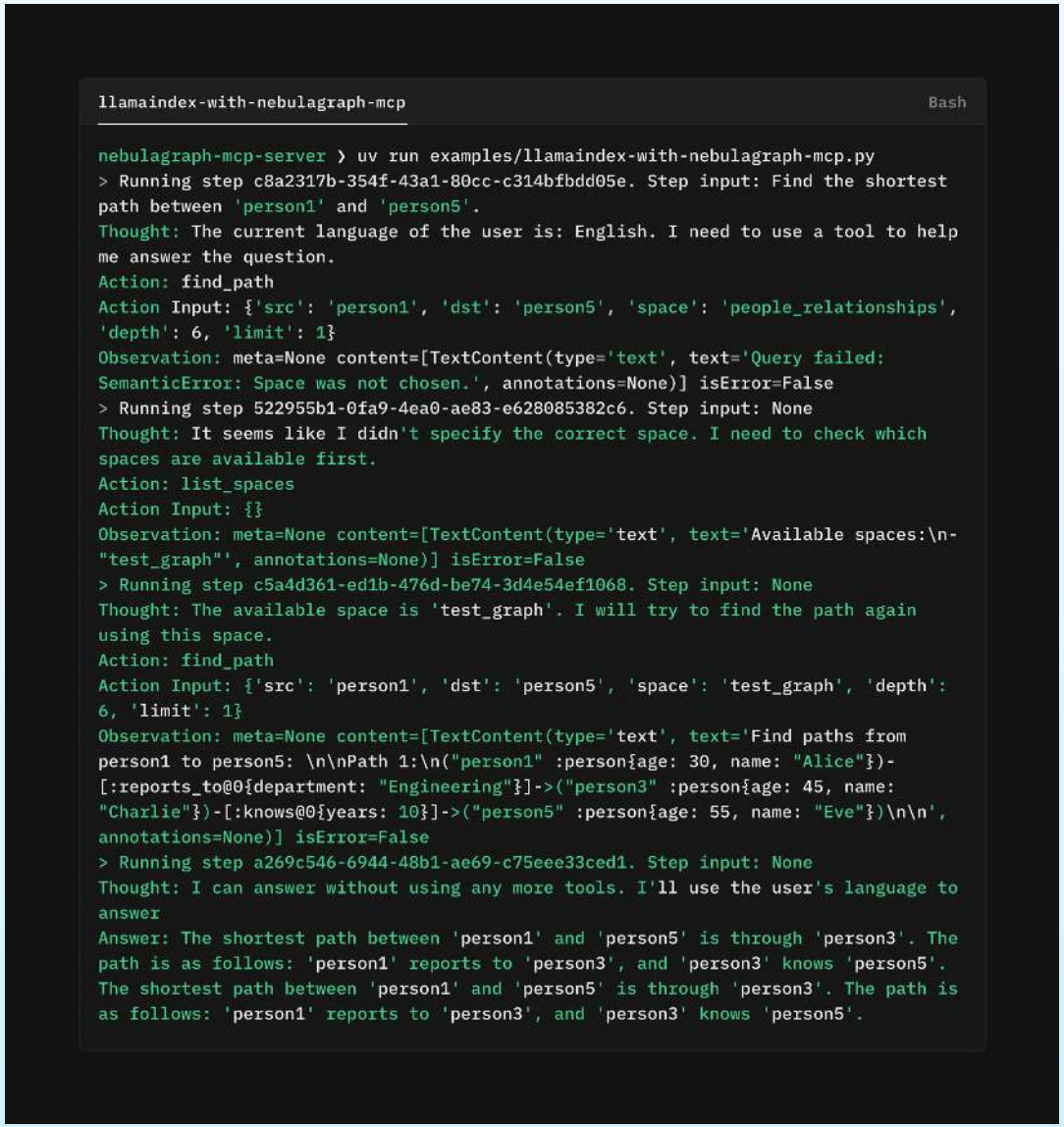




- 基于 Agentic 范式的图探索能力
- **Chain of Exploration**

Graph Insight - MCP

- 提供 NebulaGraph MCP Server
- 预先封装算子模版，如邻居发现，路径发现等
- 支持接入已有算法模板



Part 03
应用案例

NebulaGraph 的 GenAI 技术如何发挥作用

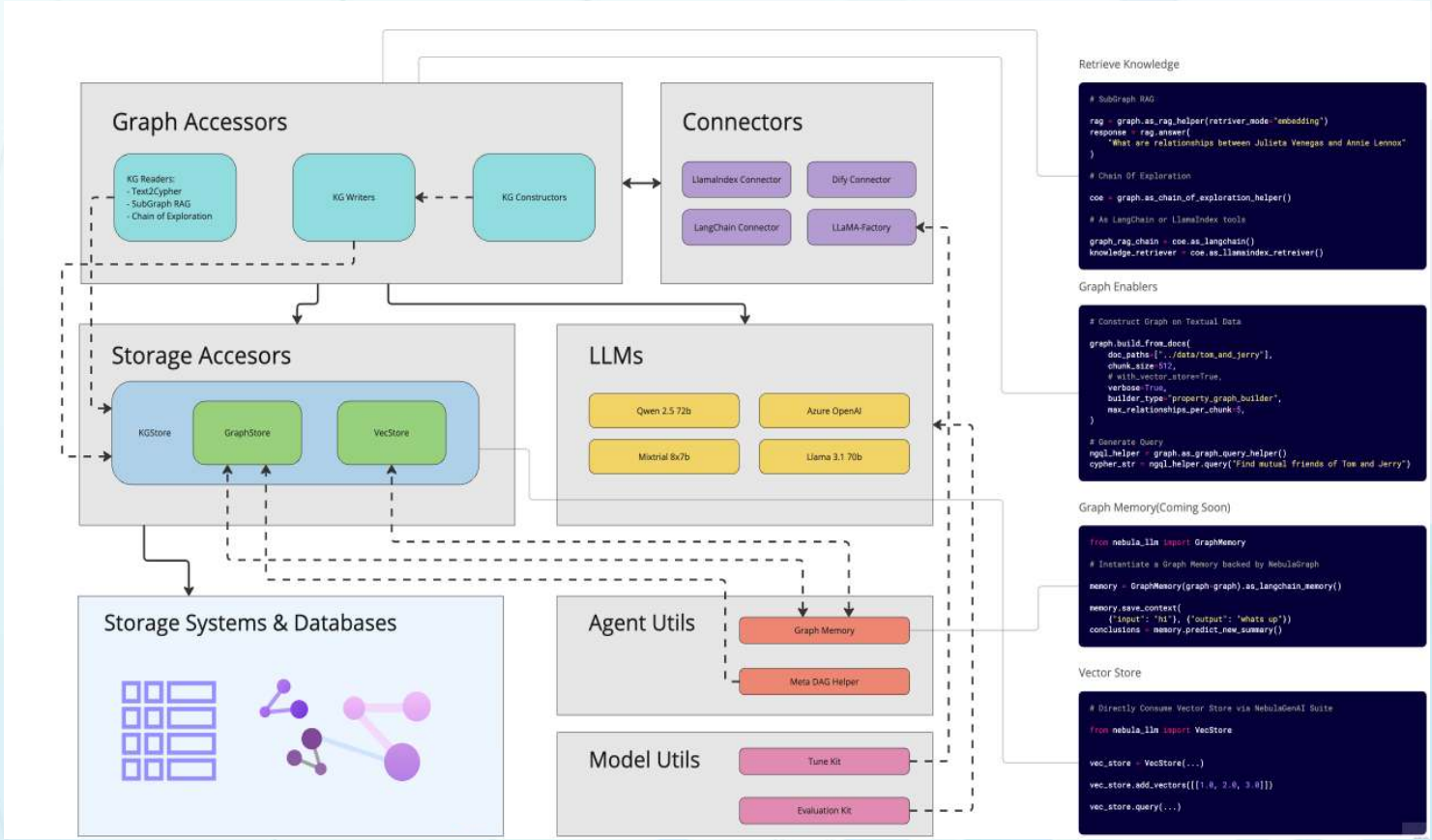
产品 - 图 AI 工具链

面向开发者的一整套工具链，
方便企业自行开发 AI 应用

- 封装 GraphRAG 核心技术
- 屏蔽复杂的 AI 应用开发的技术细节

核心功能

GraphRAG	图谱构建
图谱推理	模型微调



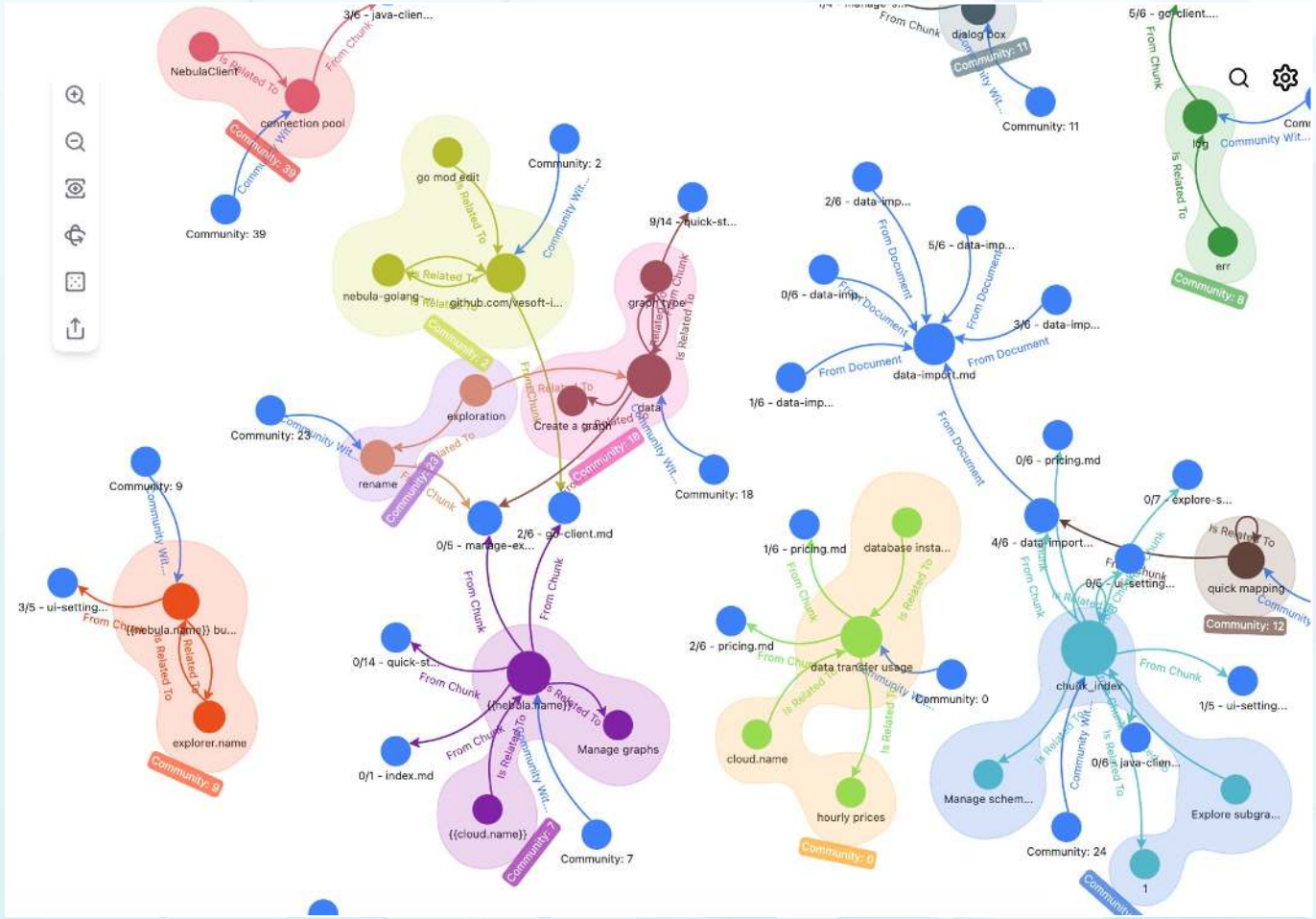
产品 – 图 AI 应用平台

面向终端用户，开箱即用，帮助用户以对话方式快速构建 AI 应用：

- 智能客服机器人
- 故障排查助手
- 研报生成机器人
- ……

特点

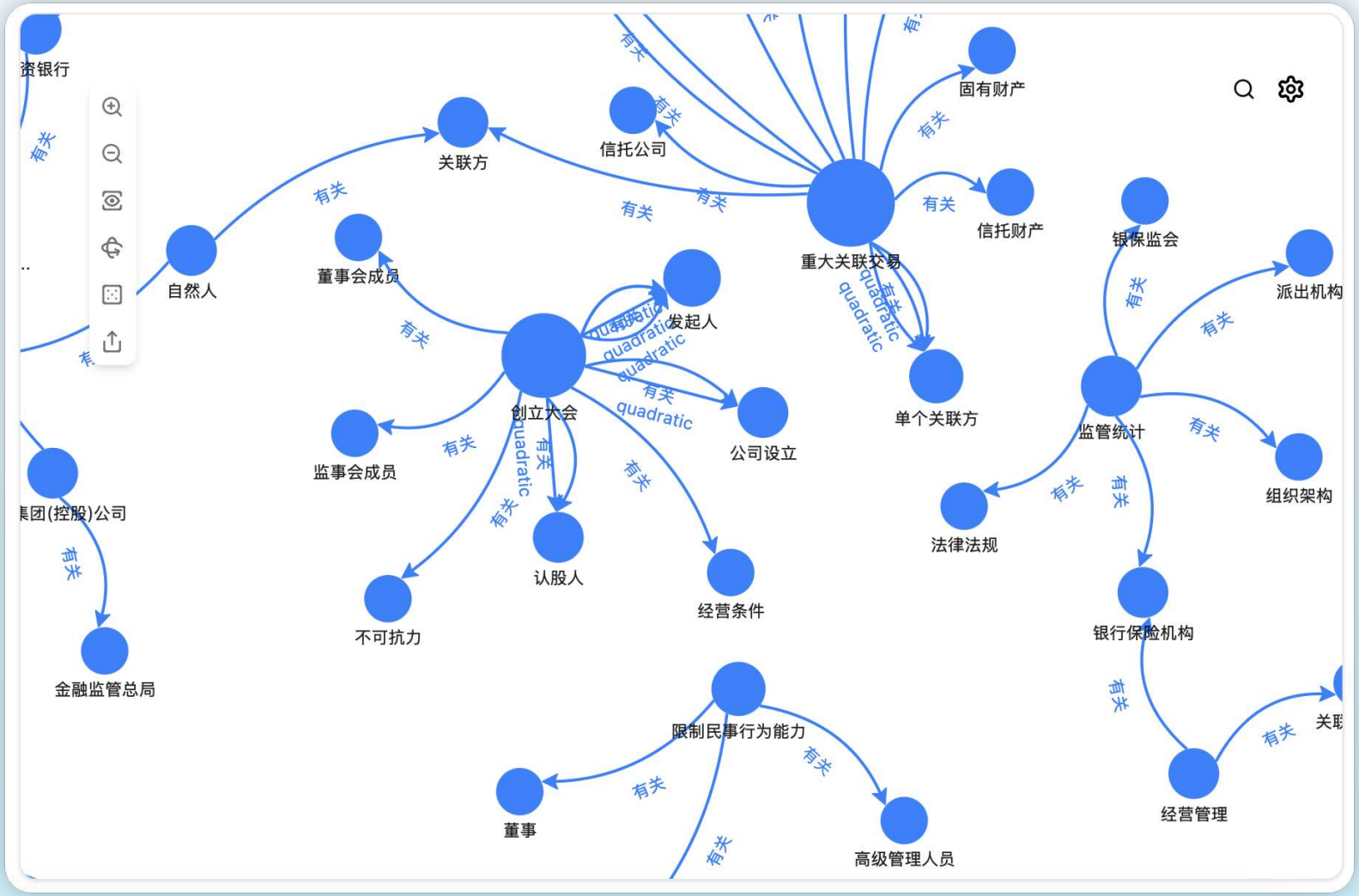
智能文档处理	**零代码**
零工作流	**零提示词设计**



工业解决方案：基于图和LLM的动态排障系统

- 基于LLM自动抽取知识：构建面向工业协同研发系统的知识图谱
- 基于知识图数据库及智能问答系统进行数据交互
- 落地行业首个生成式人工智能驱动的实际应用场景，荣获沙丘社区2024最佳案例15强
- 和行业头部企业一同协作，验证在复杂排障知识图谱上的图探索+大模型辅助系统能力，超过40万有效节点

金融行业解决方案



- 运用知识图谱自动抽取技术，基于过去和行业客户一同积累的经验洞见
- 行业伙伴共同构建和发布了基于 GraphRAG 和 Agentic Workflow 的行业解决方案

Thanks.