51CTO WOT

World Of Tech 2024

WOT全球技术创新大会

2024

智启新纪
慧创万物

51CTO WOT

LLM-based Agent在B端商业化的技术探索与实践

欧迪佐
快手-高级技术专家

分享目录

1、大模型应用建设背景

1. 快手商业化B端业务场景
2. 大模型应用技术方向选择

1.1 快手商业化B端业务场景

- 销帮帮智能客服（面向一线销售与运营人员）
- 代理商智能客服（面向代理商）
- 服务商智能客服（面向服务商）
- 广告主增长：线索分配、销售话术、外呼对话分析....
- CRM销售过程：企微对话分析、日报、拜访...
- 数据分析（few-shot）
- ……

观点：面向销售、代理商和广告主的服务中，有丰富的智能化需求和场景

1.2 大模型应用技术方向选择

1

[AIGC]
文生文/文生图
文生视频/数字人

2

[RAG]
检索增强生成
知识助手

3

[Agent/Agentic]
智能体
ToolUse
数据分析

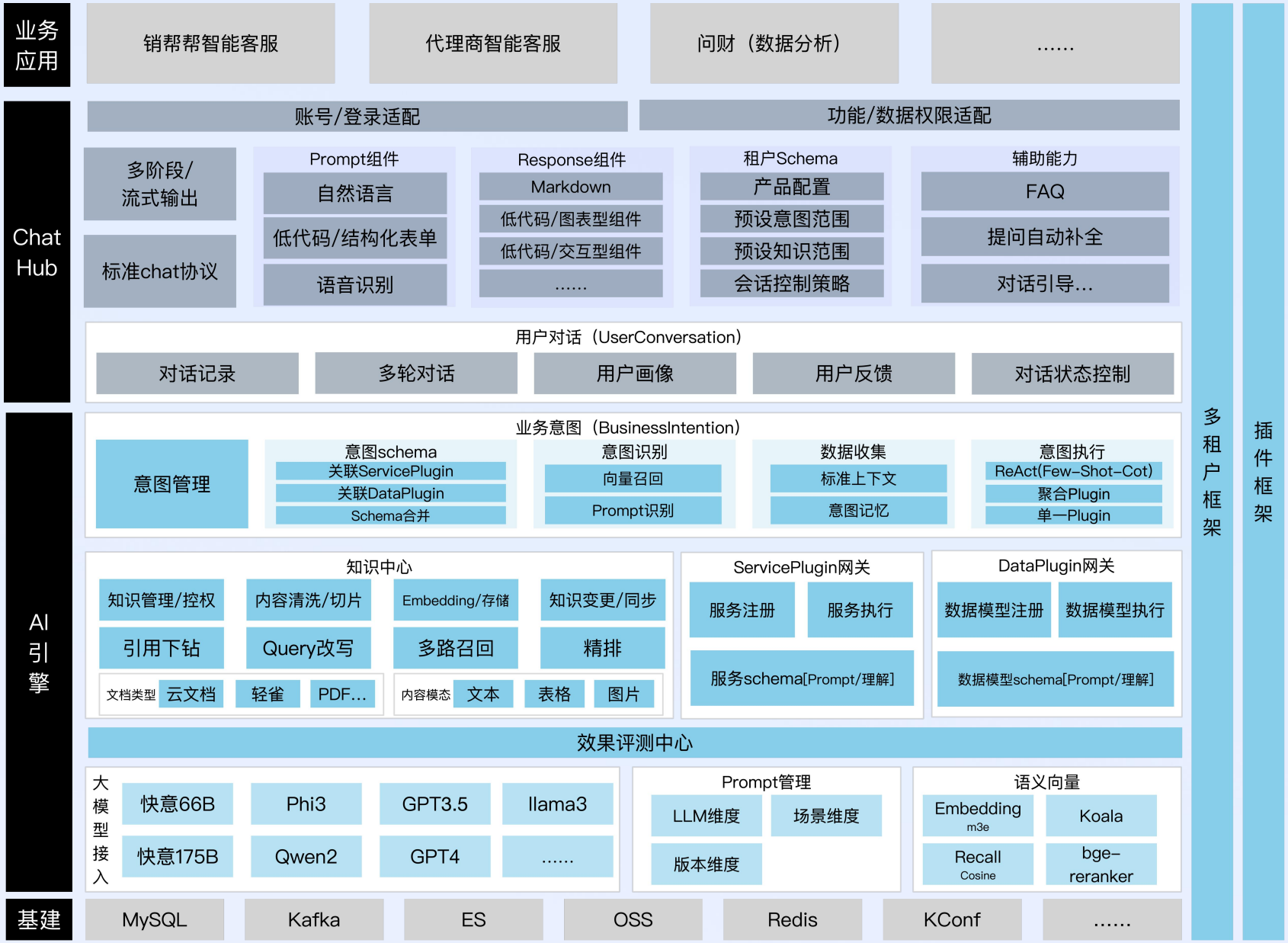
4

[垂直领域]
自动驾驶
人形机器人
.....

2、SalesCopilot技术平台

1. SalesCopilot系统架构
2. RAG实践
3. Agent实践
4. LLM相关方案
5. 效果评测中心

2.1 SalesCopilot系统架构



整体是“**三横一纵**”的系统结构

三横：

1. AI引擎，包含知识中心（RAG）、业务意图（Agent）、效果评测、大模型接入、语义向量等
2. ChatHub，智能客服核心抽象与运行框架，支持业务复用与个性化
3. 业务应用，平台之上长出来的若干应用（对应租户），归属于业务方

一纵：多租户与插件框架，这是平台化基础框架，前者支持业务接入与隔离，后者支持业务个性化

注：后面内容仅围绕AI引擎展开

2.2 RAG实践

1. LLM的优势与局限
2. RAG系统架构
3. 销帮帮若干案例
4. 销帮帮业务指标
5. 业务实践中的挑战

2.2.1 LLM的优势与局限

优势

- **常识理解与初级逻辑推理**：能够实现与人类初级水平相当的能力，高水平理解和生成文本。这使得它们在摘要、翻译、问答、意图识别、情感分析等任务中表现出色。
- **广泛的适用性**：大模型在预训练阶段接受了大量多样化的数据，这使得它们可以适应不同领域的任务需求（涌现）。从金融到医疗，从娱乐到教育，都能提供解决方案

局限性

1 **幻觉**

模型可能产生无根据的错误输出

2 **知识时效**

模型训练成本很高，周期性发布，难以即时更新最新信息

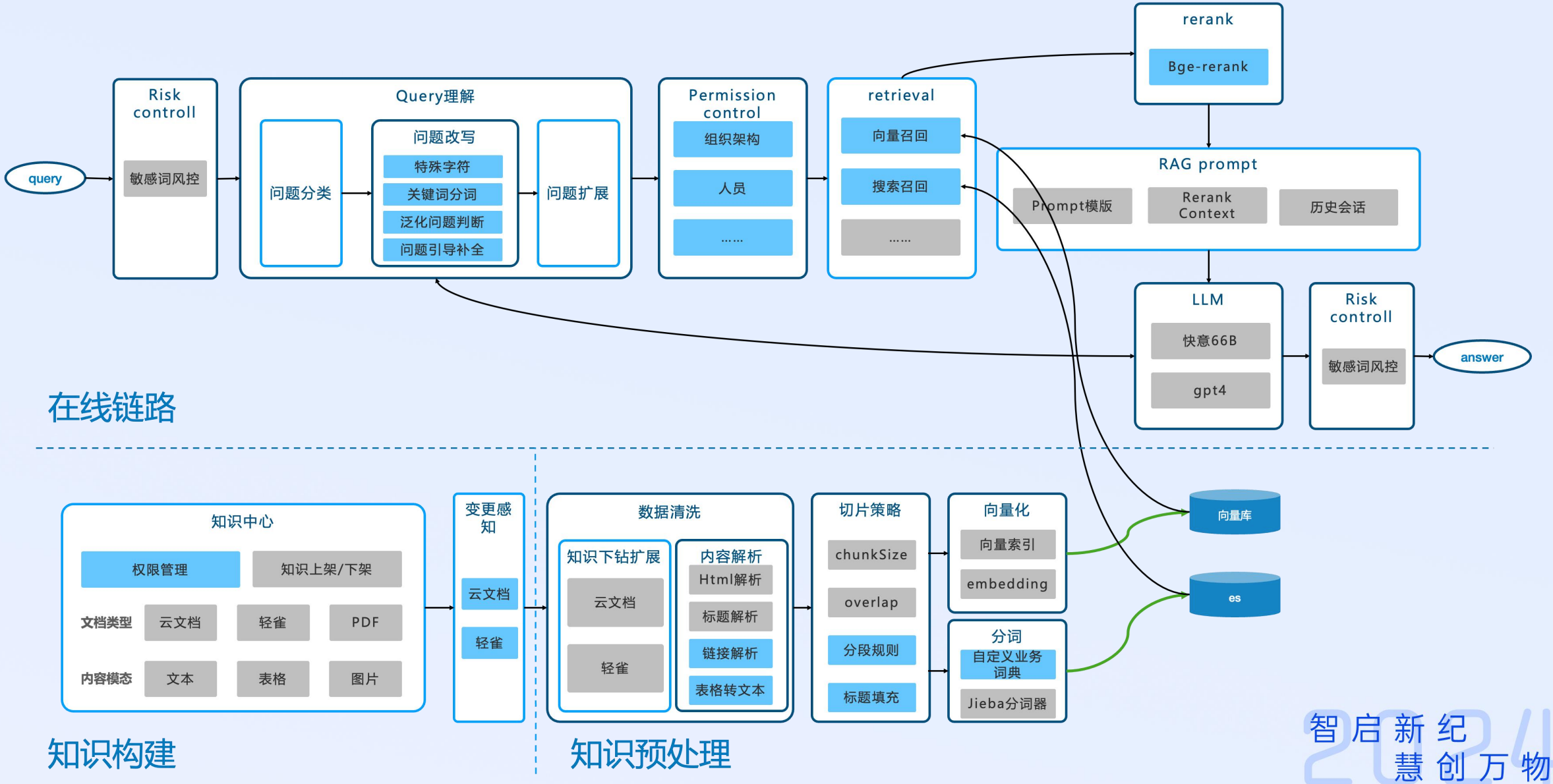
3 **记忆容量**

无状态chat服务，prompt长度有限，当前模型难以保留长记忆

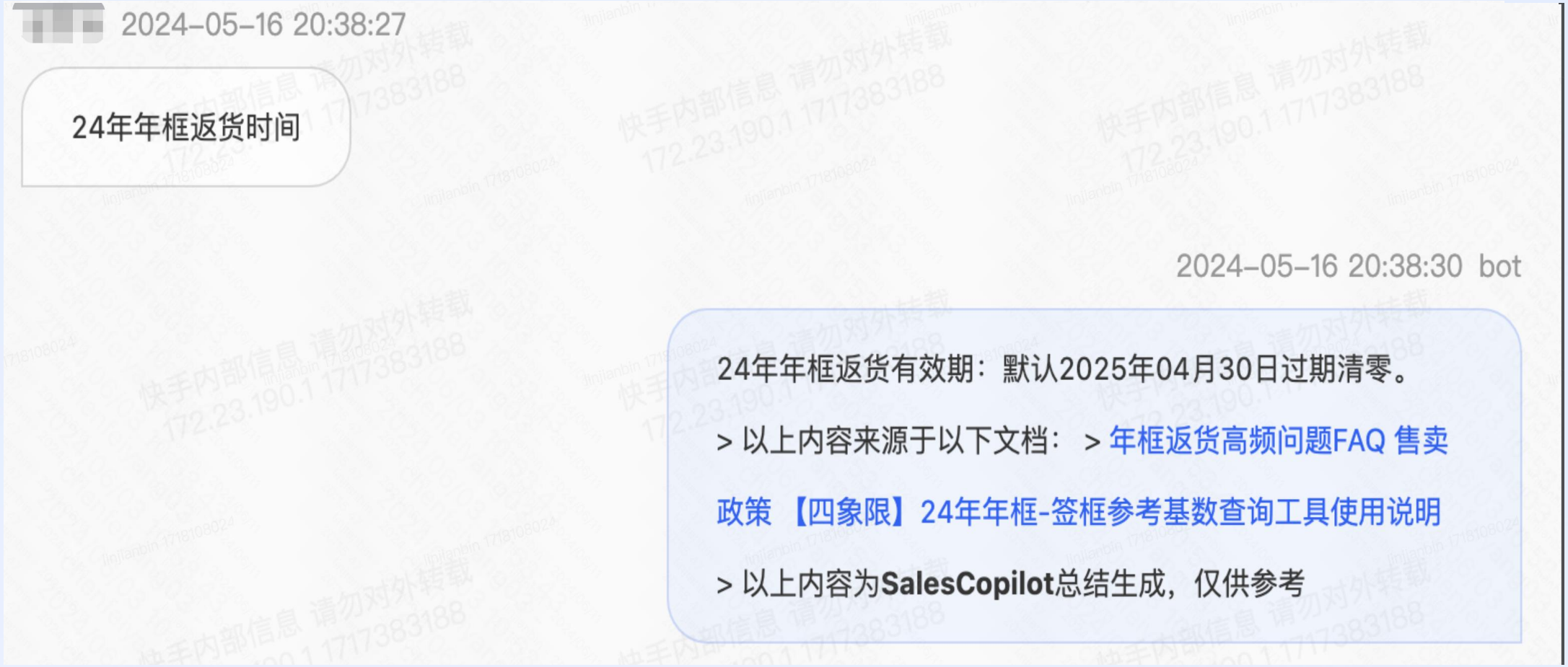
4 **数据安全**

私域数据参与训练存在安全隐患

2.2.2 RAG技术链路



2.2.3 销帮帮若干案例


2024-05-16 20:38:27
24年年框返货时间
2024-05-16 20:38:30 bot
24年年框返货有效期：默认2025年04月30日过期清零。
> 以上内容来源于以下文档： > 年框返货高频问题FAQ 售卖政策 【四象限】24年年框-签框参考基数查询工具使用说明
> 以上内容为SalesCopilot总结生成，仅供参考

2.2.4 销帮帮业务指标

- ✓ 覆盖商业化12个直客行业，销售人员覆盖率**35.6%**
- ✓ 知识库文档：**700+**(原始/34.2%、下钻/65.8%)
- ✓ 机器人拦截率：**78%**
- ✓ 多路召回和精排，在评测阶段，知识问答平均分从3.07提示至5.24，绝对值提升：2.17，百分比提升：**70.68%**

注：评测口径

1. *单条问答分数区间0-10，由大模型对问答结果打分，平均分计算公式：sum(单样本得分) / 样本总数。*
2. *样本范围取自FAQ、标注数据、线上问答*

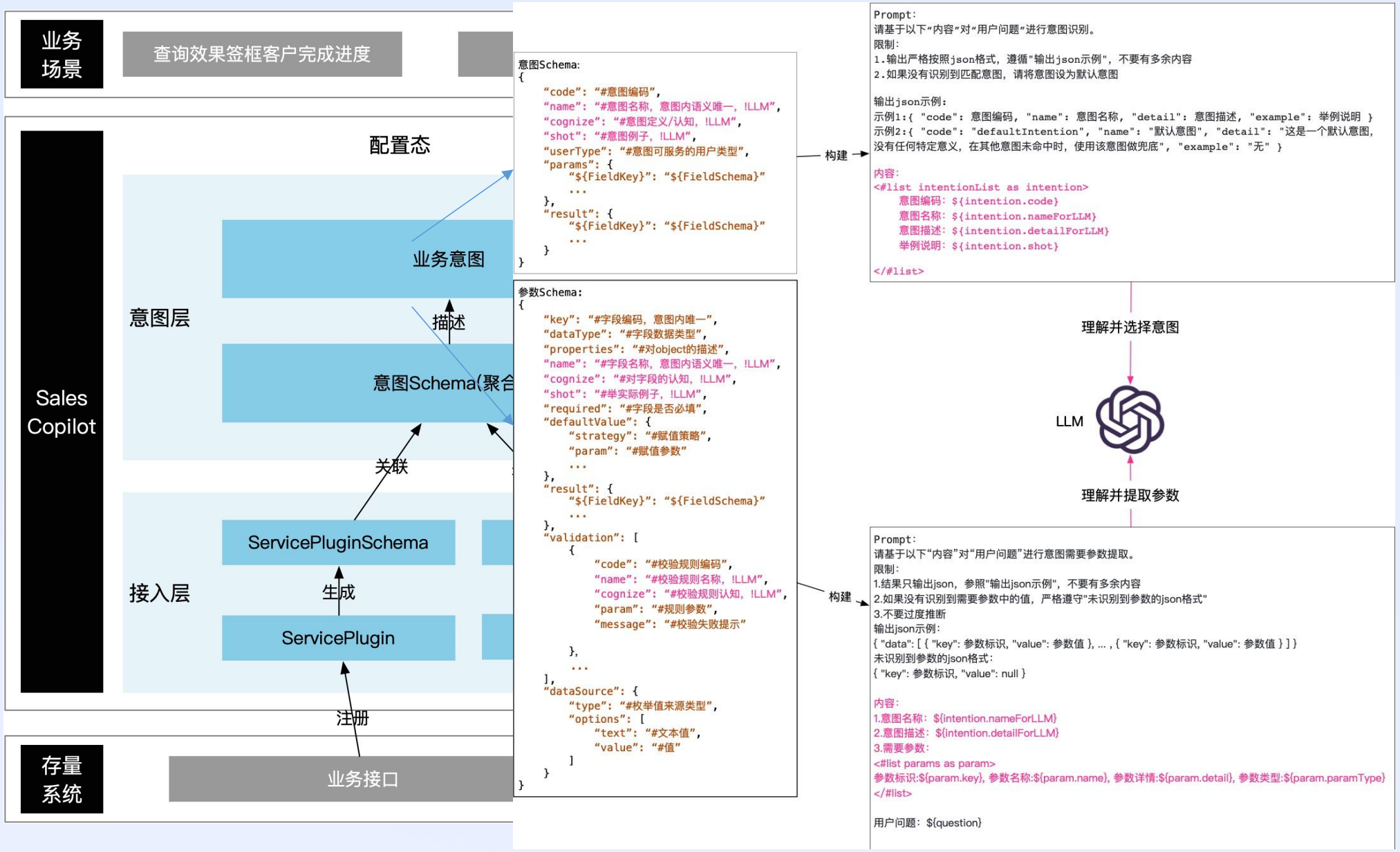
2.2.5 业务实践中的挑战

问题类型	问题定义	解决方向
RAG(4)	商业化场景下用户短问题/泛化问题的识别、处理和引导	*Query理解#问题分类* *Query理解#追问*
	漏召回和回答准确率问题，特别是知识内容规模增长带来相似语义干扰	*多路召回* *精排*
	数据安全与复杂的权限管控策略（商业化知识和业务数据）	*多租户知识中心* *商业化风控*
	无法理解专业领域黑话，以及标点符号对Query语义的影响	*Query理解#改写*
LLM(3)	Prompt和topK、topP、temperature，提示工程调优，按下葫芦浮起瓢	*建设评测工具（保障可控的效果提升）*
	大模型上下文长度有限，多轮对话时不能联系上文	*有限多轮*
	大模型幻觉和推理能力不足	*模型升级/替换*
用户需求(1)	多模态的输入，比如部分用户提问的方式是“截图+文字”，主要线上业务问题的咨询与排查，需要和业务系统和业务数据互动	*Query理解#多模态* *业务意图/Agent*

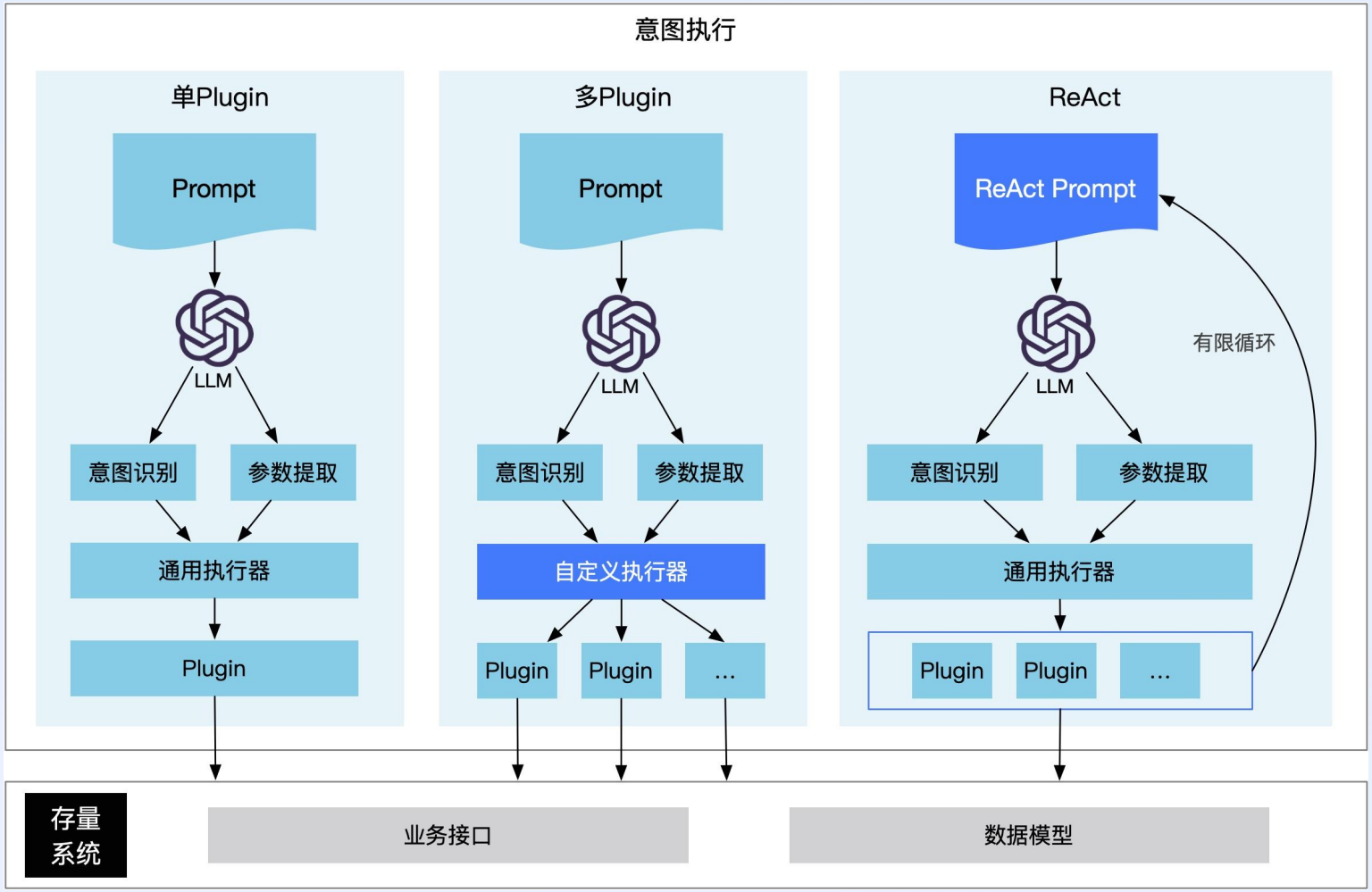
2.3 Agent实践

1. Agent-技术链路
2. Agent-意图执行
3. Agent-若干案例

2.3.1 Agent - 技术链路

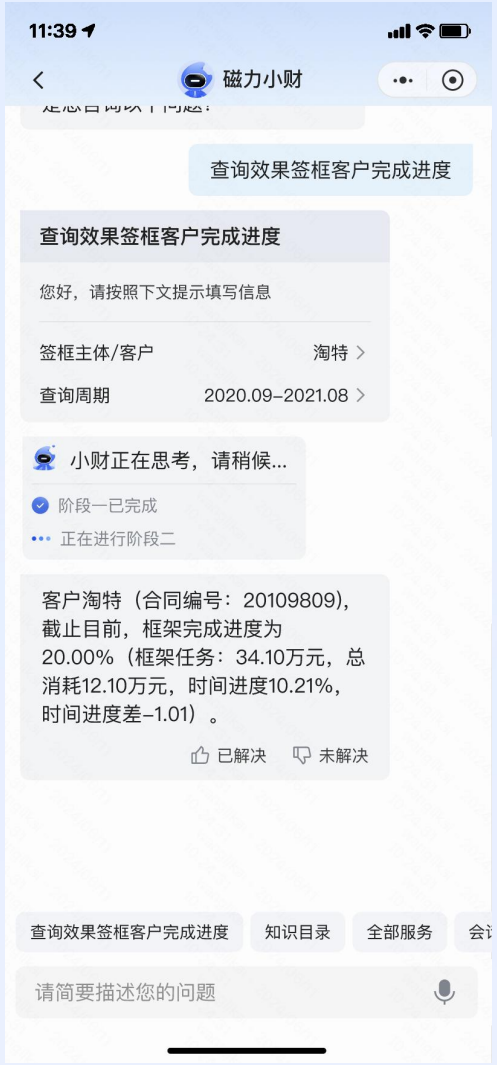


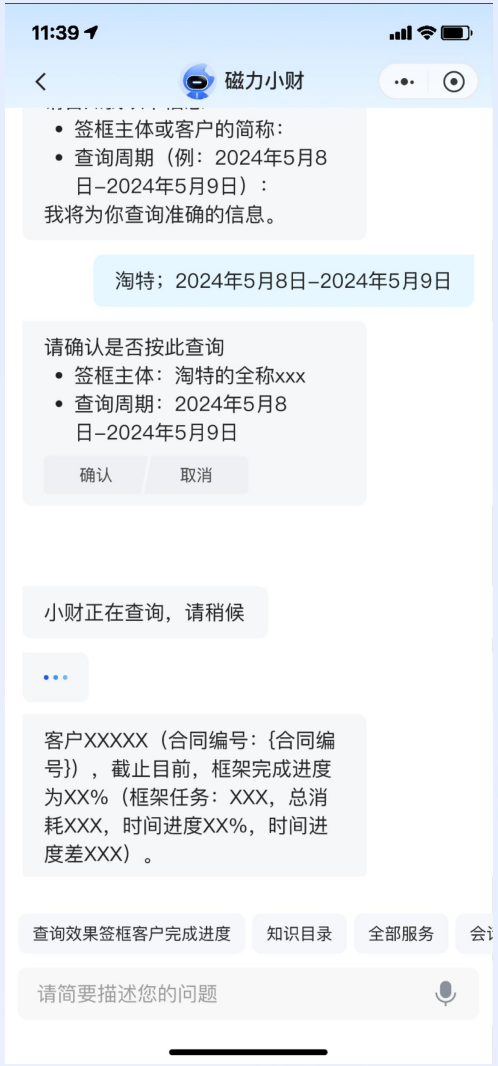
2.3.2 Agent - 意图执行

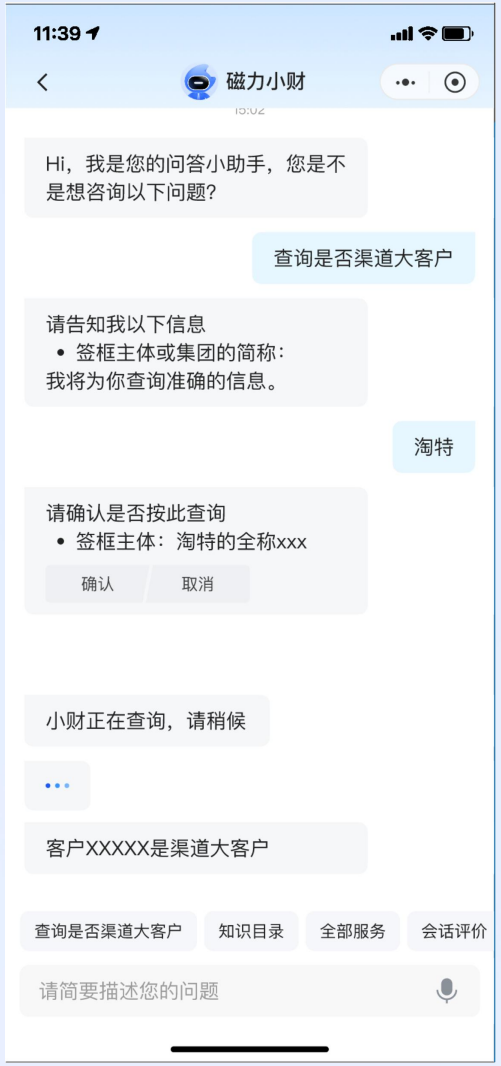


- 单Plugin
  - 使用通用执行器
  - 简单场景，无需额外开发
  - 流程确定性高
- 多Plugin
  - 需实现自定义执行器
  - 复杂场景，需要额外编码
  - 流程确定性高
- ReAct
  - 使用通用执行器
  - 复杂场景，依赖LLM能力
  - 流程确定性低

2.3.3 Agent实践 - 若干案例







2.4 LLM相关设计

可插拔

能根据需求快速替换或更新模型，支持多模型协作，让不同任务调用最适合的模型

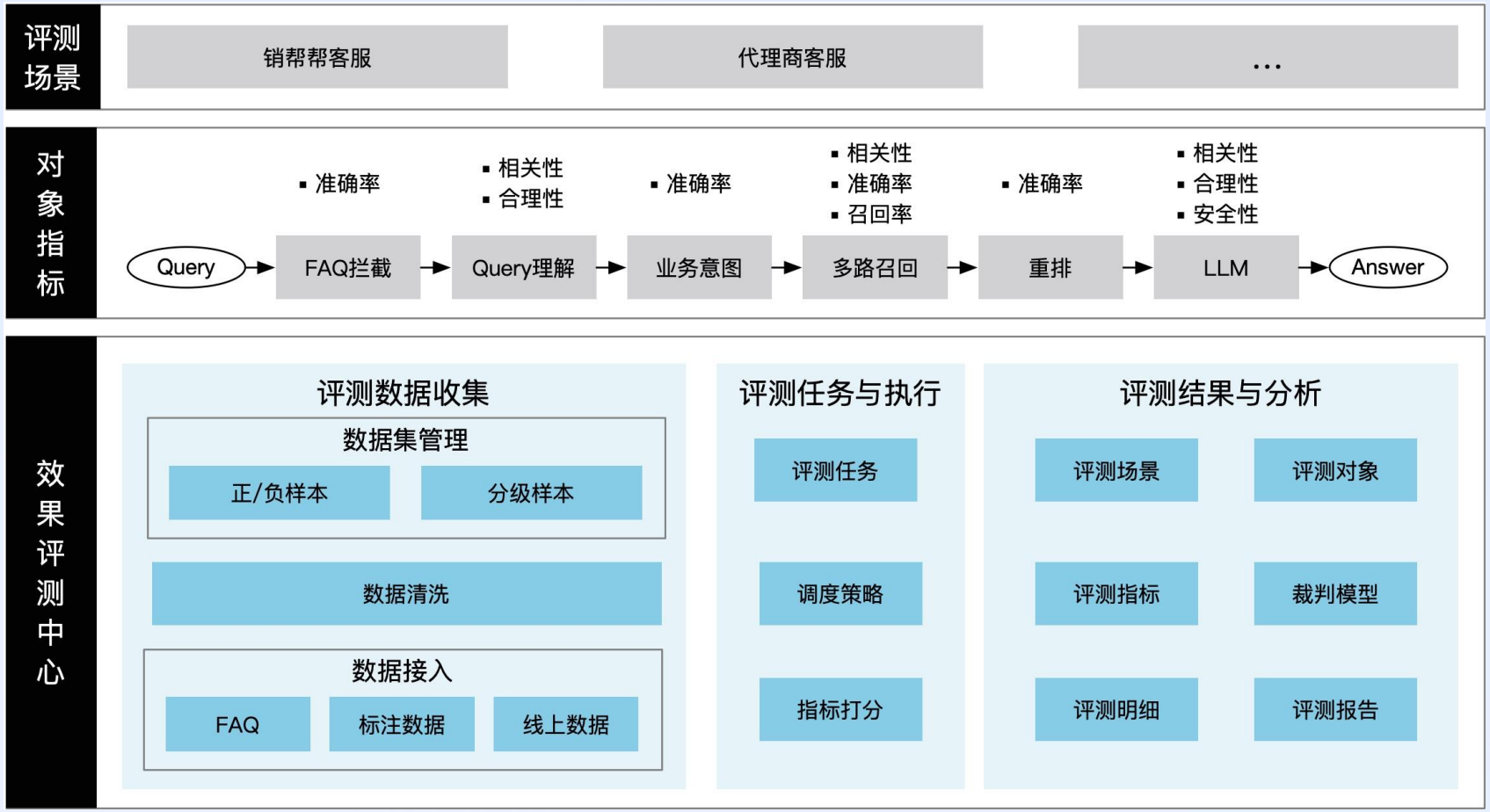
LSP

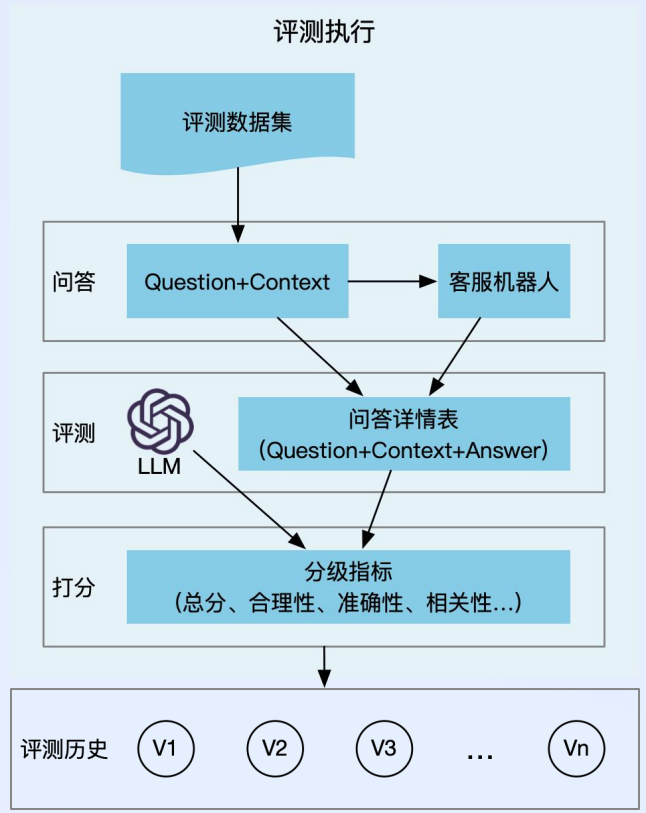
LLM Specific Prompt/模型专用提示，LLM各有调性，皆有适合自己的Prompt风格

量化LLM

量化大模型通过减少参数精度来降低资源需求，仅少量智能损失可跑高性能跑在CPU上

2.5 效果评测中心





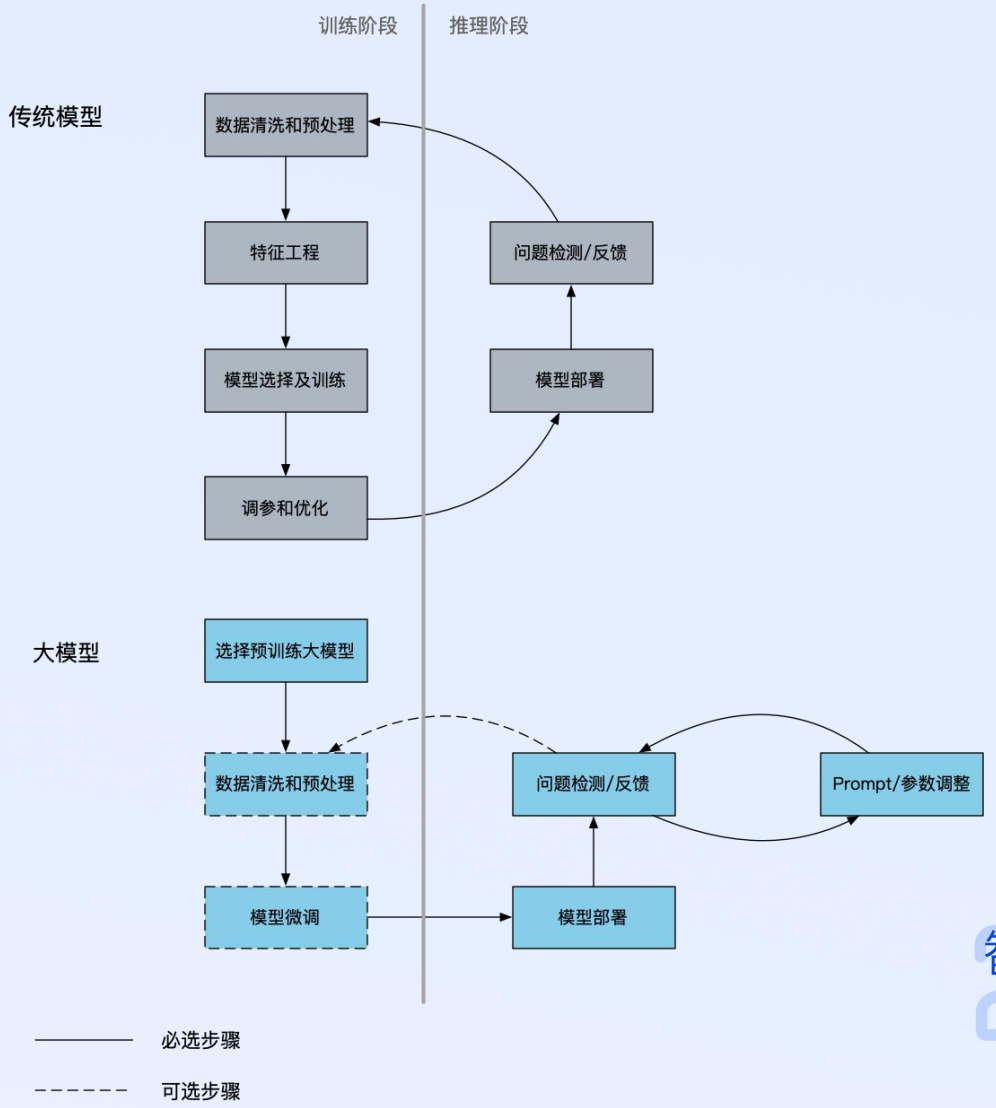
3、大模型应用研发的思考

1. 生产力：智能化技术平权
2. 路径选择：从垂直细分领域开始
3. 效果提升：乘积效应（RAG）
4. 需求趋势：多模态

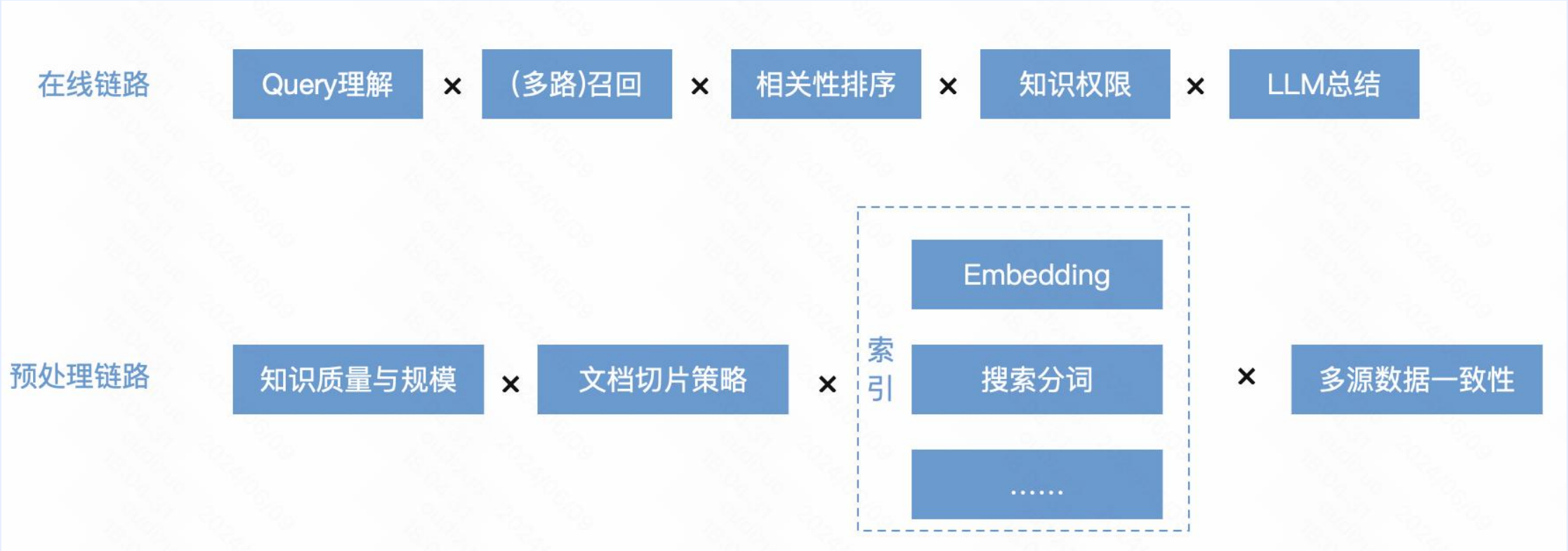
3.1 生产力：智能化技术平权

大模型通过提供先进的算法和庞大的数据处理能力，使得即使是资源较少的小企业或小团队也能利用顶级的AI技术进行产品开发和业务优化

- 例如，通过使用大模型，小公司可以实现与大公司相竞争的产品特性
- 例如，通过使用大模型，小团队就是快速做出智能化产品，不再强依赖大量算法人员



3.2 RAG效果提升：乘积效应（RAG）



RAG效果提升是一个非常系统性的工作，要做到比较好的效果，有非常多的智能化和工程策略的事情要做，没有银弹，要抓关键细节一个个去做实做深

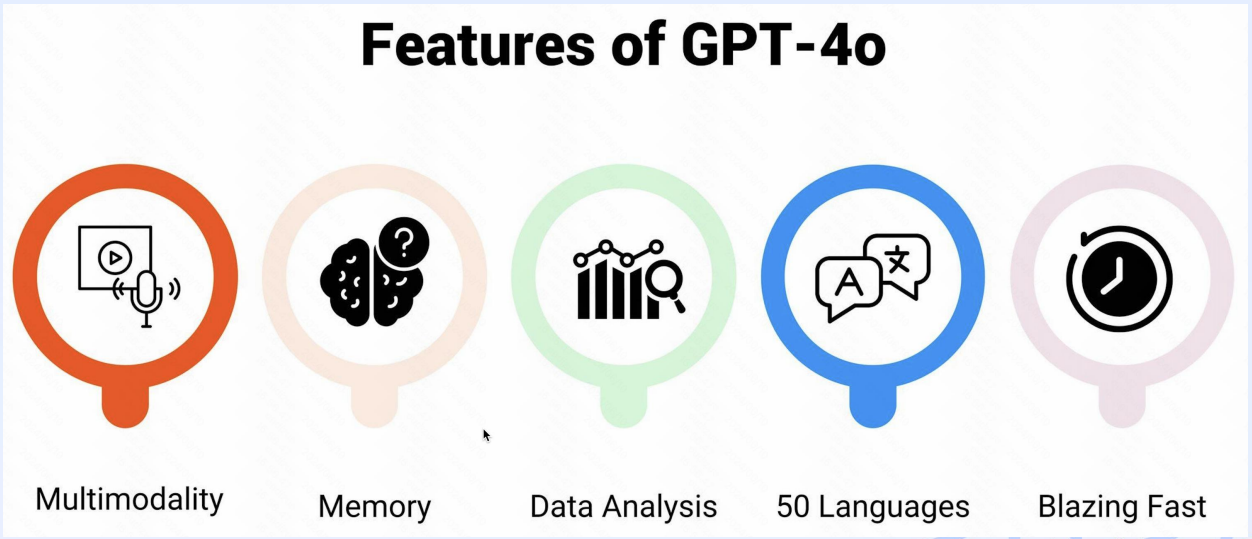
3.3 路径选择：从垂直细分领域开始

最佳路径：标杆应用验证 -> 效果提升范式（布局架构） -> 应用研发平台 -> 应用快速复制

对比维度	通用大Agent ×	垂直多Agent
【定义】	一个大模型Agent承担多种任务，覆盖广泛应用场景	多个专用大模型Agent，各自负责特定领域和任务
知识规模 知识召回	多类型和多源数据，难以统一数据预处理，Chunk整体质量提升困难；数据规模大，语义干扰多，影响召回率	专注于特定领域，数据相对集中，在数据预处理和多路召回上，可以叠加针对性策略，快速提升
意图识别	意图规模大、意图识别复杂度高，需要强大的上下文理解和推理能力	意图识别规模有限，更精准上下文简单，推理路径明确
Agent身份与职责	身份职责多样化，难以优化每个任务的表现决策逻辑复杂，易混淆	身份职责明确，优化更精准决策路径清晰，易于理解和维护
【优势】	广泛适用，覆盖更多用户需求资源集中，便于管理	专注于特定领域，性能更优快速迭代，快速验证拿反馈
【劣势】	功能过于泛化，难以做到精细化，系统复杂，难以提升效果，验证慢反馈慢	初期投入较高，需要构建多个Agent可能资源分散（SalesCopilot多租户能力）

3.4 需求趋势：多模态

- **自然交互**：多模态技术能够模拟人类自然交互，使人机互动更加顺畅
- **信息完整性**：整合多种模态提供更丰富的信息和上下文（千言万语不如一张图）
- **情感传递**：多模态技术可以通过语音语调、面部表情等方式更好地传递情感。
- **复杂任务处理**：支持更复杂的任务处理，提高系统准确性和可靠性。
- **人类五感对齐**：未来的发展方向是让终端设备更全面地对齐人类五感（视听触味嗅），实现逼真的交互体验



51CTO WOT

谢 谢 观 看

THANKS

智启新纪
慧创万物
2024

附：快手技术公众号