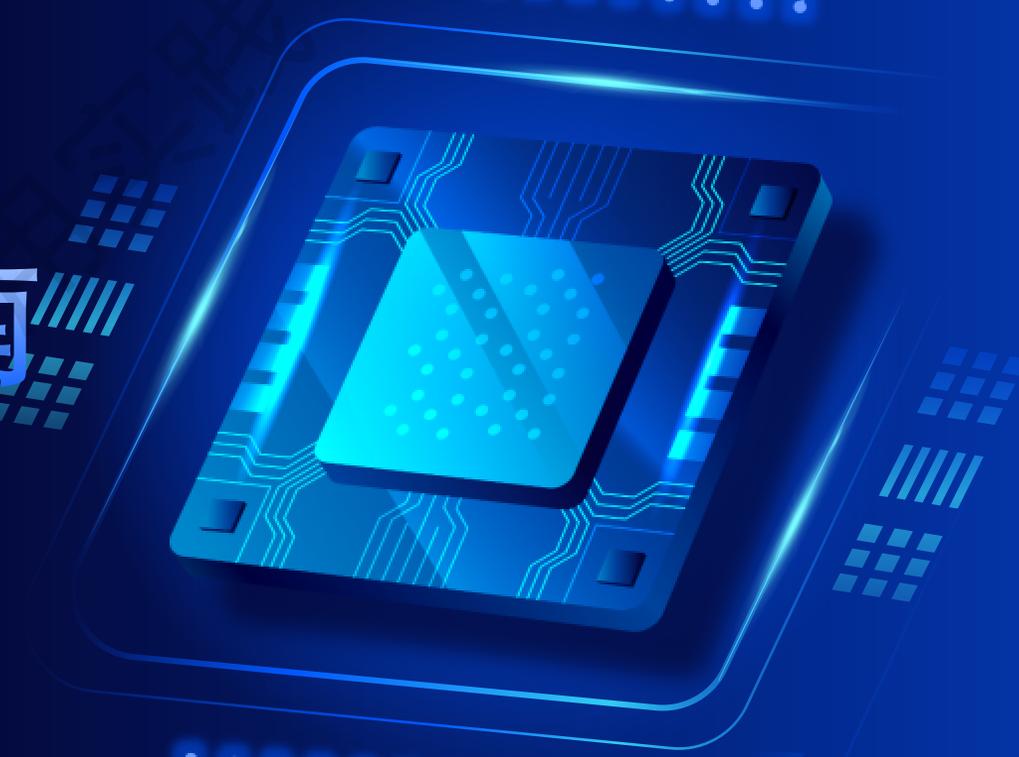


# 生成式AI与RAG： 产品经理实战指南

AI大模型应用实践



# CONTENTS 目录

01

生成式AI再认识

02

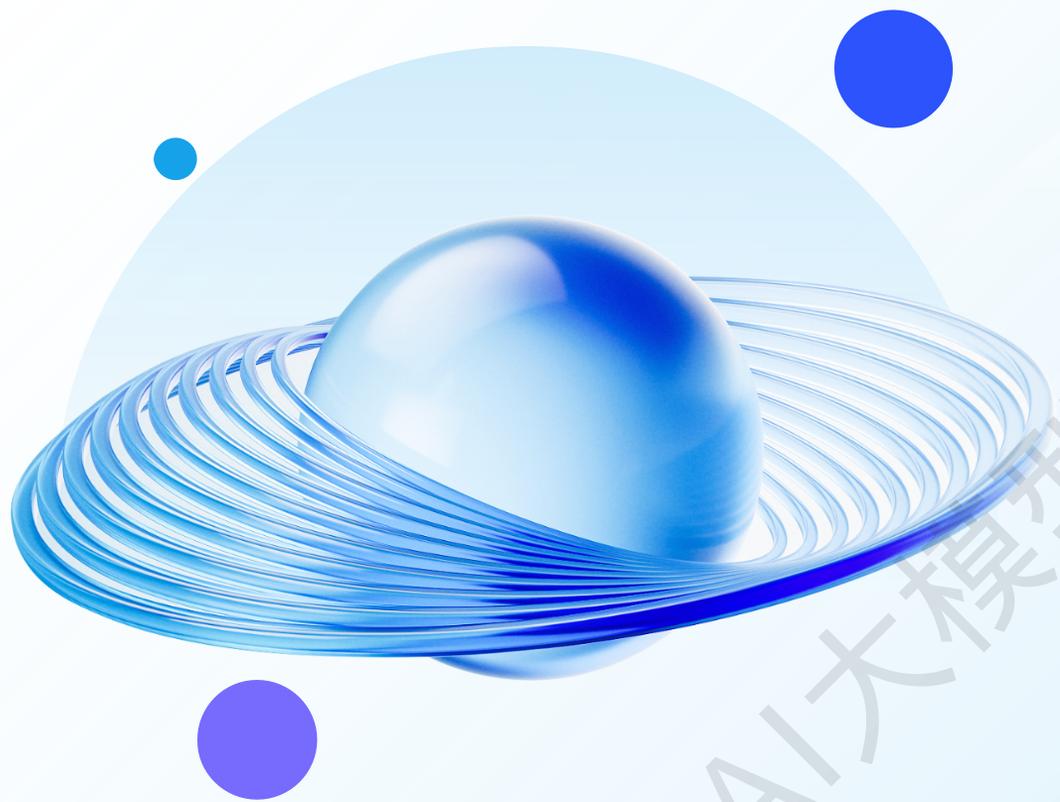
RAG架构与技术

03

RAG@企业应用

04

Q&A



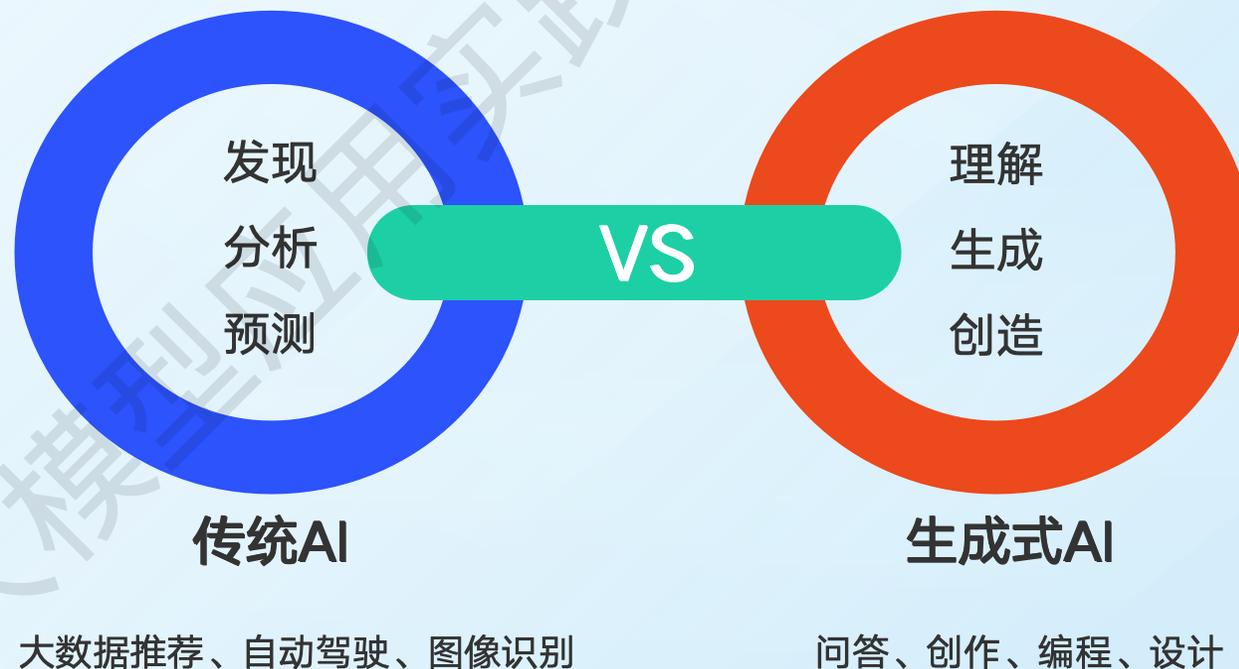
01

# 生成式AI再认识



# ：正确认识生成式AI（Gen-AI）

生成式AI（Generative AI）是一类人工智能技术，旨在生成新的、原创的文本、图像、音频、视频或其他形式的内容。生成式AI具有极强的**输入理解能力、推理能力以及创造能力**，可以生成以前从未存在过的内容。



生成式AI的基础：**大模型**（语言模型、图像模型、视频模型、多模态模型）

# 生成式AI还不完美.....



## 缺乏可解释性

生成式AI模型通常非常复杂，难以解释其内部逻辑和决策过程。



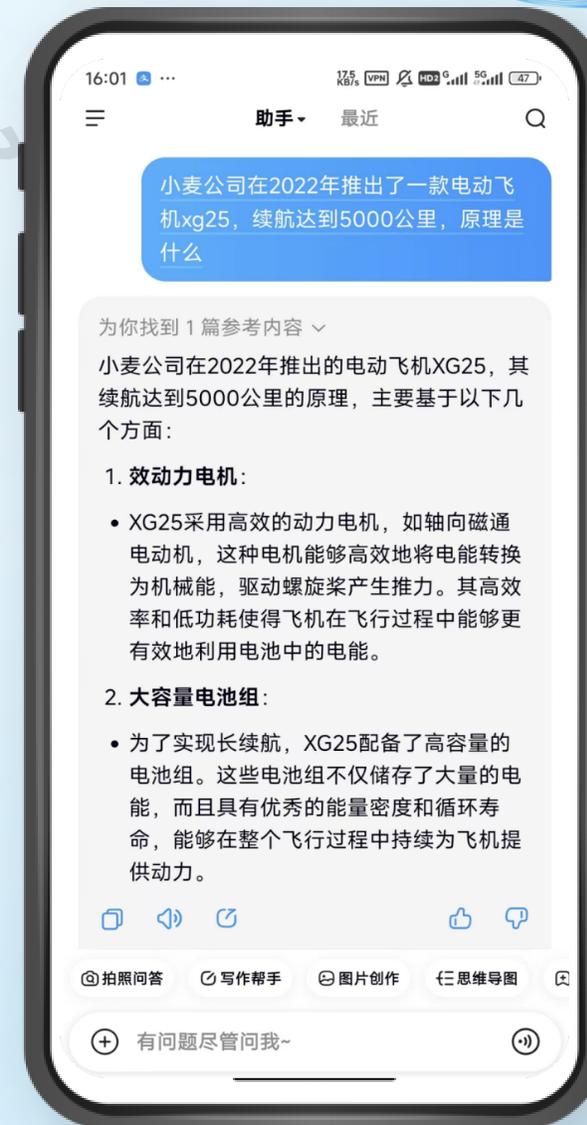
## 不确定性

生成式AI本质是一种概率模型输出，相同的输入其输出并不绝对一致。



## 幻觉问题

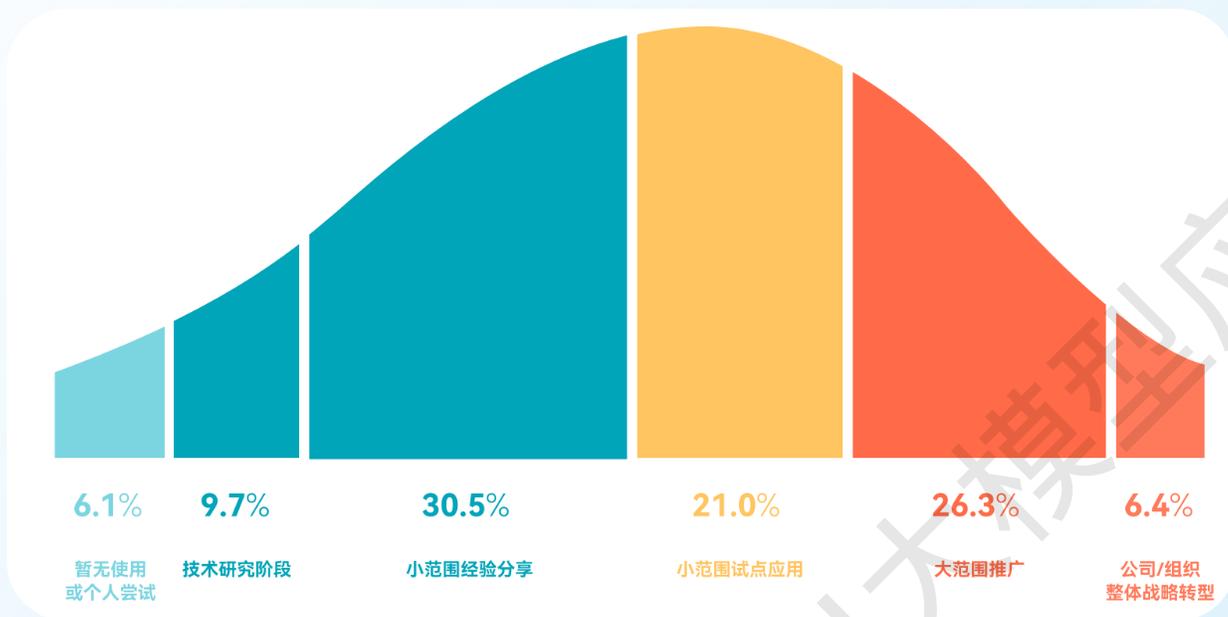
可能生成虚假的、不准确或者不存在的信息，特别是在诱导下。



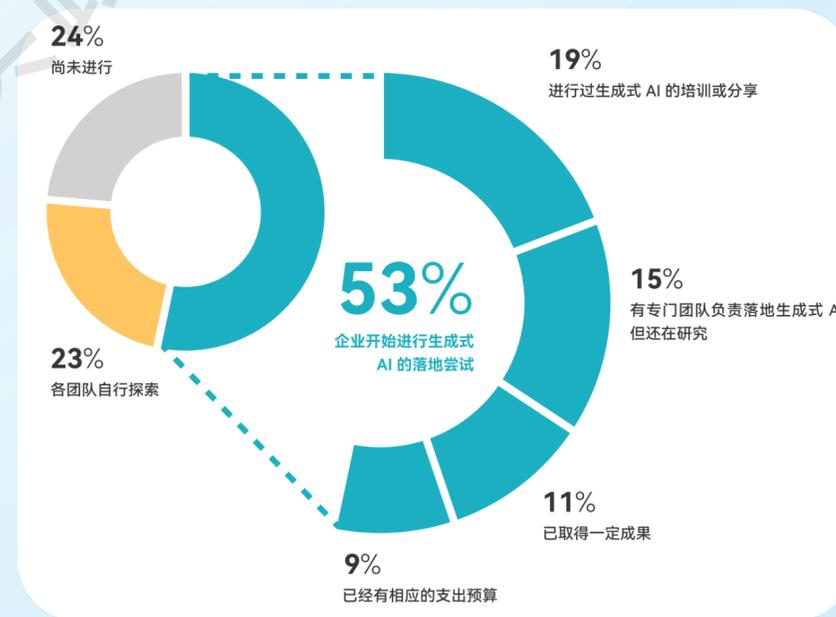
此外，知识时效性、多步推理能力、多模态能力等也制约了应用能力



# 生成式AI在中国企业的应用进入全面发展期



26.3% 的企业中高层在大范围推广生成式 AI 应用

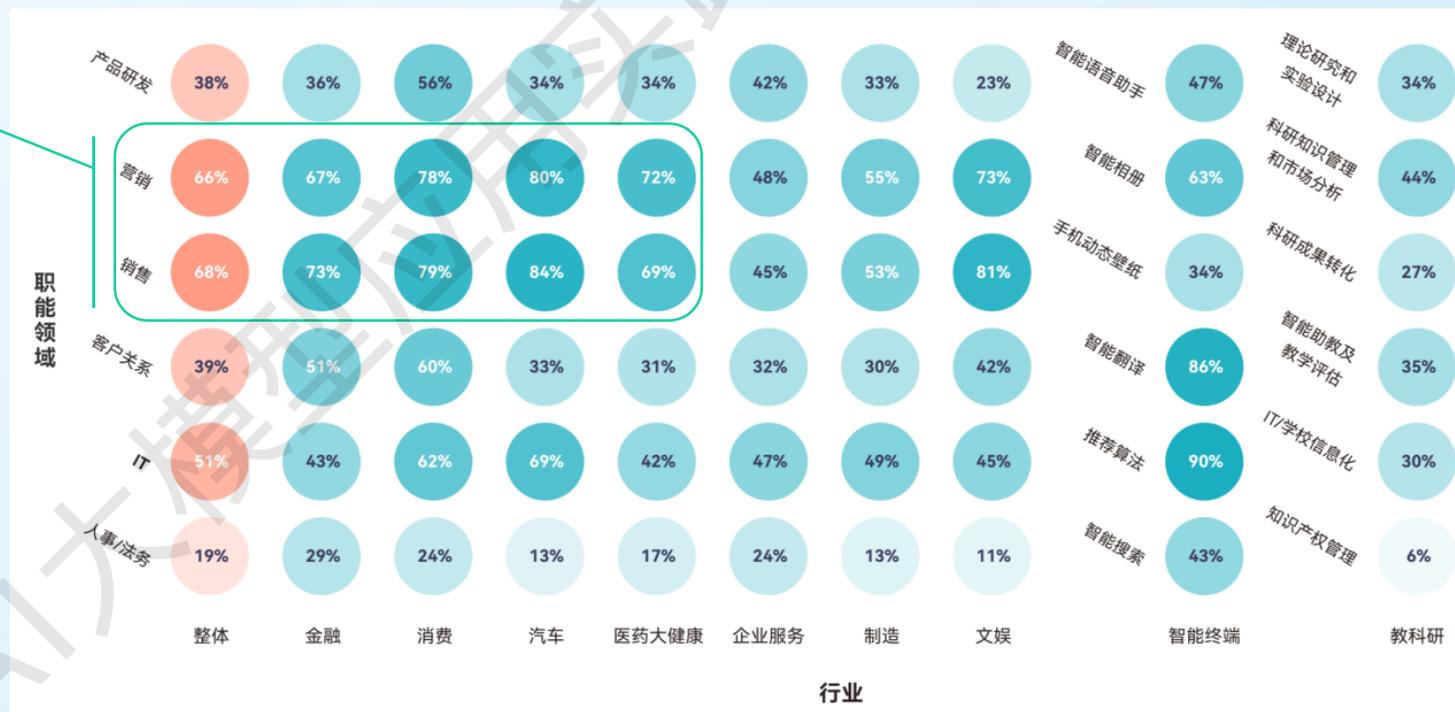


53% 的大企业开始落地生成式 AI 创新



# 国内企业生成式AI的应用领域

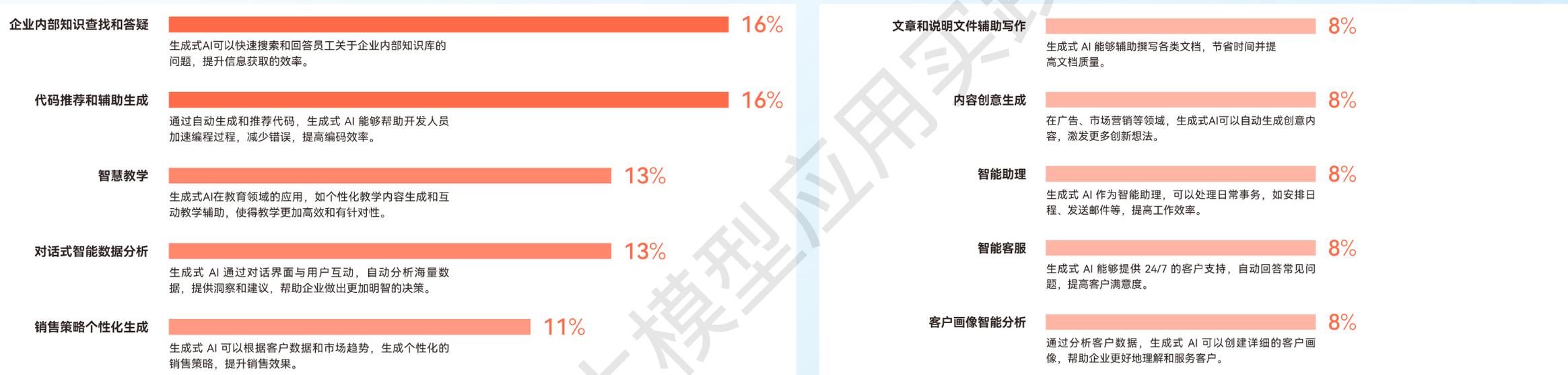
营销与销售是最热门的职能领域，主要用于各类营销过程中需要的文本、图像、甚至视频的生产



不同行业和职能领域的应用进度和场景选择，也反映出生成式 AI 技术在普及过程中的多样性和挑战。



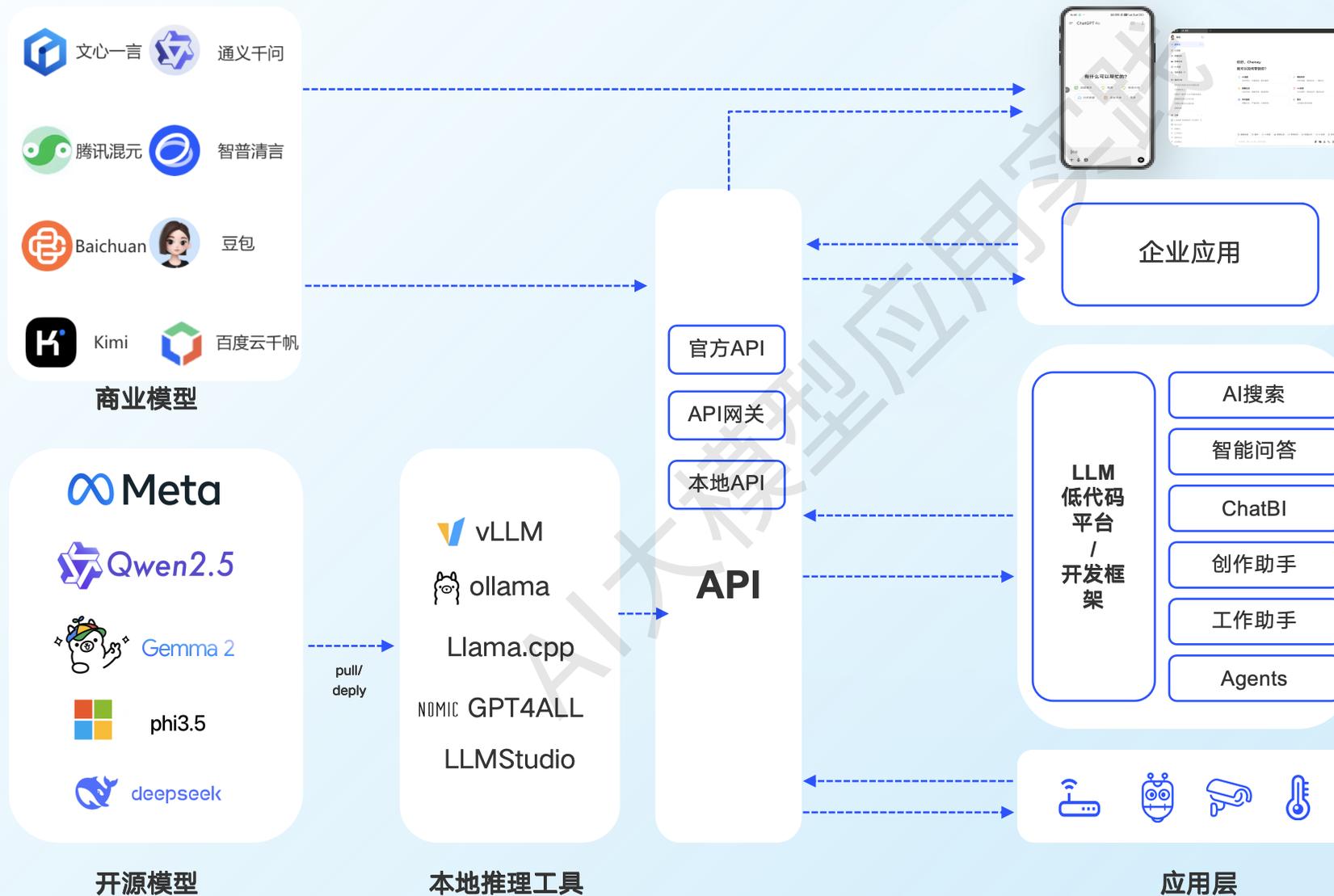
# 生成式AI在企业的应用场景 - CIO们的选择



生成式AI在企业的市场监控、产品设计、营销宣传、订单处理、客户服务、供应链管理、人力资源管理、财务管理、OA系统、经营分析等各个领域都有广阔的应用前景与发展空间。



# 了解生成式AI的应用形态



- ### ToC
- 官方客户端
  - 独立工具
  - 智能设备

- ### AI赋能
- ERP、OA、BI...
  - 行业应用

- ### 企业原生AI应用
- RAG
  - Agent
  - 定制

- ### 其他
- 边缘智能、机器人...



# 决策者最需要了解的几个事实

1 生成式AI不是AI的全部，或许你需要的并非大模型

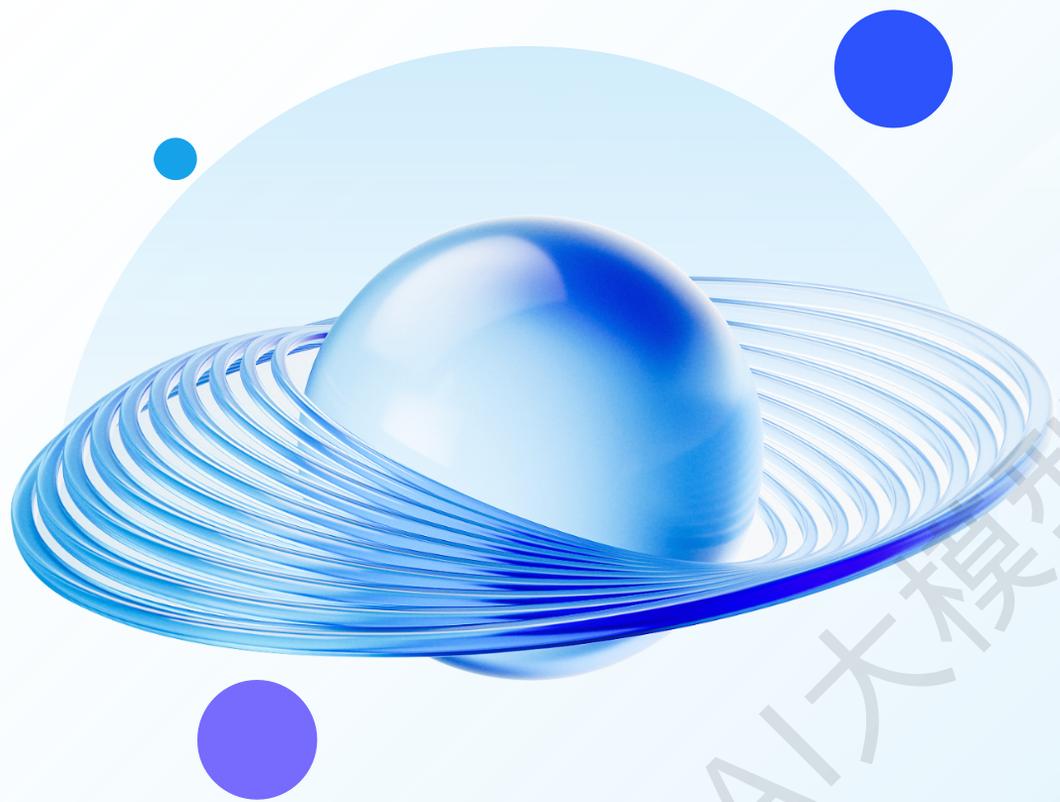
2 不要期望生成式AI有绝对的准确性

3 专业“小”模型有时候比大模型更有效

4 自然语言交互只是最“自然”，并不代表最佳体验

5 私有部署模型不一定比公有云模型更安全

6 不要被原型与Demo所迷惑，适当降低期望值



02

# RAG架构与技术



# 大模型为什么还会需要RAG



老师教错了



训练知识偏差

以偏概全了



过度泛化推理

缺乏深入理解力



理解局限性

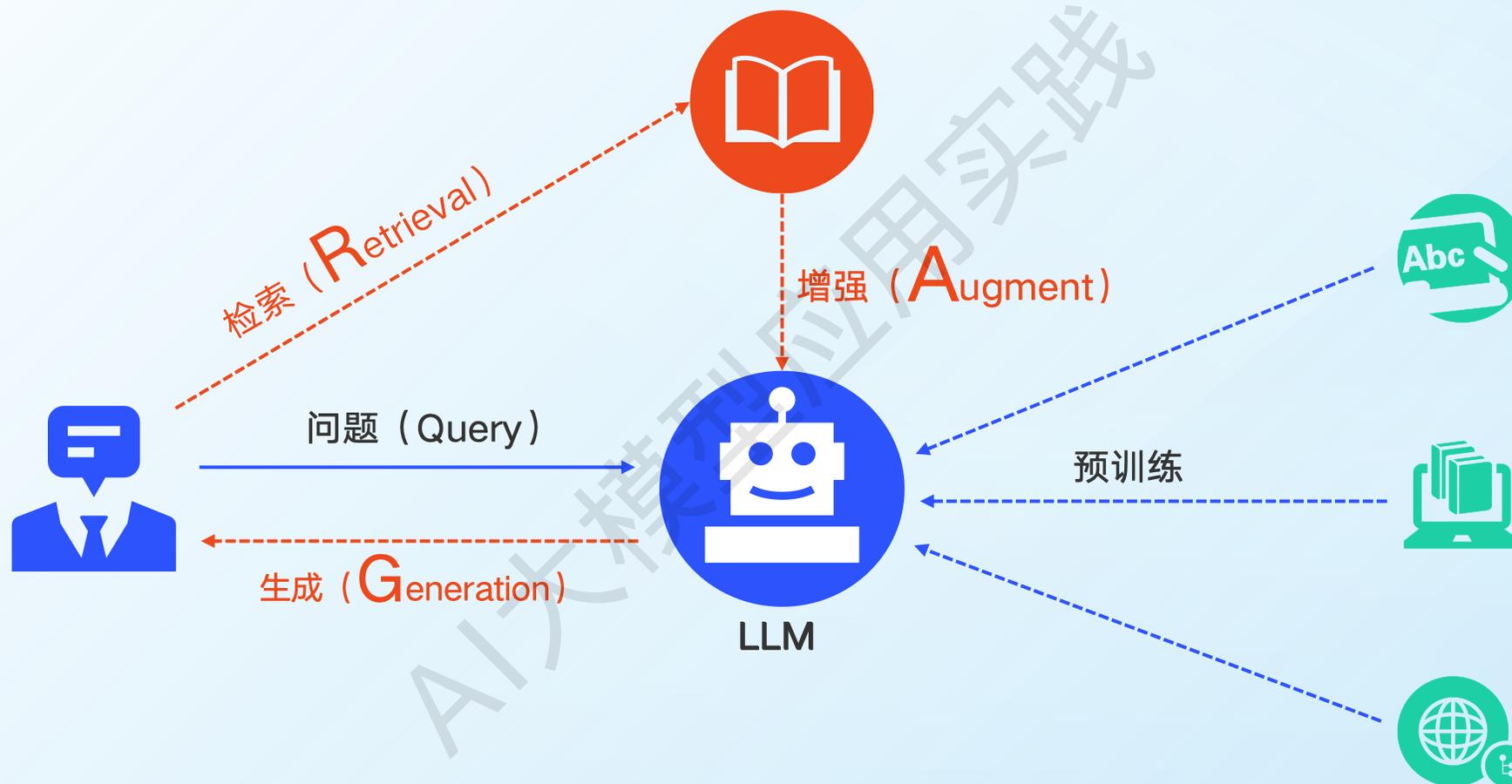
没学过只能瞎编



缺乏领域训练知识

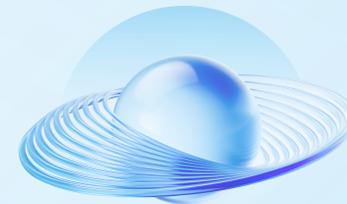
特别在企业级应用中，直接的大模型生成无法满足准确性的要求，因此需要“增强”的生成。

# ：RAG如何优化“幻觉”问题：连接私域知识



RAG = Retrival (检索) + Augment (增强) + Generation (生成)

# 直接LLM生成 vs RAG生成



问题

请告诉我我们公司 2023 年 Q4 的营收增长数据以及主要贡献部门？



LLM

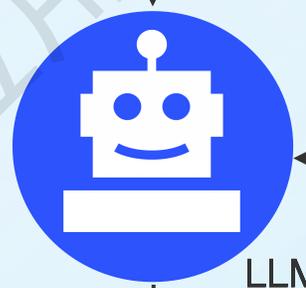
响应

我没有访问公司具体数据的能力，因此无法提供您公司的 2023 年 Q4 营收增长数据以及贡献部门的信息。您可以查阅公司财报或联系公司财务部门获取相关数据。



Direct LLM

请告诉我我们公司 2023 年 Q4 的营收增长数据以及主要贡献部门？



LLM

根据公司财务数据展示，2023 年 Q4 的营收增长为 18%。其中，主要贡献部门是销售部门（12% 增长）和技术支持部门（6% 增长）。



RAG



检索

公司整体营收增长数据：  
• 2023 年 Q4 总营收：1.2 亿人民币  
• 同比增长：18%  
• 环比增长：7%

2. 主要贡献部门及详细数据：

- 销售部门：
  - 营收贡献：7200 万人民币
  - 同比增长：12%
  - 主要驱动因素：
    - 推出了新产品线，获得 3 个大客户订单，总计 2000 万。
    - 提高了客户续费率，从 85% 提高到 92%。
  - 关键地区贡献：华东地区占总销售额的 40%，同比增长 15%。
- 技术支持部门：
  - 营收贡献：3000 万人民币
  - 同比增长：6%
  - 主要驱动因素：
    - 成功 upsell（追加销售）了技术支持高级套餐，为 25 家现有客户提供定制服务。
    - 优化服务流程，工单处理效率提高 20%。

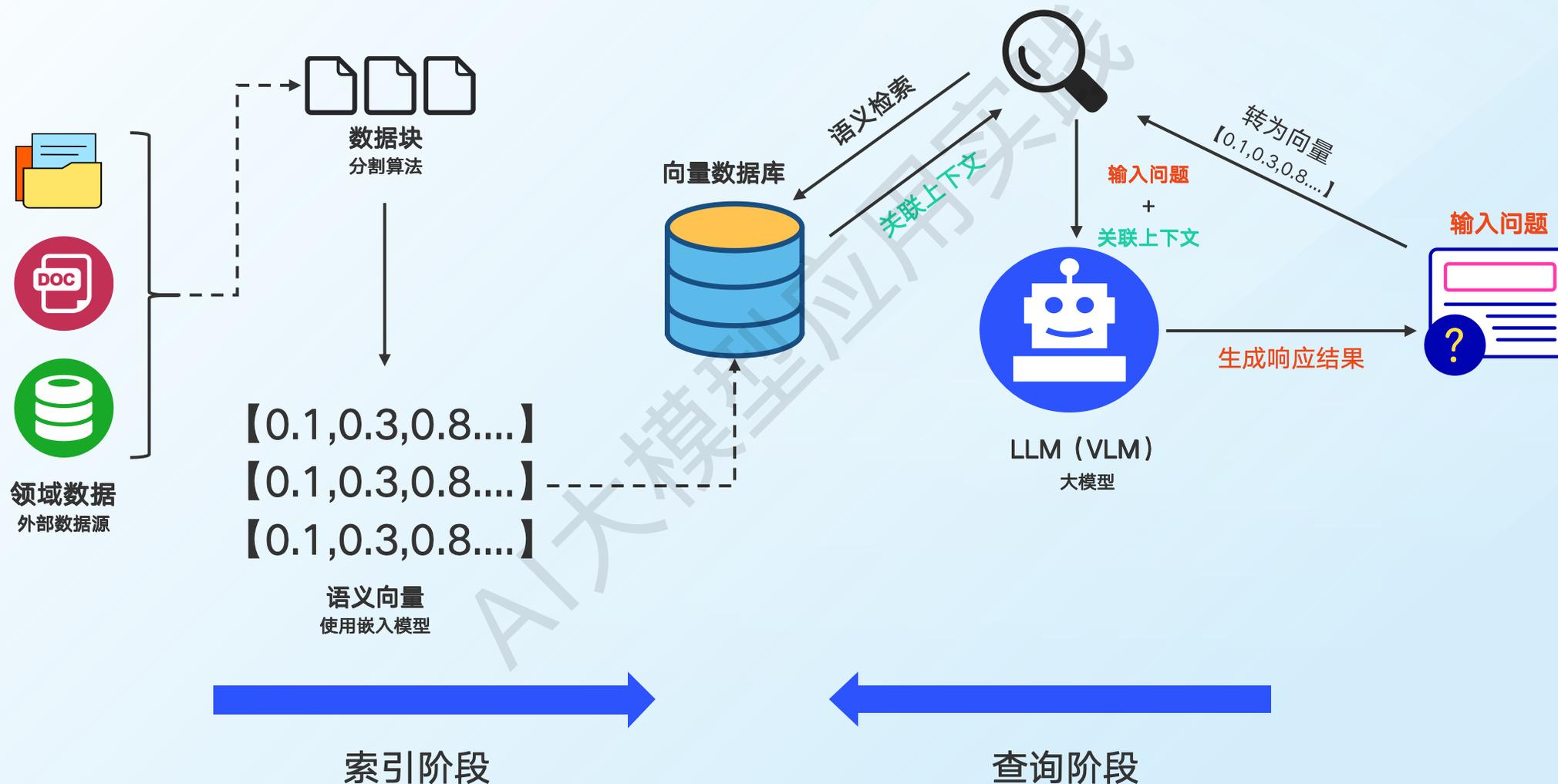
3. 其他部门贡献：  
• 研发部门：通过推出新功能为销售提供了强力支持，贡献间接营收 5%。  
• 市场营销部门：

- 营销活动影响：营销活动带来的线索转化率提升 8%。

关联知识



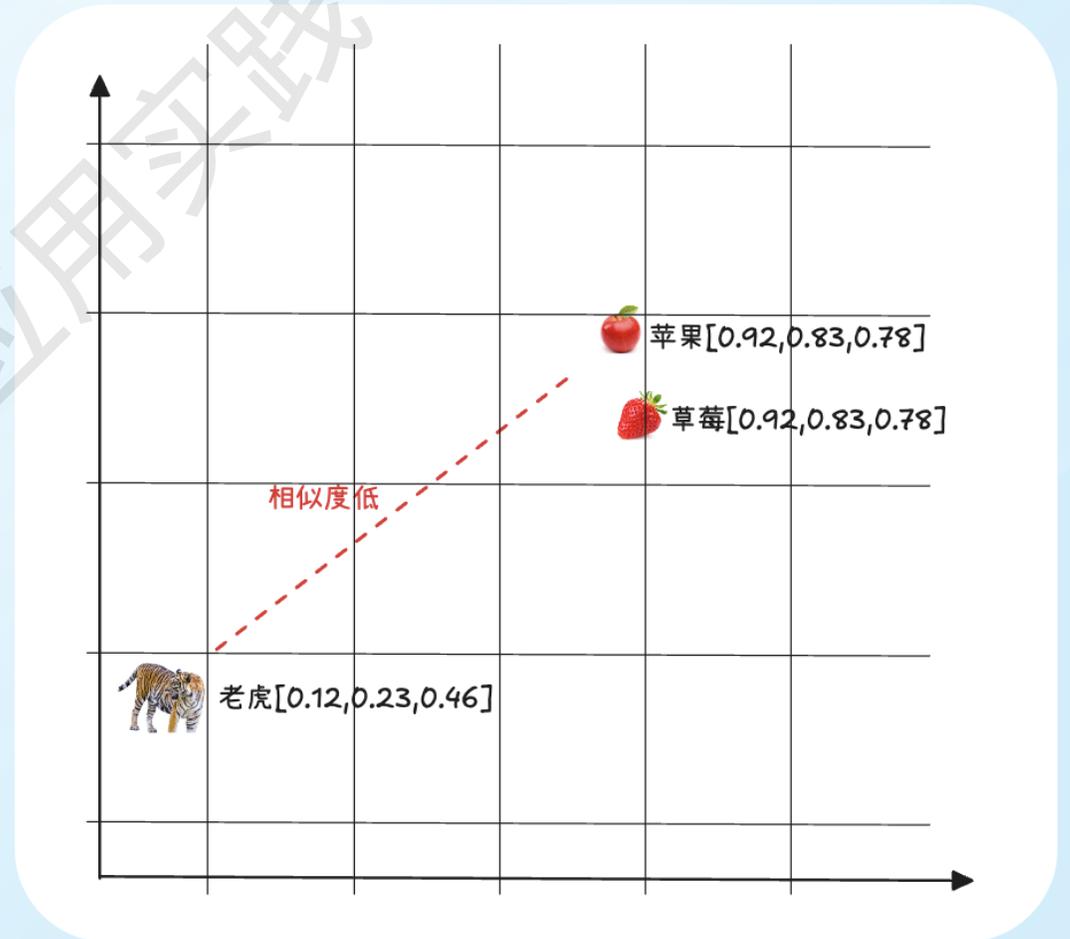
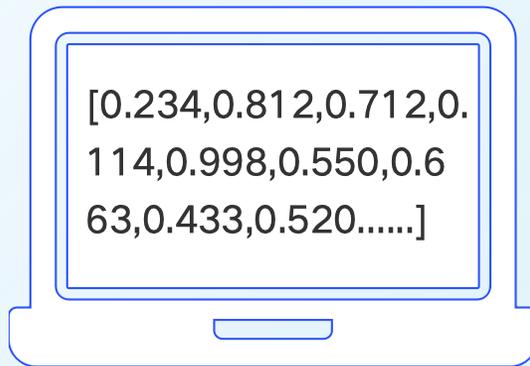
# 经典RAG的架构与工作流程



# 认识经典RAG的核心概念：向量与嵌入 (Embedding)



Embedding  
Model (嵌入模型)



**向量**：可以理解计算机能处理的一组有意义的数字

**向量的意义**：可以通过算法计算事物之间的“相似性”

# 经典RAG的重要基础设施：向量数据库（VectorDB）



## 关系型数据库（RDBMS）



Oracle、MySQL、PostgreSQL:

- 结构化数据为主
- 行列表示，存在主外键关系
- SQL + 查询条件
- 例：“查询销售金额大于100万的产品”

## 向量数据库（VectorDB）



Milvus、ChromDB、Pinecone:

- 非结构化数据为主：文本、图像
- 高维向量+元数据，存在相似关系
- 向量相似度算法查询最近邻
- 例：“查询与图片A最相似的5张图片”



# 经典RAG在复杂应用环境中的挑战

过去一年中，企业实施RAG的成功率（基本达到预期目标）不足50%



## 检索召回的精度问题

检索与排序的精准性往往取决于很多因素



## 大模型抗干扰能力问题

大模型对于上下文的理解与推理能力存在不确定性



## 响应质量与性能矛盾问题

较高的响应质量往往带来更低的响应性能



## 企业数据并非一个文档

多种形态、多来源、多种特点、数量庞大



## 输入任务的多样性

不要期待使用者按照你设想的去提问

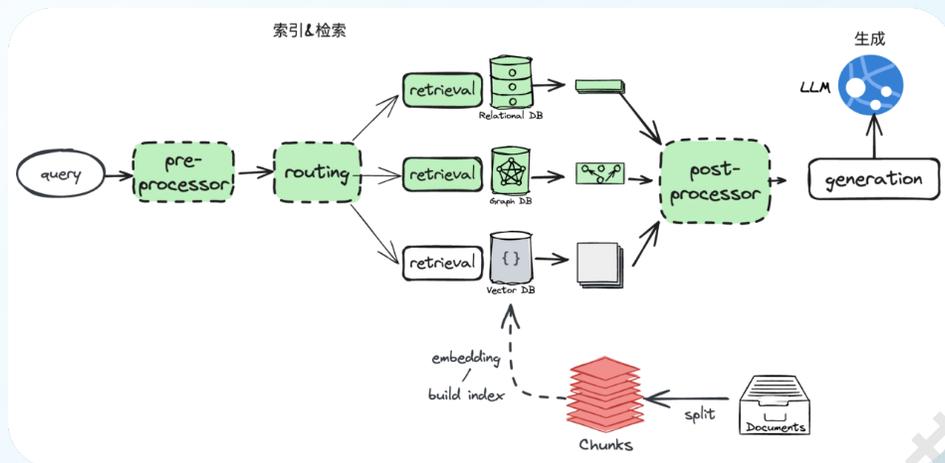


## 更高的工程化要求

更高的准确性、确定性、可观测性、可解释性

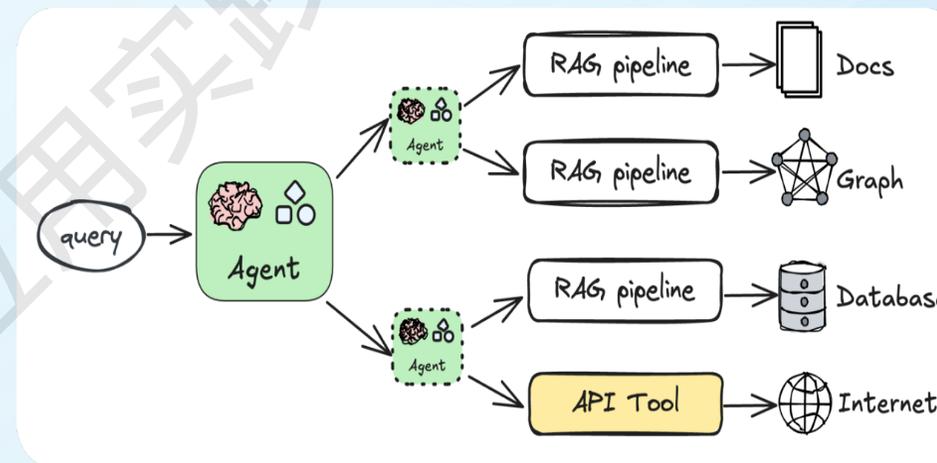


# RAG架构进化的两个重要方向



## 高级RAG/模块化RAG

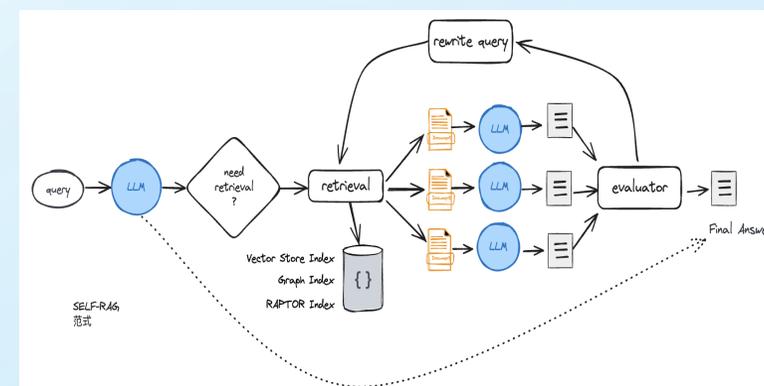
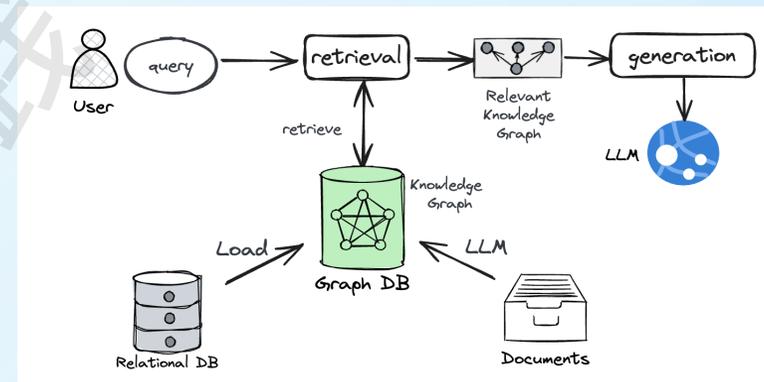
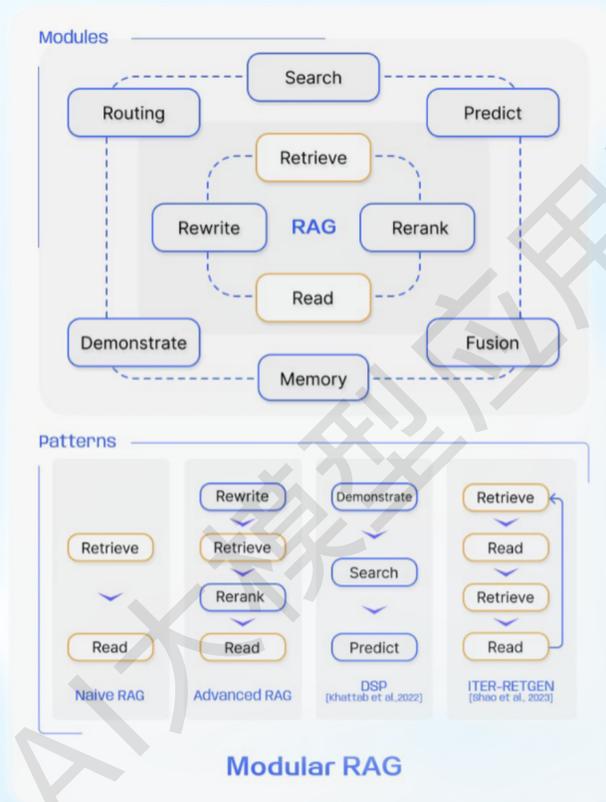
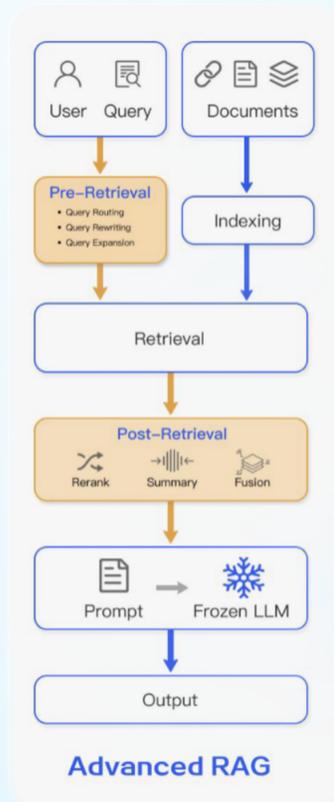
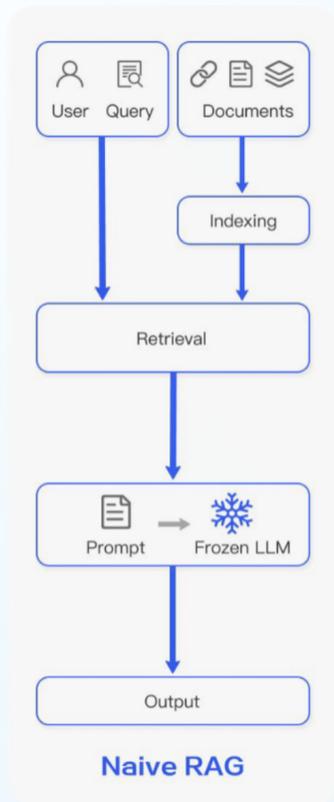
- ✓ 新的索引
- ✓ 新的算法
- ✓ 新的模块
- ✓ 新的流程



## Agentic RAG

- ✓ 具有Agent特征的RAG
- ✓ 适应多样复杂的查询任务
- ✓ 自主的规划与使用RAG管道
- ✓ 可集成外部工具

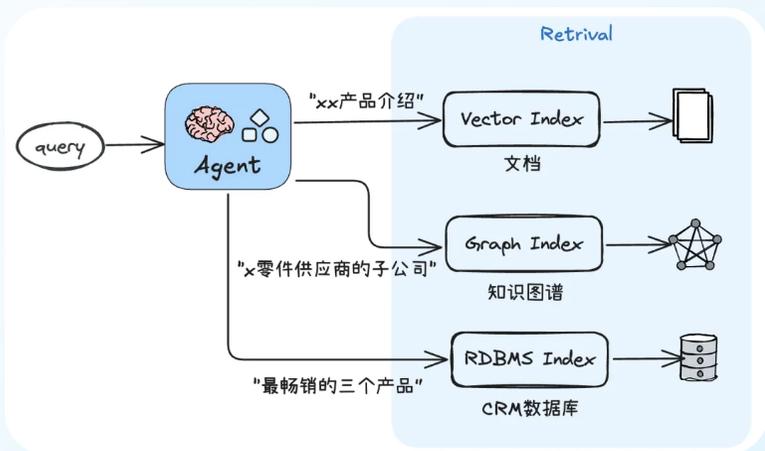
# RAG架构的进化 - 高级RAG/模块化RAG



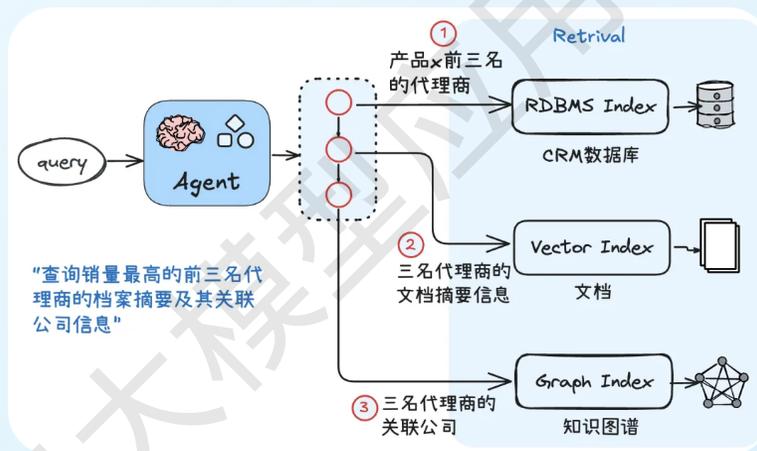
# ：RAG架构的进化 - Agentic RAG



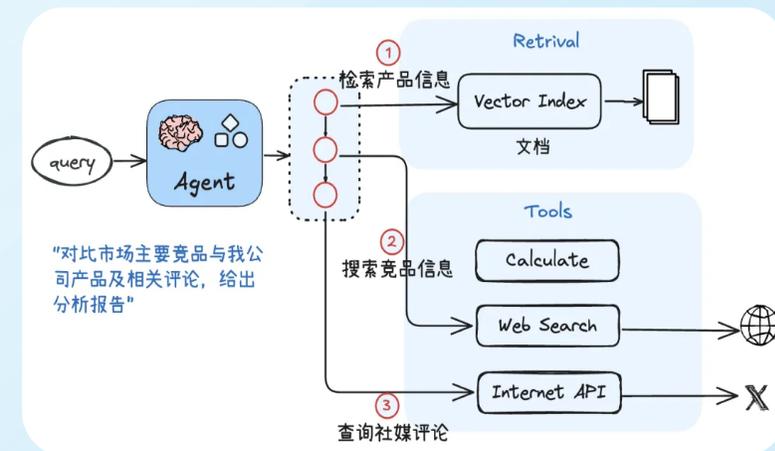
将AI智能体的自主规划（如路由、行动步骤、反思等）能力带入到传统的RAG，以适应更加复杂的RAG查询任务。



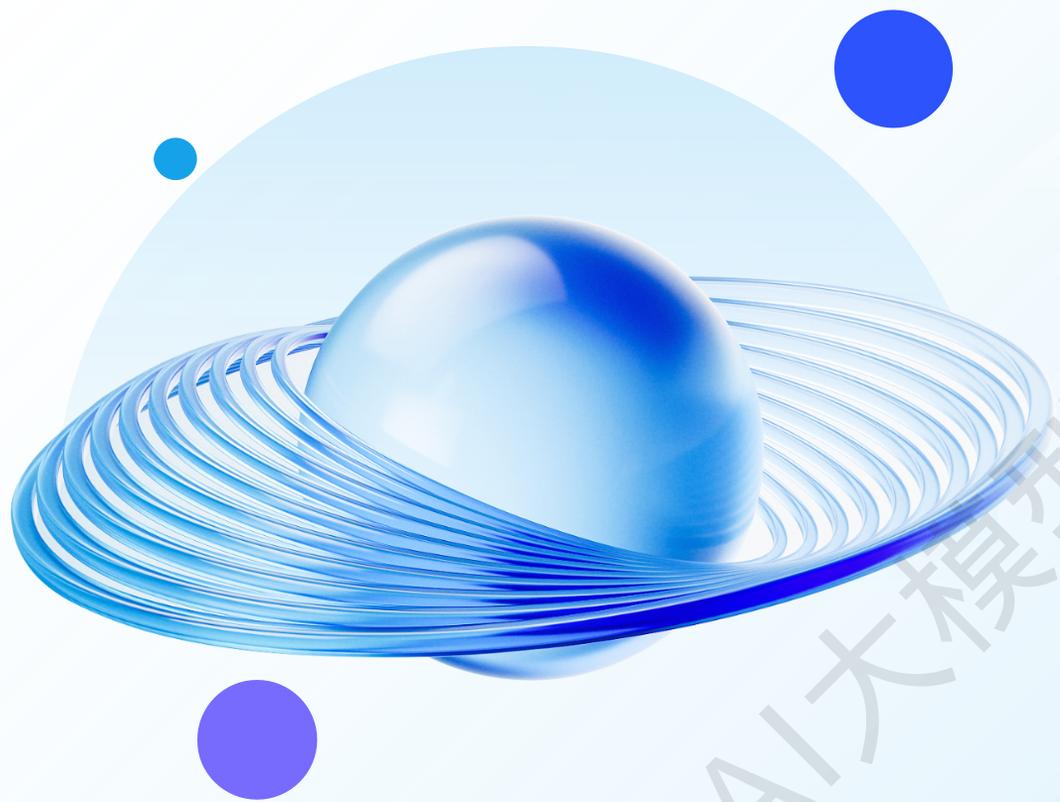
在不同类型的RAG管道间路由，  
以适应任务的多样性



融合多种类型的RAG管道与数据源，  
以适应综合性复杂查询任务



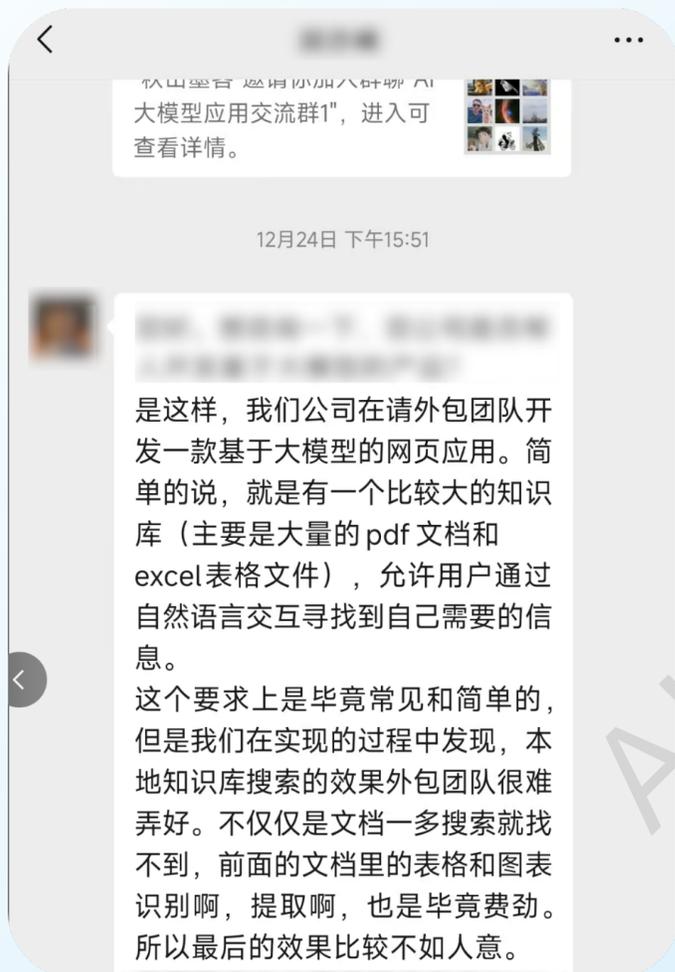
与必要的外部检索工具协作，  
以增强输出的准确性



03

**RAG@企业应用**

# ：RAG@企业应用：现实很骨感



“效果不好，文档一多就找不到答案...”

“时好时坏，有时候换个问法就好了...”

“开发一天，优化十天，经验难复用...”

企业的自身条件与使用场景的差异，加上自然语言的复杂性，大模型自身的不确定性，带来企业RAG的复杂性。

# ：RAG在企业落地的决策关键



01

## 应用场景

应用的业务价值、最终用户、  
任务场景、使用形式

02

## 数据条件

数据的来源、形态、内容、质  
量、安全、更新周期等

03

## 工程要求

准确性、性能、可靠性、可解  
释性、可维护性、扩展性等



# 企业RAG：选择合适的应用与任务场景

## 优先考虑

- ✓ 没有绝对的准确性衡量指标
  - 内容生成、改写、总结、摘要、设计、角色模拟等
- ✓ 能够容忍一定的错误概率
  - 智能搜索、咨询、问答、识别、分类、推荐等
- ✓ 发生错误但能够被自动拦截
  - 涉及工具使用、结构化输出等可校验的场景
- ✓ 发生错误但有人类做二次处理
  - 代码辅助、功能导航、论文辅助、Excel辅助等

## 有绝对准确性要求且没有保护机制

- ✓ 有绝对准确性要求且没有保护机制
  - 交易生产型应用
  - 关键决策支持型应用
- ✓ 对响应性能要求极高的场景
  - 实时事件的流式应用
  - 边缘计算应用
- ✓ 对内容创作但又没有严格的过滤审核机制

## 慎重考虑

根据统计，目前生成AI在企业应用的主要价值体现在提高效率与降低成本，而非直接创造收入。



# 企业RAG：认识与评估数据条件





# 企业RAG：明确合理的工程要求





# 企业RAG应用的开发与部署平台



## 01

### 云平台（低代码）

- 字节Coze
- 百度千帆AppBuilder
- 阿里云百炼
- .....

- 简单易用、快速构建
- 灵活性差，无法深度定制
- 第三方模型支持较少
- 云端部署，安全性

## 02

### 开源（低代码）

- FastGPT
- Dify
- RAGFlow
- .....

- 较简单、少量开发部署
- 较灵活，必要时可改代码
- 本地部署，安全性高
- 深度定制较复杂，高级功能收费

## 03

### 开源开发框架

- LlamaIndex
- LangChain
- DSPy
- .....

- 较复杂，需要自主开发
- 最灵活，深度按需定制
- 本地部署，安全性高
- 功能强大，支持大量第三方

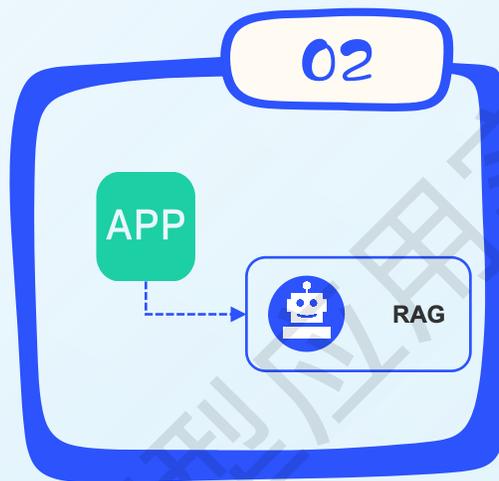
# 企业RAG应用的三种存在形式



## 独立RAG应用

面向使用者独立的端到端RAG应用。

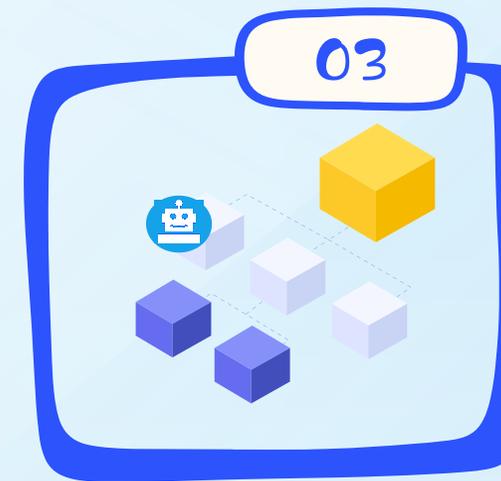
- ✓ 企业AI智能搜索
- ✓ 企业内部知识助手
- ✓ IT技术支持工具
- ✓ ...



## RAG API赋能

提供RAG的API为其他企业应用做AI赋能。

- ✓ 呼叫中心坐席助手
- ✓ 智能在线客服
- ✓ CRM智能体
- ✓ ...

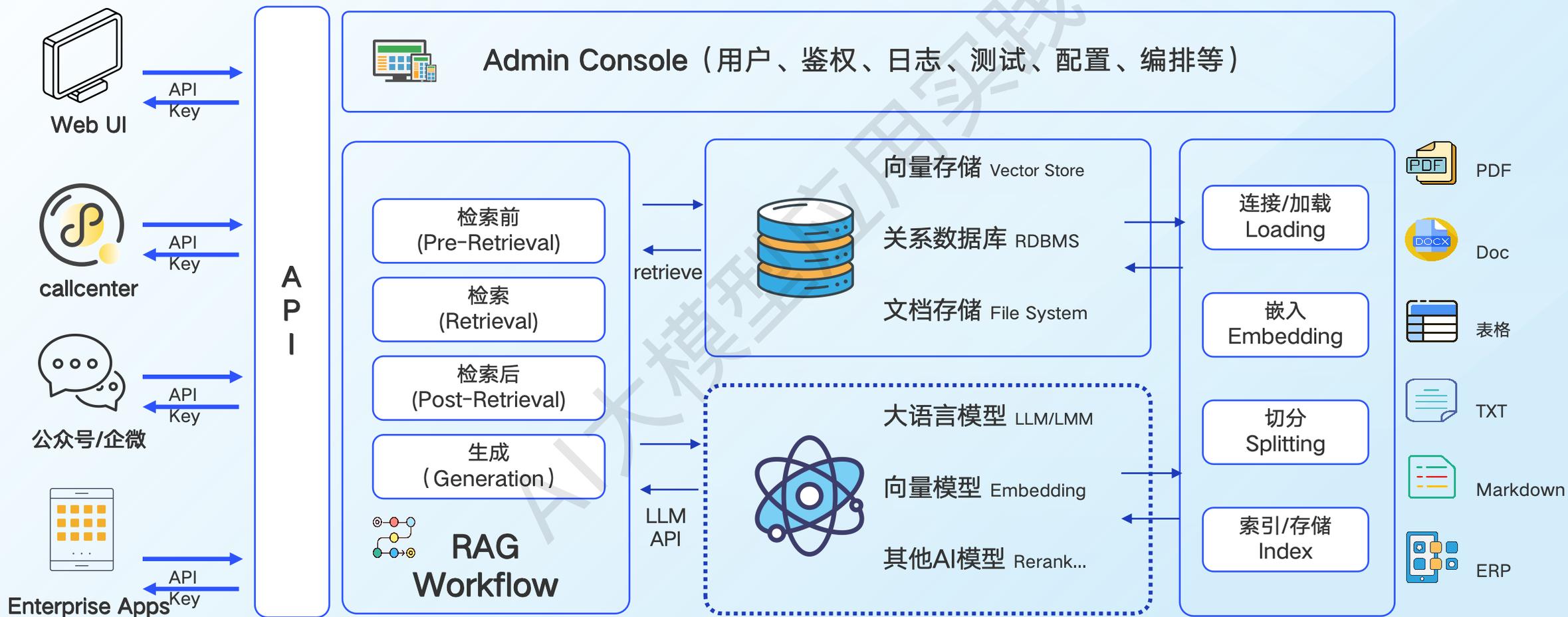


## 嵌入式RAG模块

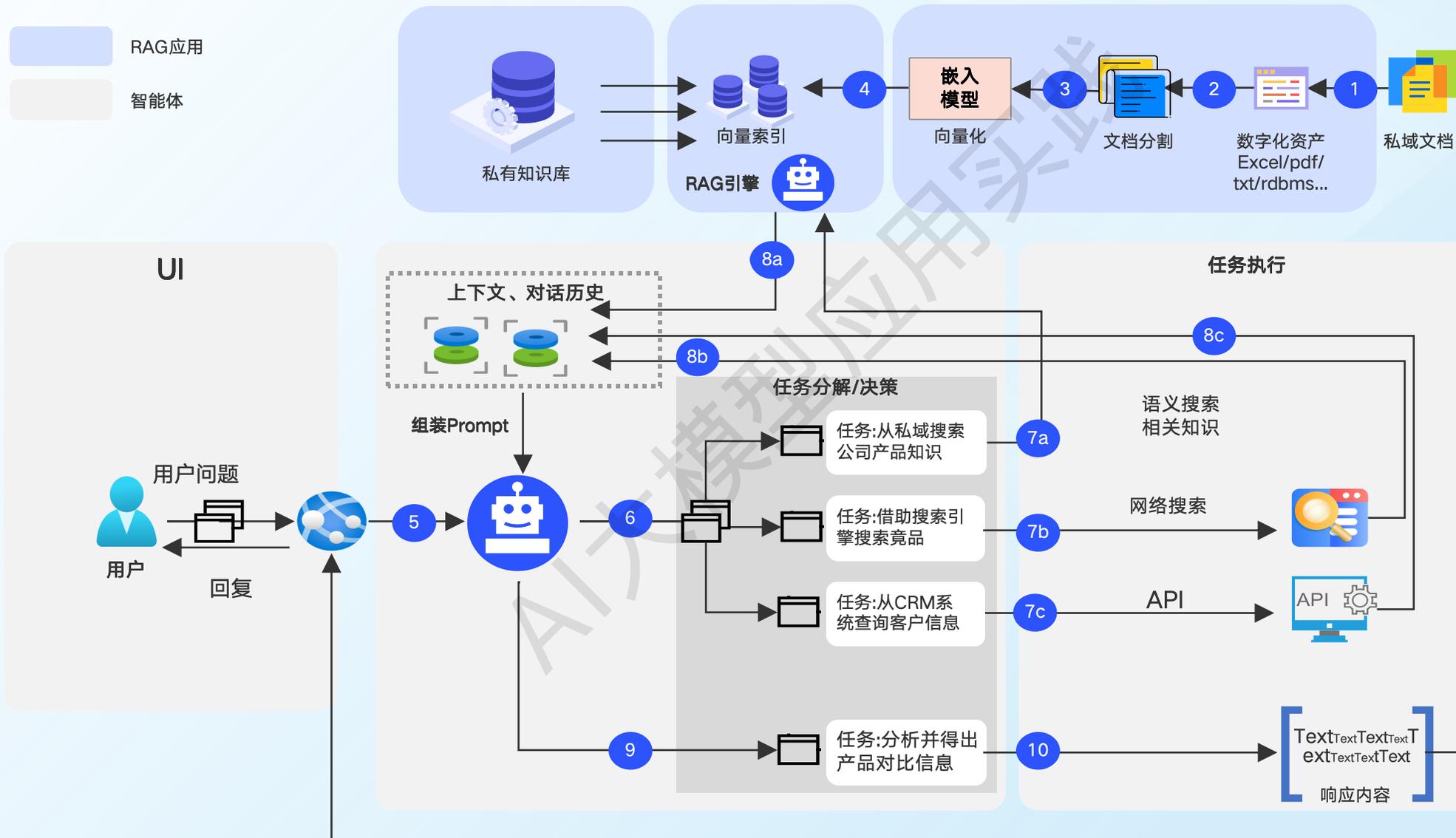
其他AI应用（如Agent）借助RAG思想增强与优化。

- ✓ 优化Agent工具调用
- ✓ 实现AI个性化记忆
- ✓ 优化NL2SQL应用
- ✓ ...

# 端到端企业RAG应用的整体架构

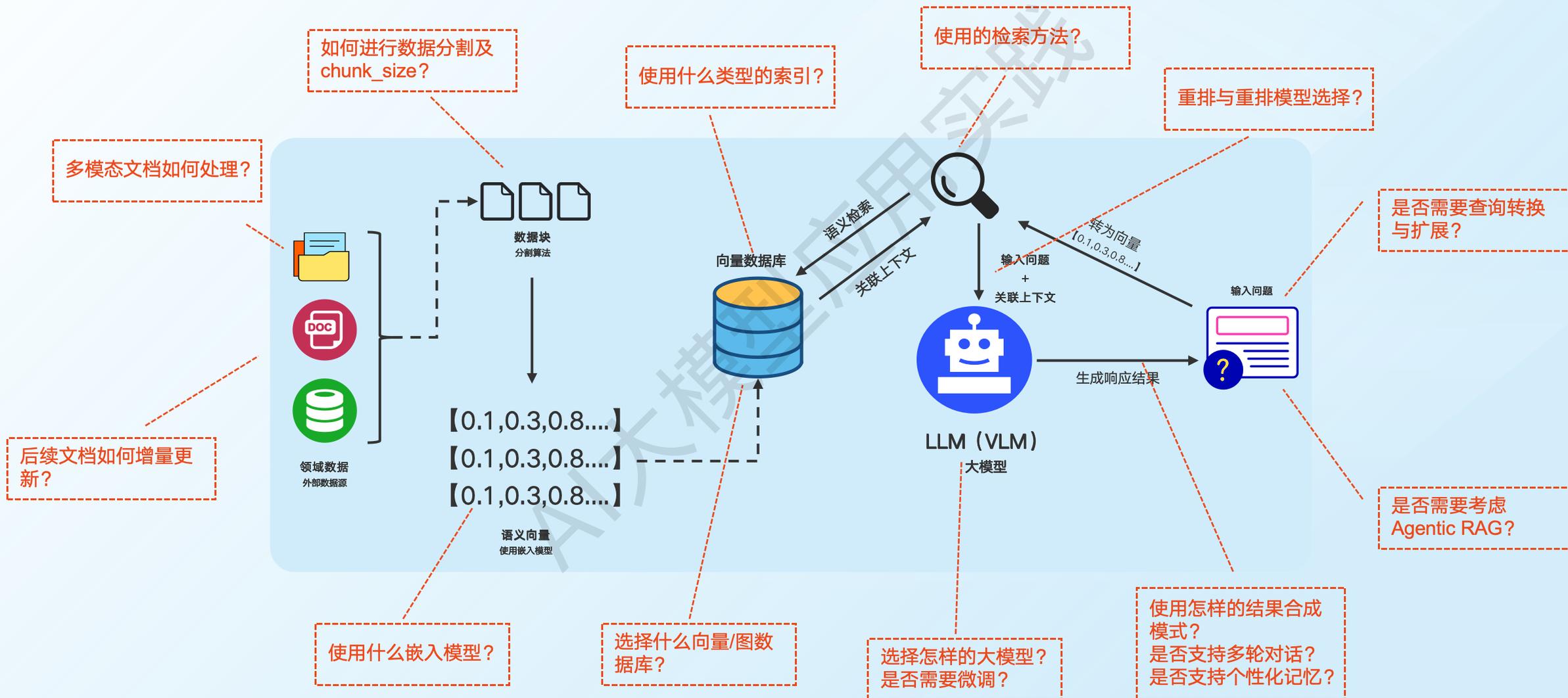


# RAG API赋能其他企业应用





# 企业RAG应用的技术优化选项



# 企业RAG应用的底层模型



## 商业大模型

云端



性能优越，易于使用，持续更新，  
无需硬件投入



数据隐私风险高，长远来看成本可能较高，  
缺乏定制能力



快速上线、低数据敏感性场景



## 开源模型+微调

本地部署/云端部署

数据自主可控，高定制化，长期成本可控

初始成本高，维护难度大，技术门槛较高

数据敏感、高定制化需求场景



## 商业+开源

灵活性高，兼顾性能与隐私需求，  
优化成本

实现复杂度高，延迟和依赖问题可能存在

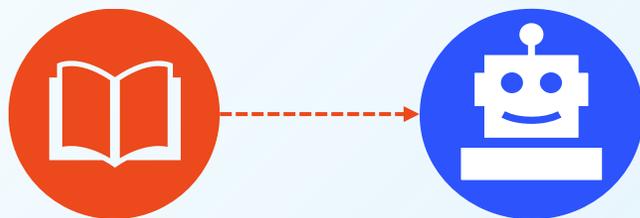
隐私敏感且对性能要求高的场景



# 模型微调的选择：根据任务的特点与要求

## RAG

考试时提供某个领域的参考书



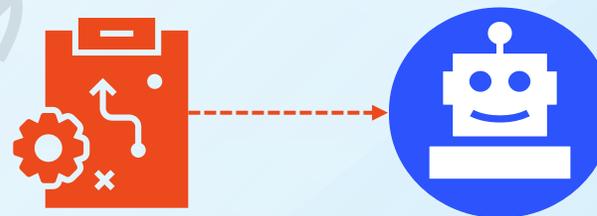
- 可随时利用外部信息、随时调整提示与策略、更灵活
- 动态更新，不依赖于固定的训练数据集，适应性强



- 依赖外部数据源，存在延迟或数据更新不及时的问题
- 效果依赖于检索系统的质量，对应用调优要求高

## 微调 (Fine-tuning)

考试前辅导使他成为某个领域的专家



- 针对特定任务训练的模型，对下游应用更简单友好
- 不依赖外部数据源与上下文，推理阶段成本更低

- 一旦完成训练，知识就固定了，知识的时效性较差
- 需要大量的标注数据和计算资源，成本与技术要求高

微调：相对稳定、迭代周期较长的垂直领域知识（如医学、法律）、精准度要求极高的关键任务。

# 企业RAG应用的工程化新挑战



## LLM应用的新挑战：

- ✓ 难以调试与跟踪
- ✓ 不确定且难以评估的输出
- ✓ 自然语言带来的不确定性
- ✓ 需要持续的迭代优化
- ✓ 控制成本的需要





# · RAG应用的工程化挑战：涵盖不同的生命周期阶段

原型

开发测试

预发布

生产

跟踪调试

提示词 (Prompt) 管理

测试数据集 (Dataset)

用户反馈搜集/手工标注

自动化评估

日志、监控与故障分析

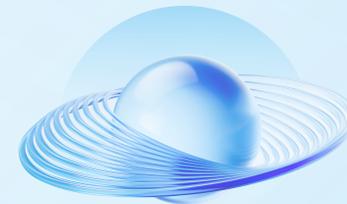
识别用户模式与边缘行为

监控延迟与成本

多版本跟踪、A/B测试



工程化平台：LangSmith、LangFuse等



# RAG应用工程化的关键环节：测试评估

## 大模型的不确定性

开发时准确不代表生产可用  
少量准确不代表大部分准确

## 维护与改进的需要

应用的持续优化，需要科学、  
快速、可复用的方式评估效果



## 知识库的更新需要

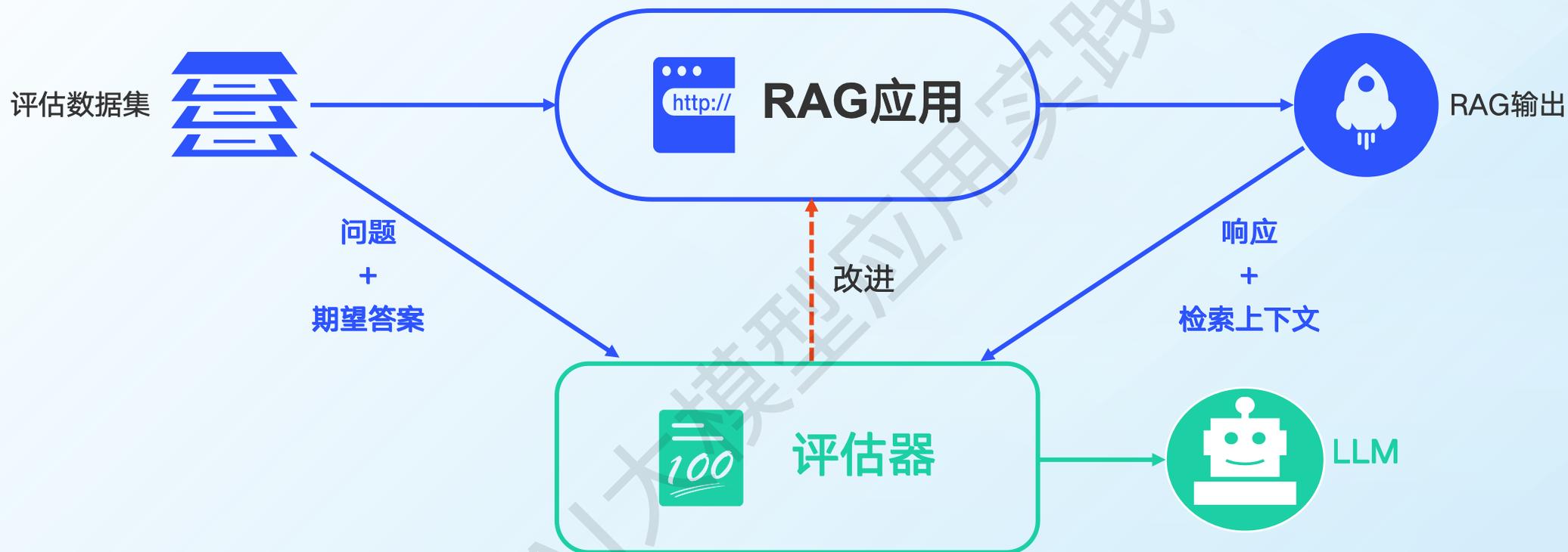
新的知识不断更新，可能会产生  
干扰与影响

## 模型选择的需要

科学的评估有助于  
选择最合适的大模型



# RAG应用的评估 - 借助科学的评估框架与工具



评估对象：

- ✓ 检索器
- ✓ 生成器

评估指标：

- ✓ 上下文相关度、召回率、精度
- ✓ 答案正确性、相关性、忠实度

评估方法：

- ✓ 开发框架模块（如LangSmith）
- ✓ 独立评估框架（Ragas、Langfuse）



# 最后：企业RAG项目落地的总体建议

## 审慎选择应用场景

选择明确有用的业务场景  
避免为了AI而AI

## 选择最适合的技术策略

合适的就是最好的，不盲目跟风  
充分考虑数据、场景等条件

## 小步快走，持续迭代

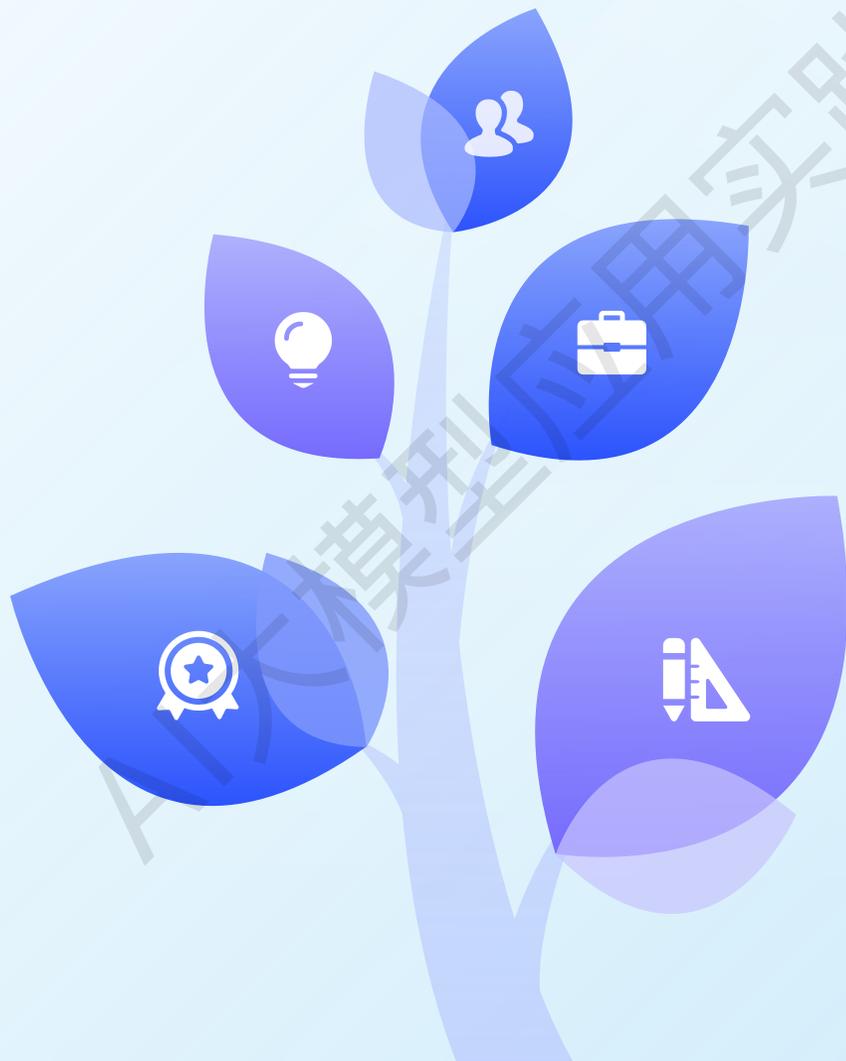
小规模开始，持续迭代，逐步扩大  
降低风险，保护投资

## 保持合理的期望值

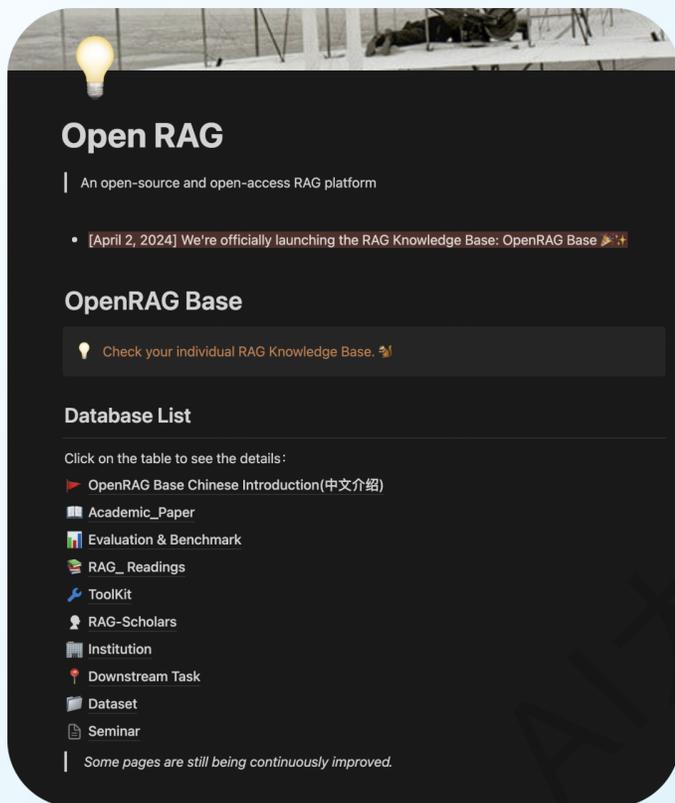
充分认识生成式AI的优缺点  
“三天上线，半年调试”是常态

## 不要忽视数据工程

意识到数据对RAG的重要性  
数据不仅是IT，需要多部门协作



# 推荐资源



Open RAG Base

<https://github.com/Tongji-KGLLM/RAG-Survey/>



《基于大模型的RAG应用开发与优化- 构建企业级LLM应用》



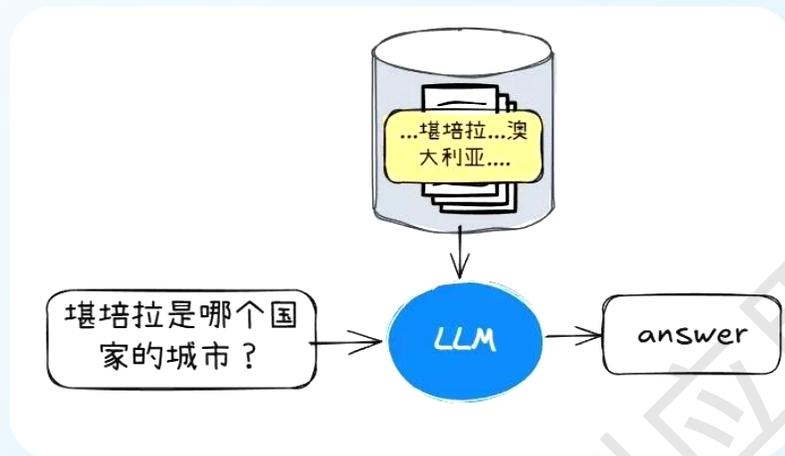
# THANK YOU

AI大模型应用实践

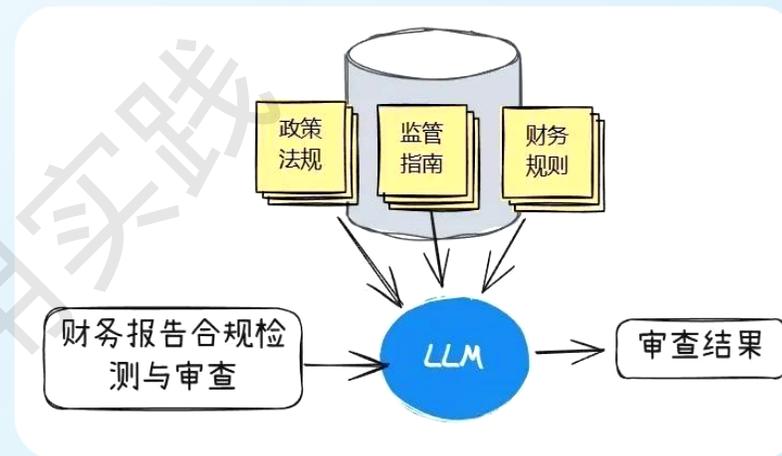


AI大模型应用实践

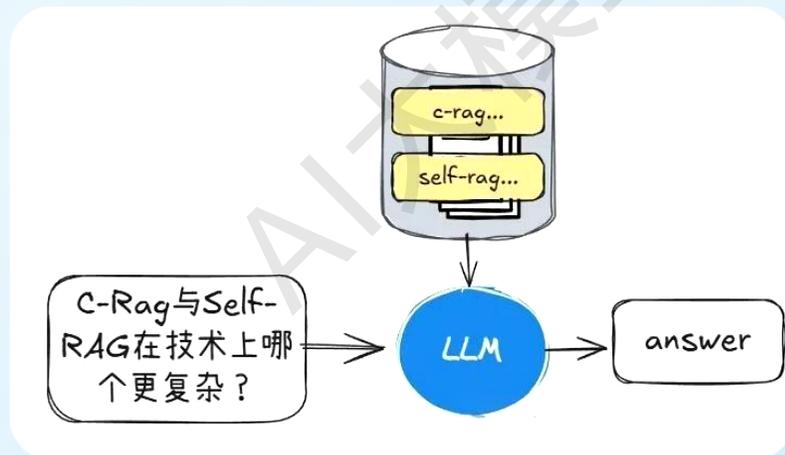
# 数据增强型LLM应用的4种任务级别



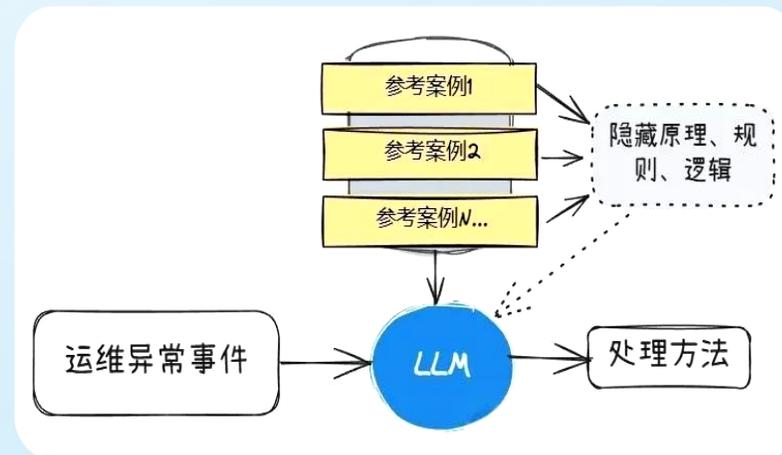
基于显式事实的查询: RAG



基于可解释原理的查询: 提示微调、CoT提示、Agentic



基于隐式事实的查询: 高级RAG/Graph/Text2SQL等



基于隐藏原理的查询: 上下文学习、微调、小模型辅助

# 数据质量

