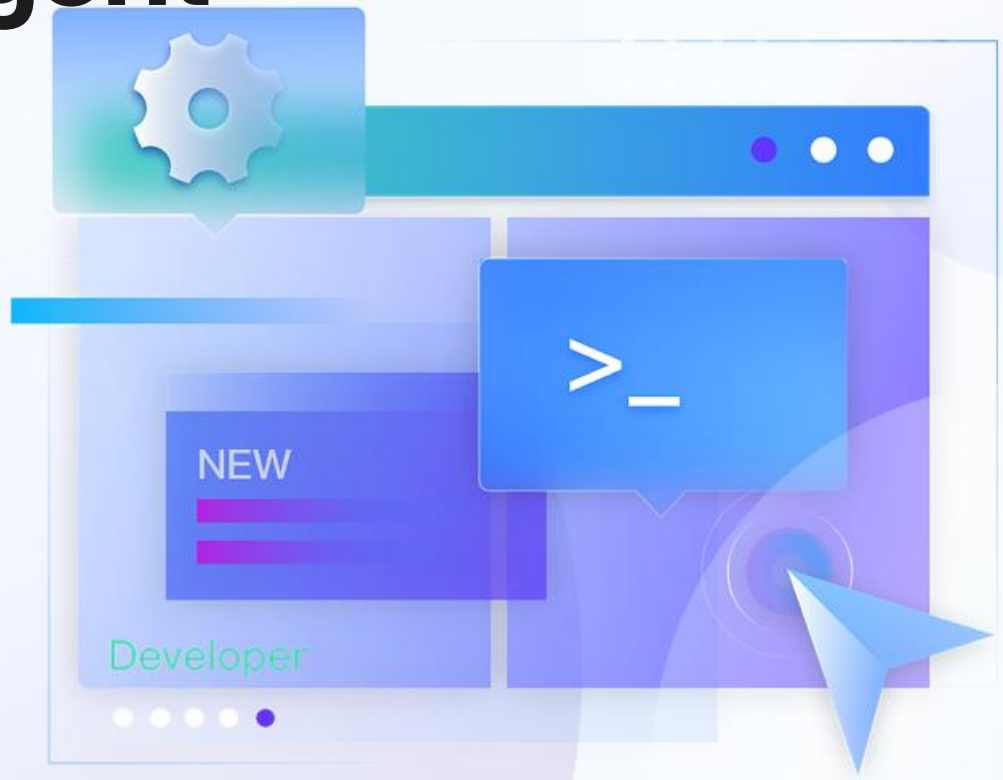
面向数据分析领域的Agent思考和探索



演讲人

屿你数智AI负责人：eason

2024/07/06

Contents
目录

Agent定义和特性

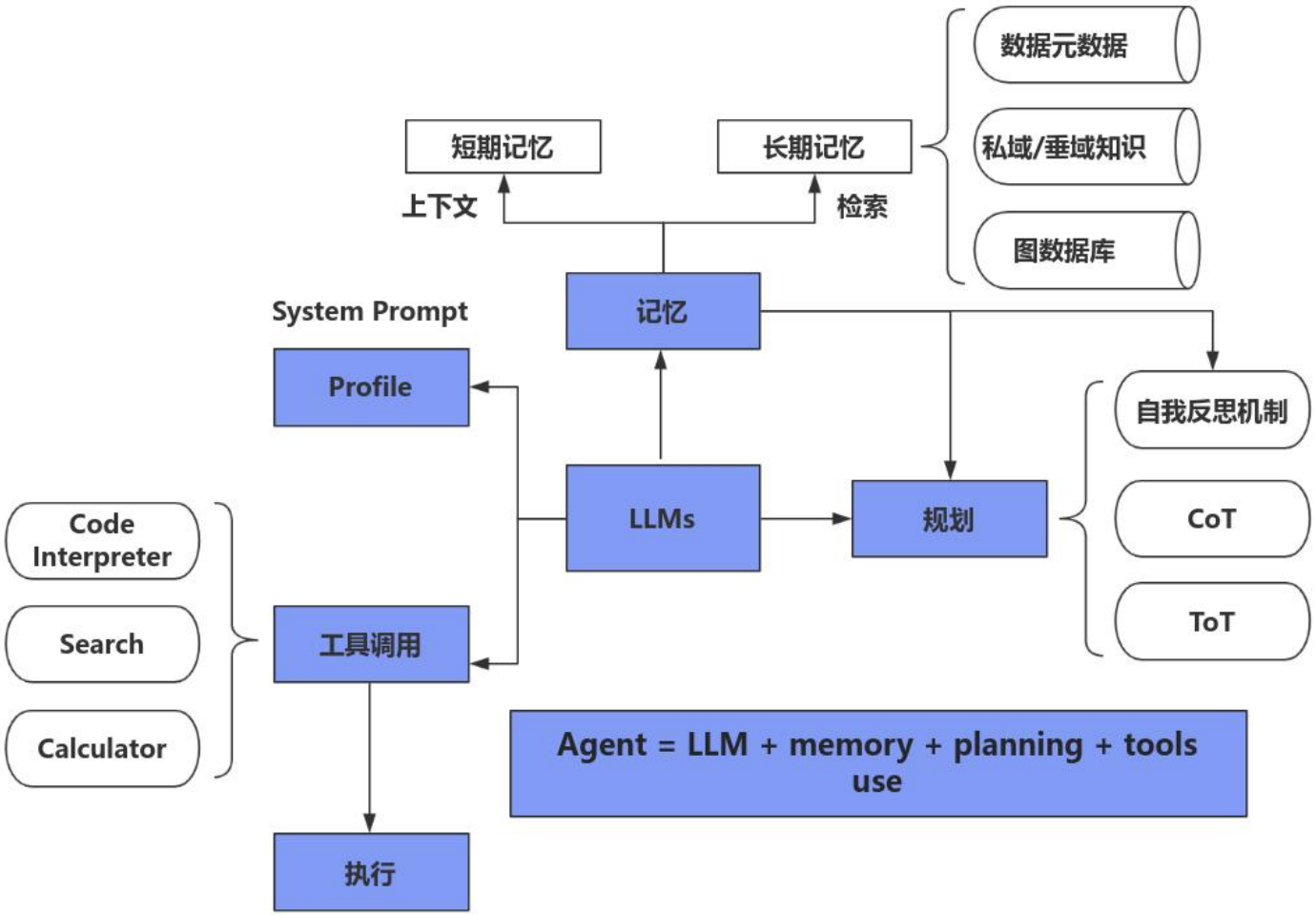
在人工智能领域，人们所指的Agent可以实现自主的意图理解、规划拆解问题、并调用合适的工具完成目标。

自主性

Agent智能体能独立运行，不需要人为干预，可以根据预设规则或学习算法自主决策

社会性

多个Agent智能体能够相互通信、协调与合作，共同完成任务。



反应性

Agent智能体能感知环境变化，并对其作出反应，通过不断调整自身行为以适应环境。

适应性

Agent智能体能通过学习、进化等方式不断提升自身性能，更好地完成任务。

Contents
目录

toC vs. toB

	通用场景 toC	企业内场景 toB
领域	解决通用领域的问题，常识性问题，比如旅游推荐、写作辅助等，一般采用基座大模型	解决某领域的专业性问题，需要结合垂域和私域内容解决问题，一般采用微调的手段
架构	一般采用Prompt+大模型端到端解决问题	需对大模型进行微调、结合传统的搜推算法、现有工具串联组合调度以支撑复杂业务场景
严谨性	追求输出的创意性、多样性，严谨性要求低	对准确性要求很高（85线是企业可尝试的基准线）、同时要求输出稳定、完整和结果可解释
安全性	大多是公开数据用来训练基座模型，但是需要关注种族歧视和伦理道德	核心关注数据安全和隐私问题

数据、知识、场景是企业落地Agent的关键要素

企业生产往往都是复杂场景

2024W13，XX产品在IOS端的销售额是多少?
难点：时间识别、维度识别、指标识别、过滤条件识别

2024W13，XX产品在IOS端卖的怎么样？同比上一个周是降了还是升了?
难点：时间识别、维度识别、指标识别、语义模糊、多任务拆解

简单场景

简单

中等

偏难场景

偏难

复杂

场景复杂度，由易到难

中等场景

2024W13，XX产品在IOS端卖了多少钱?
难点：时间识别、维度识别、指标识别、语义模糊

复杂场景

2024W13，大盘的xx指标是否正常？如果有异动，是由什么导致的？可以用连环替代法来分析。
难点：时间识别、维度识别、指标识别、语义模糊、多任务拆解、连环替代法、计算、归因分析

Contents
目录

各个场景用什么大模型合适?

测评方法

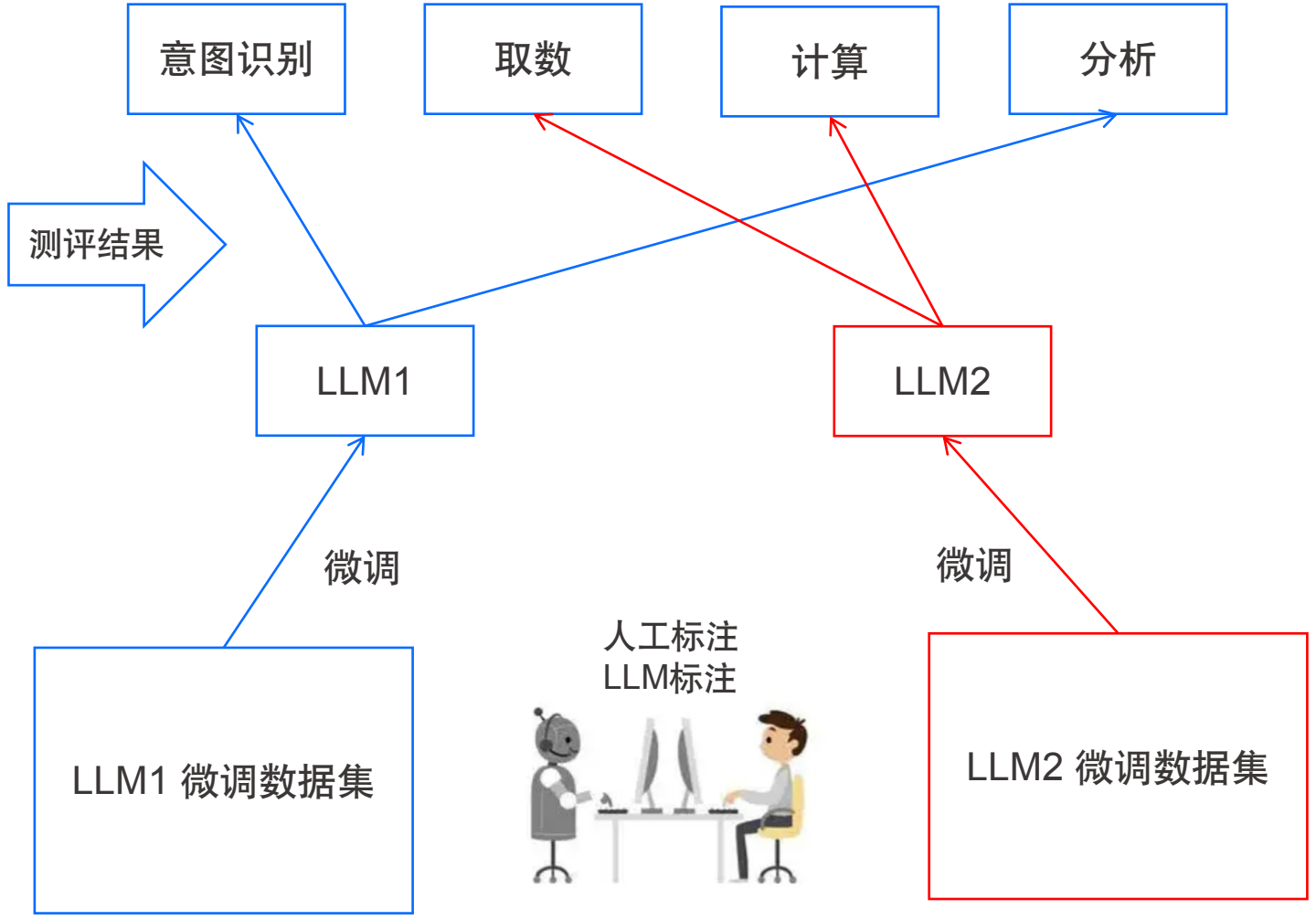
人工测评　　自动测评

测评集

RAG top10召回测评集
NL2params测评集
工具调用能力测评集
code能力测评集（NL2SQL、NL2Python）
通用能力测评集

测评内容

自然语言处理
鲁棒性、真实性
社会科学
数学、通用科学和工程
代理应用



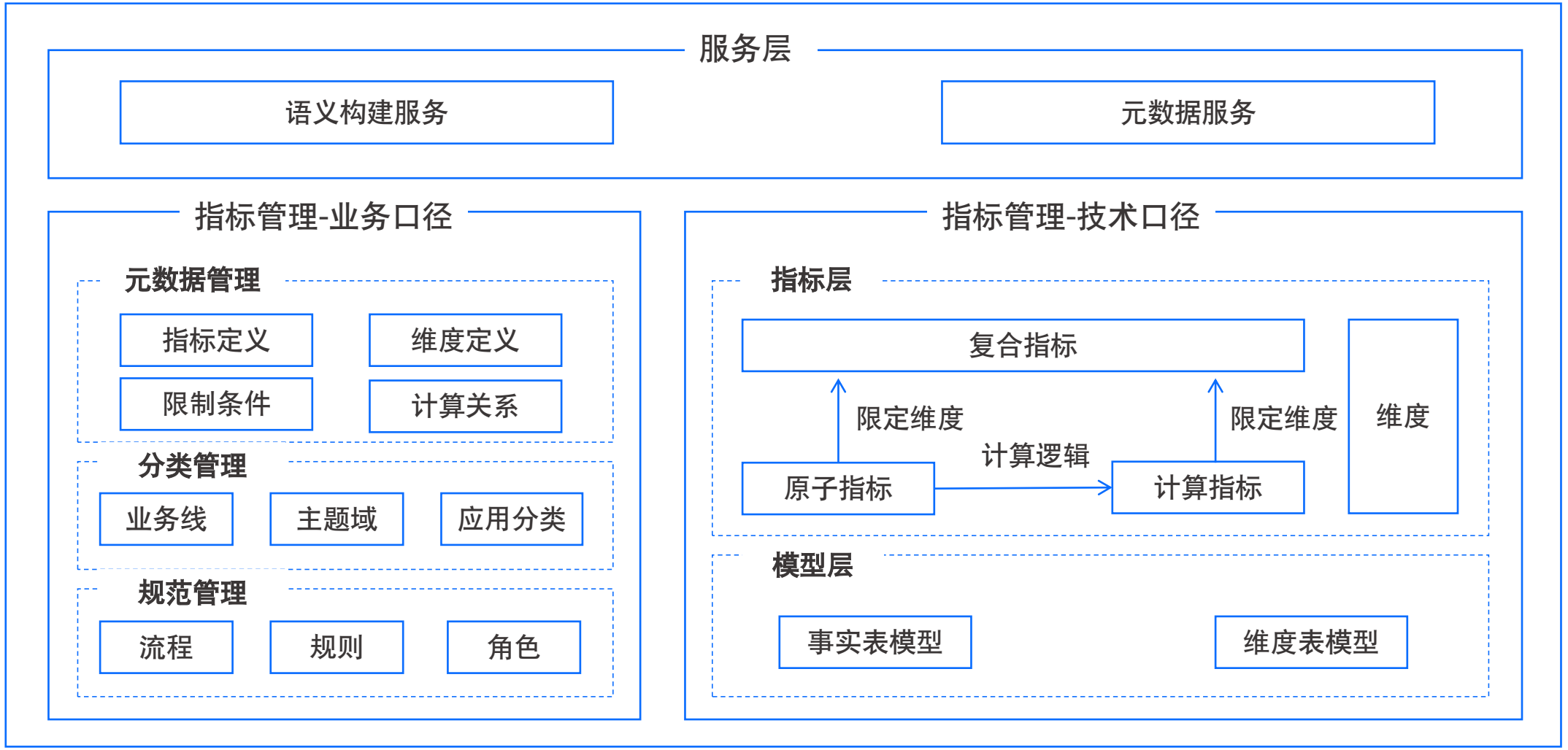
知识可信加工

Garbage in，Garbage out

Garbage in，Quantity out

以上是通用的清洗模块和流程，可以保障知识库入库的基线，若要更好的效果，还需要针对公司内、领域内流行格式、排版做对应的清洗规则。

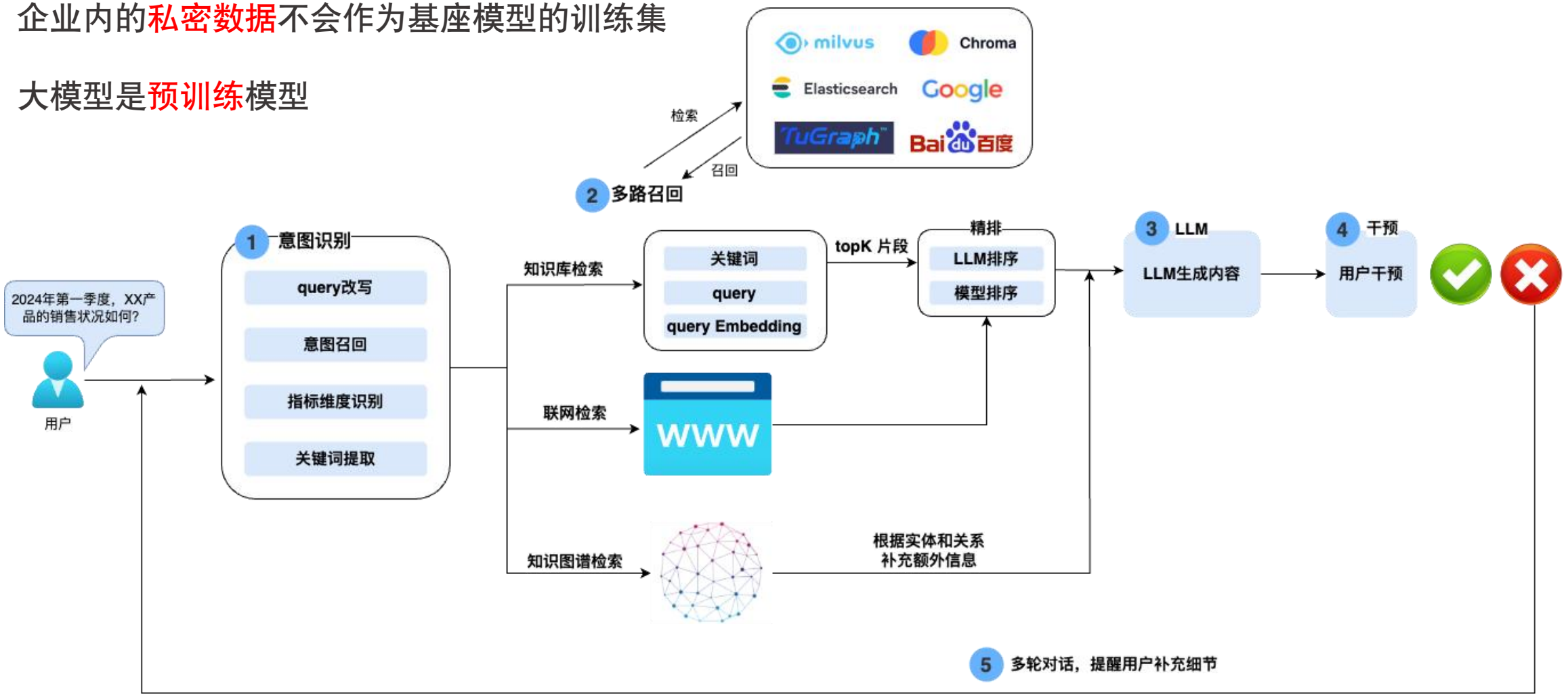
指标管理平台


服务层
语义构建服务
元数据服务
指标管理-业务口径
元数据管理
指标定义
维度定义
限制条件
计算关系
分类管理
业务线
主题域
应用分类
规范管理
流程
规则
角色
指标管理-技术口径
指标层
复合指标
限定维度
计算逻辑
限定维度
维度
原子指标
计算指标
模型层
事实表模型
维度表模型

基于知识图谱的RAGs

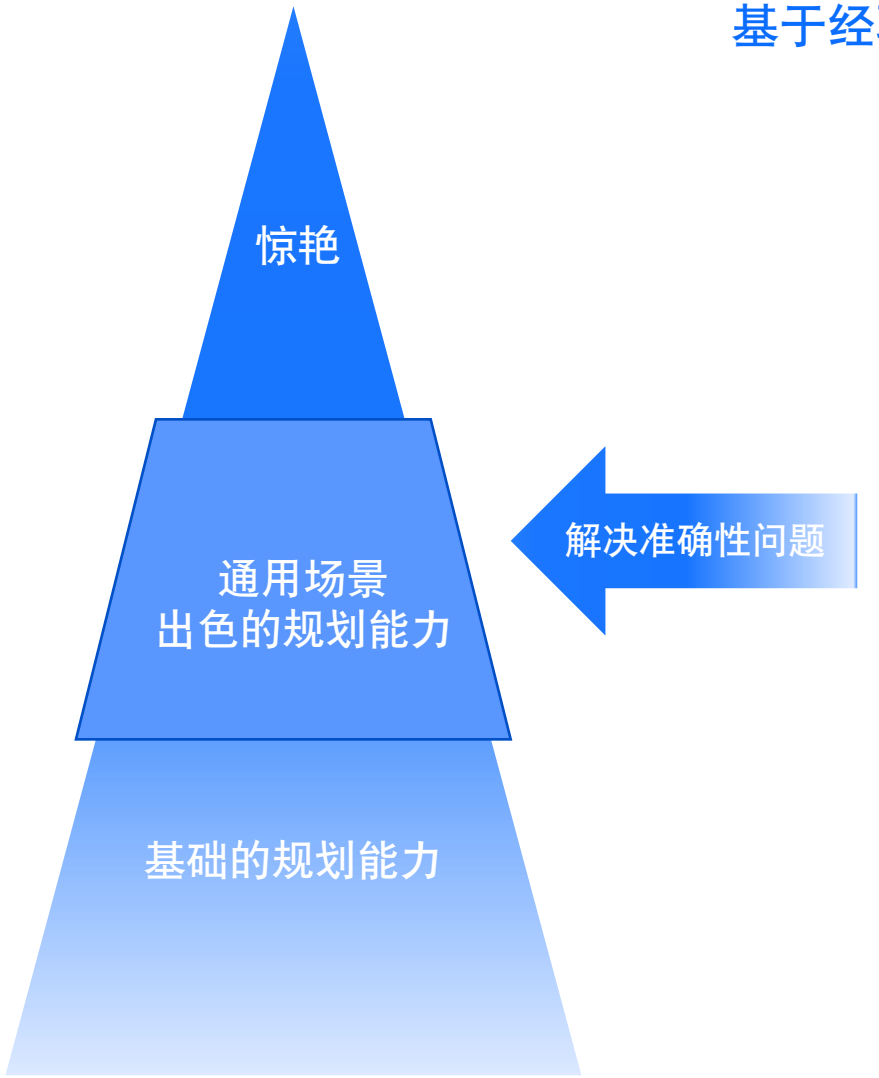
企业内的私密数据不会作为基座模型的训练集

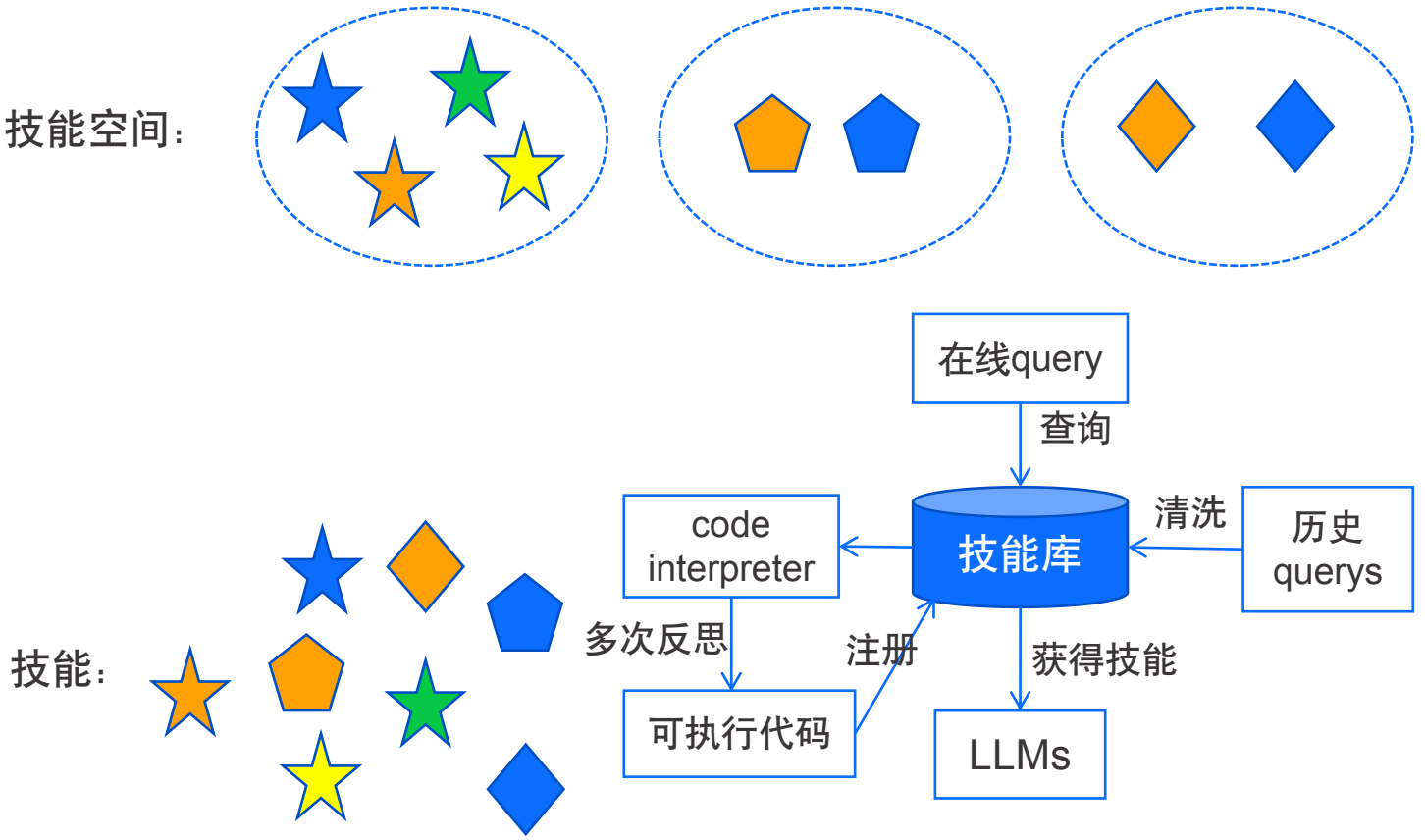
大模型是预训练模型



技能库与技能空间

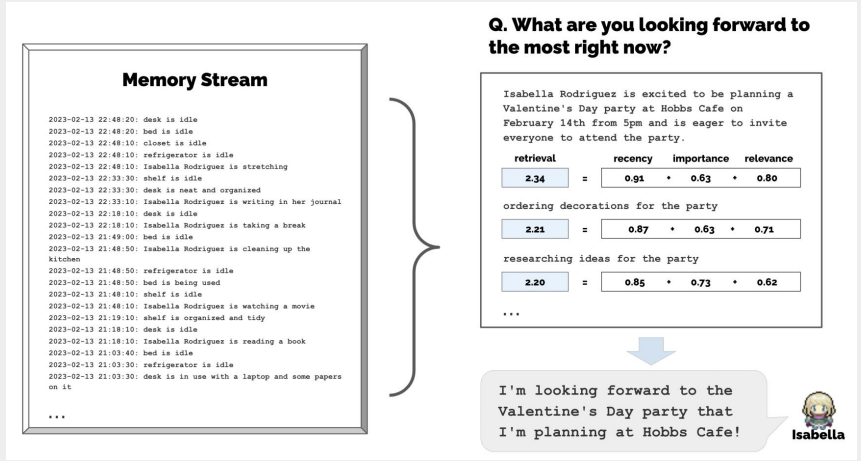
基于经验，枚举通用场景的问题，缩小问题空间





记忆召回

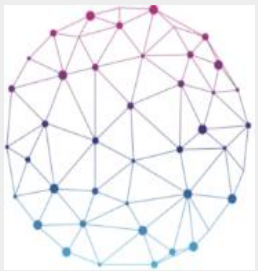
短期记忆召回



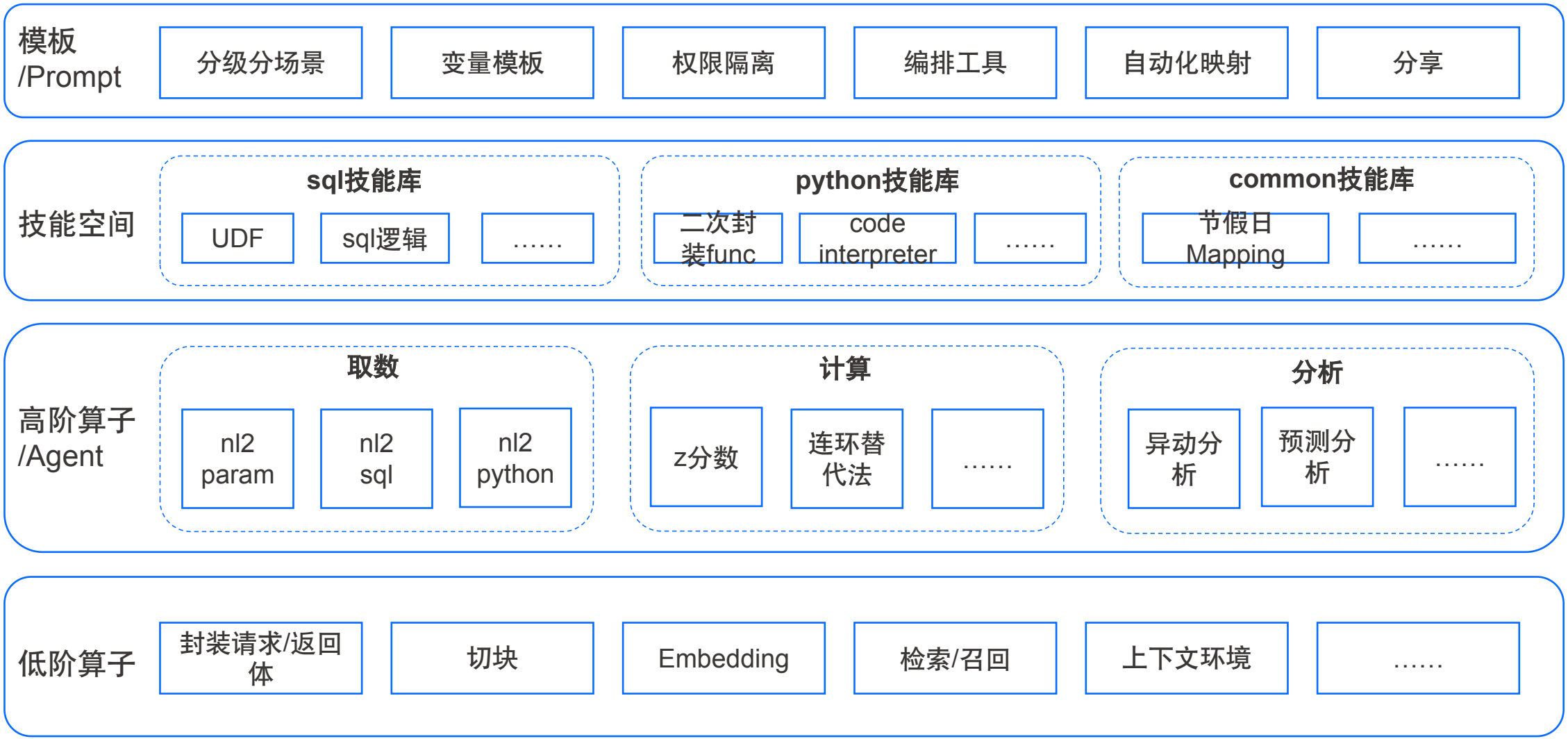
$$\gamma_E = \alpha_{recency} \cdot S_{recency} + \alpha_{relevance} \cdot S_{relevance} + \alpha_{importance} \cdot S_{importance}$$

取数和分析重要度高，计算重要度低

长期记忆召回

字典 keywords	语义记忆 embedding	知识图谱 graph
elastic	Milvus	
关键词索引 BM25算法	语义相似匹配	关系检索 属性检索

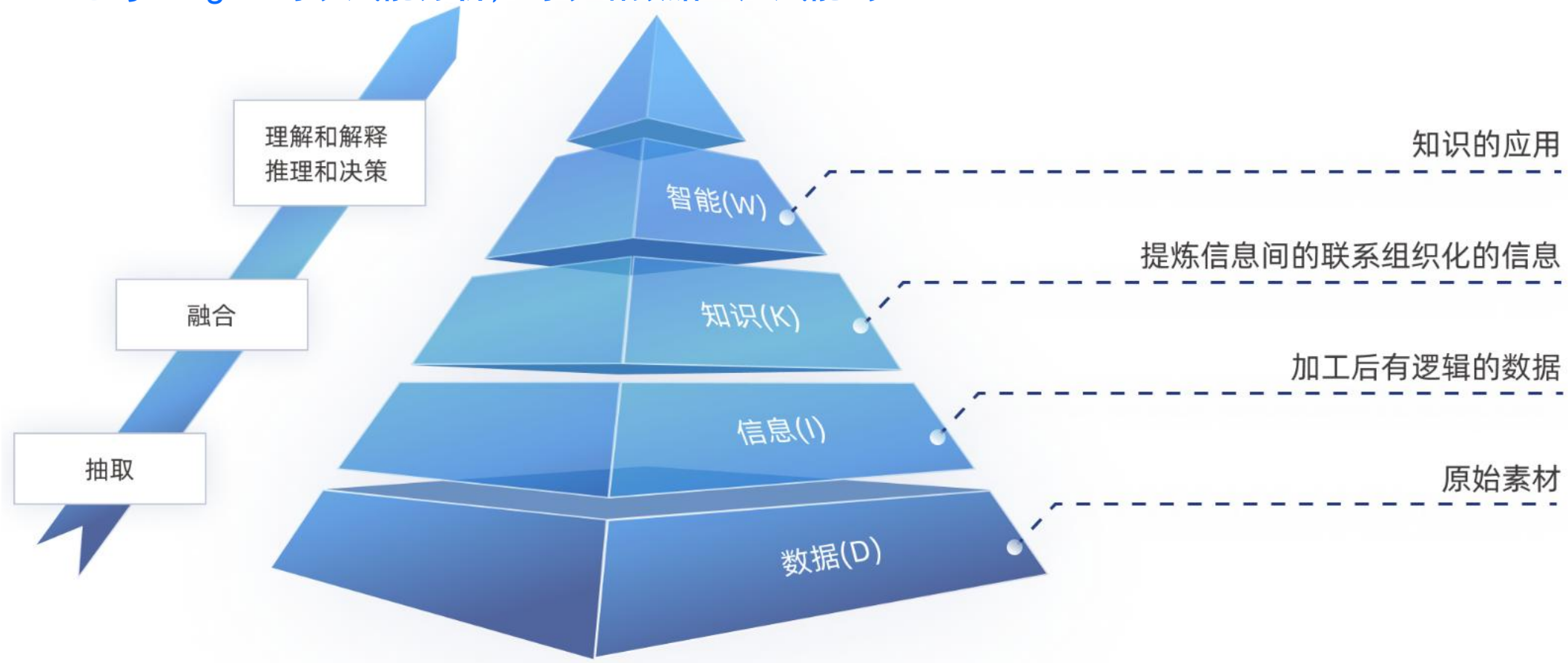
复杂场景解决方案：Agentic WorkFlow



总结

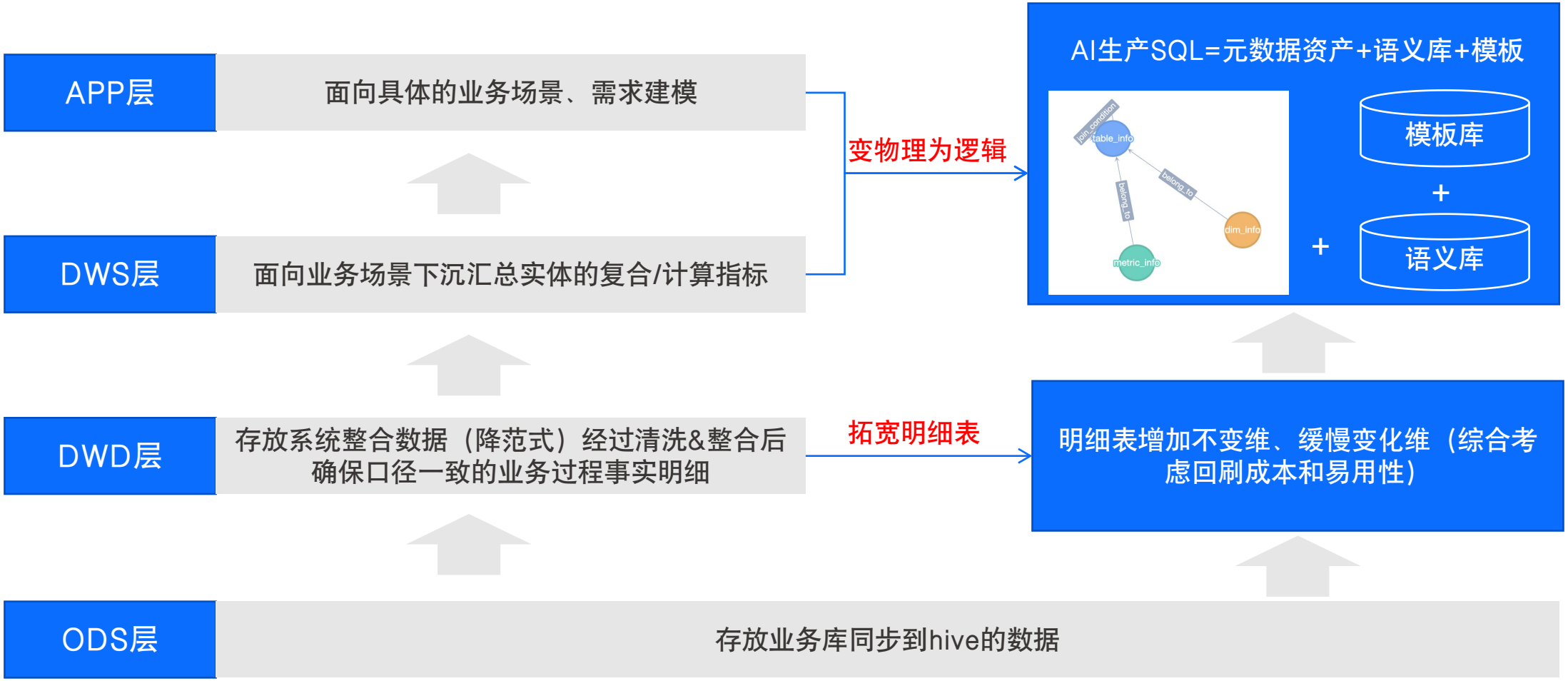
之前会分析的人，可以通过Tigi将方法论具象化、模板化。
之前不会分析的人，在Tigi的加持下，变得会分析。

思考：Agent可以赋能分析，可以给数据生产赋能吗?



Contents
目录

变革一：做厚数据集市 -> 做宽明细数据层



变革二：进一步挖掘元数据的价值

人能看懂 → AI能看懂

数据治理

统一metadata，保障数据一致性和准确性，驱动数仓建模规范落地
优化查询逻辑，保障数据查询及时性

应用治理

通过热度元数据，可以识别冷热看板，为后续的看板整合下线提供输入

数据生产

将元数据资产整合成语义库、模板库和知识图谱直接作用于AI生成SQL的输入

AI加持下，元数据将直接应用于数仓生产

传统元数据的价值体现在数据和应用的治理上

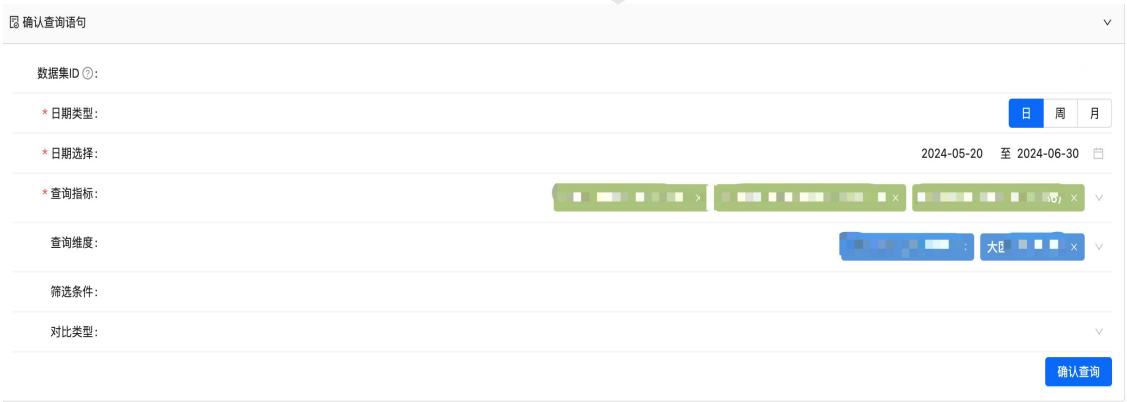
元数据（用户历史的query请求、表schema、指标业务口径、计算公式、维度枚举值、血缘、热度、查询耗时等）

Contents
目录

大模型幻觉

流程上每个节点的幻觉都会累加

思路



关键环节需要人工二次确认，不断优化人工确认环节的用户体验（减少确认次数）

RAGs建模

元数据采集问题	元数据建模不规范问题	元数据重复存储问题
元数据散落在各个平台上，治理平台、指标管理平台，数据挖掘加工的结果	元数据组织模式尚处于探索中，多种范式并存，给使用带来复杂度	元数据存储了多份，并且随着存储引擎的增加而增加

思路：

元数据采集同步规范
元数据加工规范
元数据建模规范

探索新型数据库
(兼备Embedding和图构建)

防安全风险

训练中敏感数据清洗

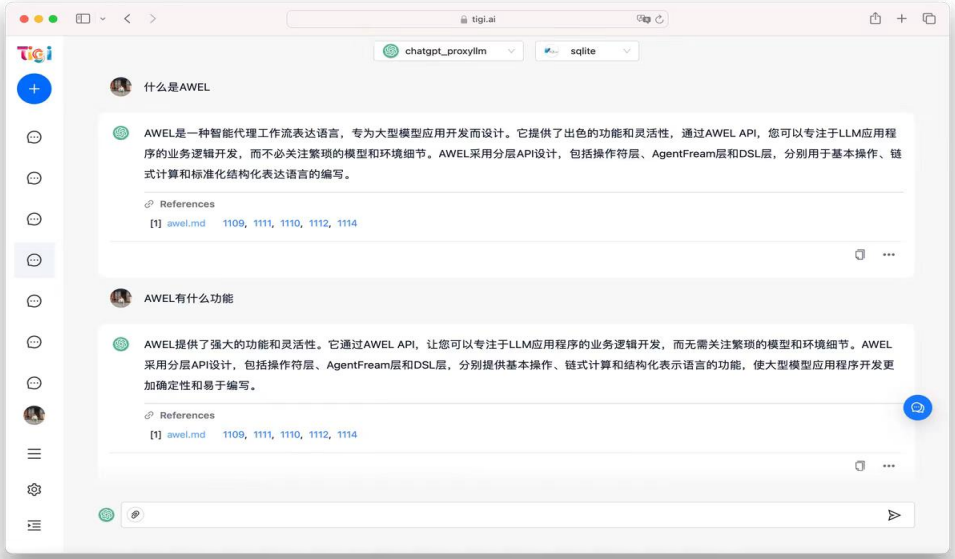
宗旨：用户隐私信息不入训练集，密文信息训练时不降低可用性

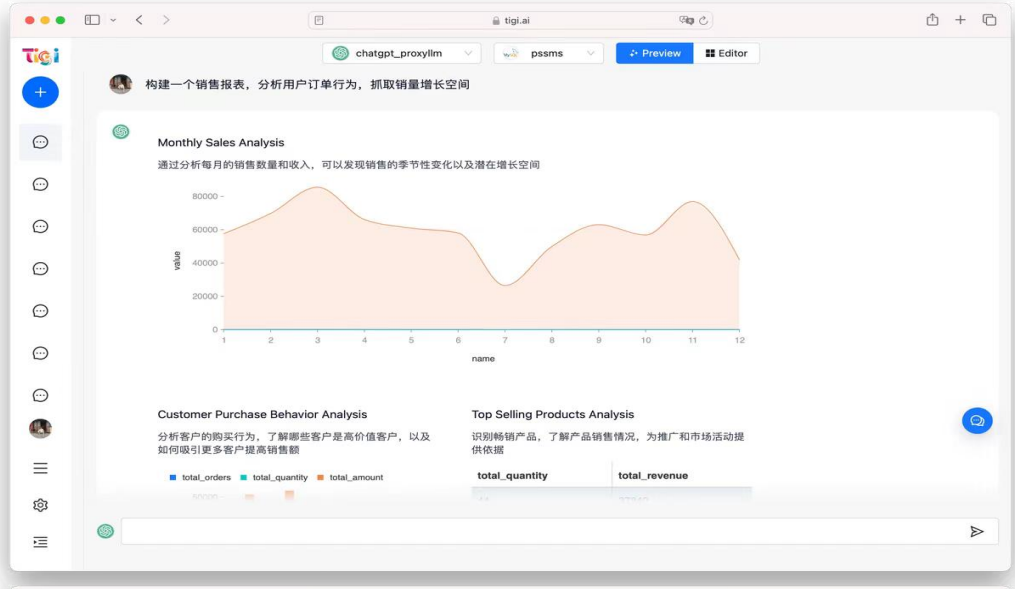
推理中敏感数据防外泄

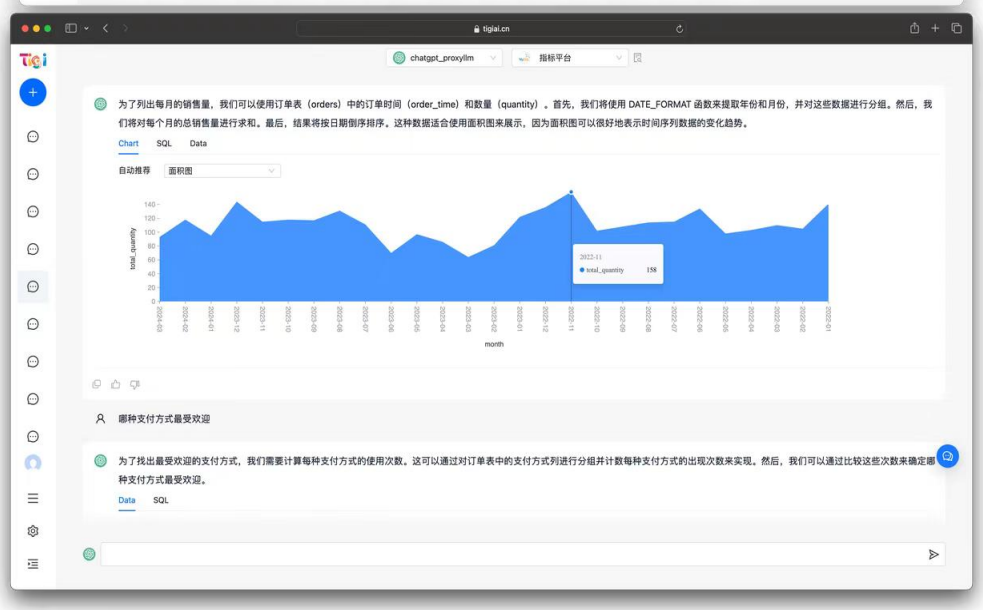
宗旨：用户不感知加密解密过程，大模型基于密文输入进行推理不损失效果

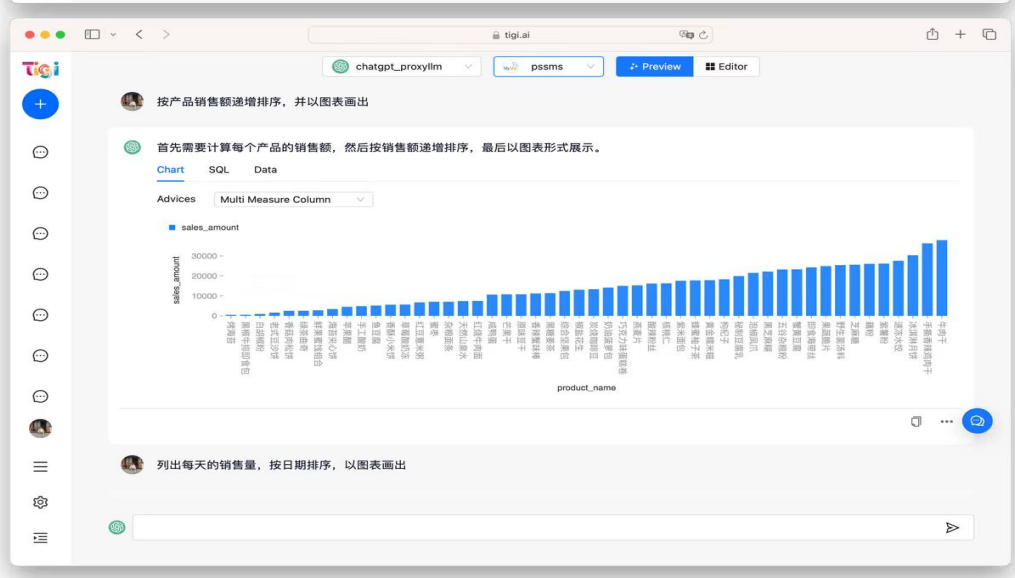
效果测评	ABTest对比加密前后的大模型输出效果
加密组件	加密码表Mapping、数值混淆加密等
识别隐私数据组件	正则、字典表、模糊匹配、OCR等
输入数据	个人信息：身份证、手机号、Email、地址等 ID类：订单ID、商品ID、门店ID等 标识类：MAC地址、IP地址、WX_OPEN_ID等

产品演示

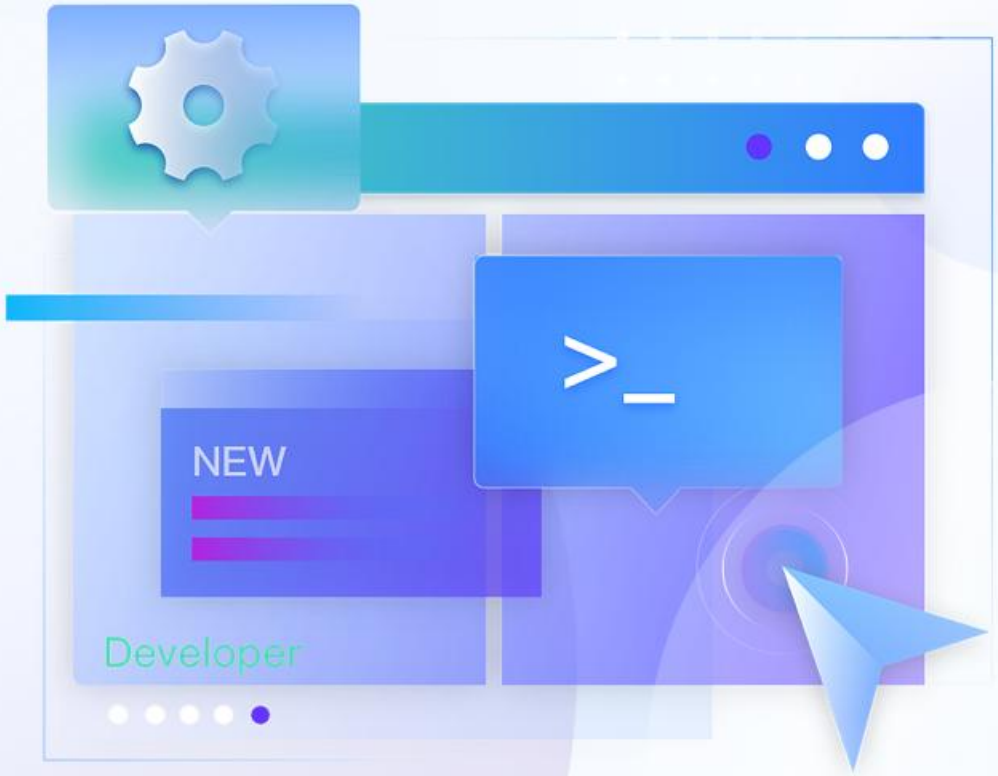

tigi.ai
chatgpt_proxyllm
sqlite
什么是AWEL
AWEL是一种智能代理工作流表达语言，专为大型模型应用开发而设计。它提供了出色的功能和灵活性，通过AWEL API，您可以专注于LLM应用程序的业务逻辑开发，而不必关注繁琐的模型和环境细节。AWEL采用分层API设计，包括操作符层、AgentFream层和DSL层，分别用于基本操作、链式计算和标准化结构化表达语言的编写。
References
[1] awel.md 1109, 1111, 1110, 1112, 1114
AWEL有什么功能
AWEL提供了强大的功能和灵活性。它通过AWEL API，让您可以专注于LLM应用程序的业务逻辑开发，而无需关注繁琐的模型和环境细节。AWEL采用分层API设计，包括操作符层、AgentFream层和DSL层，分别提供基本操作、链式计算和结构化表示语言的功能，使大型模型应用程序开发更加确定性和易于编写。
References
[1] awel.md 1109, 1111, 1110, 1112, 1114


tigi.ai
chatgpt_proxyllm
pssms
Preview
Editor
构建一个销售报表，分析用户订单行为，抓取销量增长空间
Monthly Sales Analysis
通过分析每月的销售数量和收入，可以发现销售的季节性变化以及潜在增长空间
value
name
Customer Purchase Behavior Analysis
分析客户的购买行为，了解哪些客户是高价值客户，以及如何吸引更多客户提高销售额
total_orders
total_quantity
total_amount
Top Selling Products Analysis
识别畅销产品，了解产品销售情况，为推广和市场活动提供依据
total_quantity
total_revenue


tigai.cn
chatgpt_proxyllm
指标平台
为了列出每月的销售量，我们可以使用订单表（orders）中的订单时间（order_time）和数量（quantity）。首先，我们将使用 DATE_FORMAT 函数来提取年份和月份，并对这些数据进行分组。然后，我们将对每个月的总销售量进行求和。最后，结果将按日期倒序排序。这种数据适合使用面积图来展示，因为面积图可以很好地表示时间序列数据的变化趋势。
Chart
SQL
Data
自动推荐
面积图
total_quantity
month
2022-11
total_quantity 158
哪种支付方式最受欢迎
为了找出最受欢迎的支付方式，我们需要计算每种支付方式的使用次数。这可以通过对订单表中的支付方式列进行分组并计数每种支付方式的出现次数来实现。然后，我们可以通过比较这些次数来确定哪种支付方式最受欢迎。
Data
SQL


tigi.ai
chatgpt_proxyllm
pssms
Preview
Editor
按产品销售额递增排序，并以图表画出
首先需要计算每个产品的销售额，然后按销售额递增排序，最后以图表形式展示。
Chart
SQL
Data
Advices
Multi Measure Column
sales_amount
sales_amount
product_name
列出每天的销售量，按日期排序，以图表画出

期待与大家
一起学习和交流



微信关注Tigi，了解更多数智决策解决方案

Tigi体验链接：https://tigiai.cn/

一起探索Agent在数据分析领域的落地