AIGC 在 DataLeap 大数据研发治理套件的实践

火山引擎 DataLeap 资深架构师 / 王慧祥

火山引擎 DataLeap 技术专家 / 丁桂涛

目录

- DataLeap整体介绍
- AIGC在DataLeap数据资产方向的实践 —— 找数助手
- AIGC在DataLeap数据研发方向的实践 —— 研发助手
- 未来规划

DataLeap整体介绍

DataLeap 大数据研发治理套件

数据研发全链路管理

整合全域数据，支持20+多源异构数据集成，灵活对接各类业务系统。敏捷开发CI/CD，覆盖需求、开发、测试、发布、运维等研发全链路管理。

数据全生命周期治理

结合基线监控、数据质量、 SLA治理等能力，提供事前预警、事中处理、事后复盘及推荐优化的全生命周期的数据治理能力

沉淀数据规范

统一数据标准及数据查询出口，沉淀数仓建设规范的最佳实践，提升数据开发效率，保证数据质量，快速精准为业务赋能

保障数据安全

更细粒度的行、列权限控制，表及字段级别的血缘管理，加上行为监控等功能，构成真正意义上的数据安全屏障

多云多引擎

提供公有云PaaS服务及灵活的私有化部署方案。可低成本、高效适配客户已有大数据平台，控制迁移成本，降低业务影响

DataLeap智能助手

火山引擎大数据研发治理套件DataLeap智能助手基于自研方舟MAAS，经过海量代码和语料训练，支持根据自然语言理解，提供拟人化的逻辑推理总结、自动生成代码构建优化和管理。资产知识库经过对话式语义检索，高效聚焦全链路的搜索过程，以低门槛、自助式的数据探索，极大提升企业数据研发和数据消费能力

找数助手

对话式的数据检索能力，解决用户找数据与用数据诉求。通过AI加持推动让搜索过程更聚焦。同时伴随模型语义理解能力的逐步提升，其全链路的检索效率更高，使得资产以低成本管理、促进自助式数据消费

研发助手

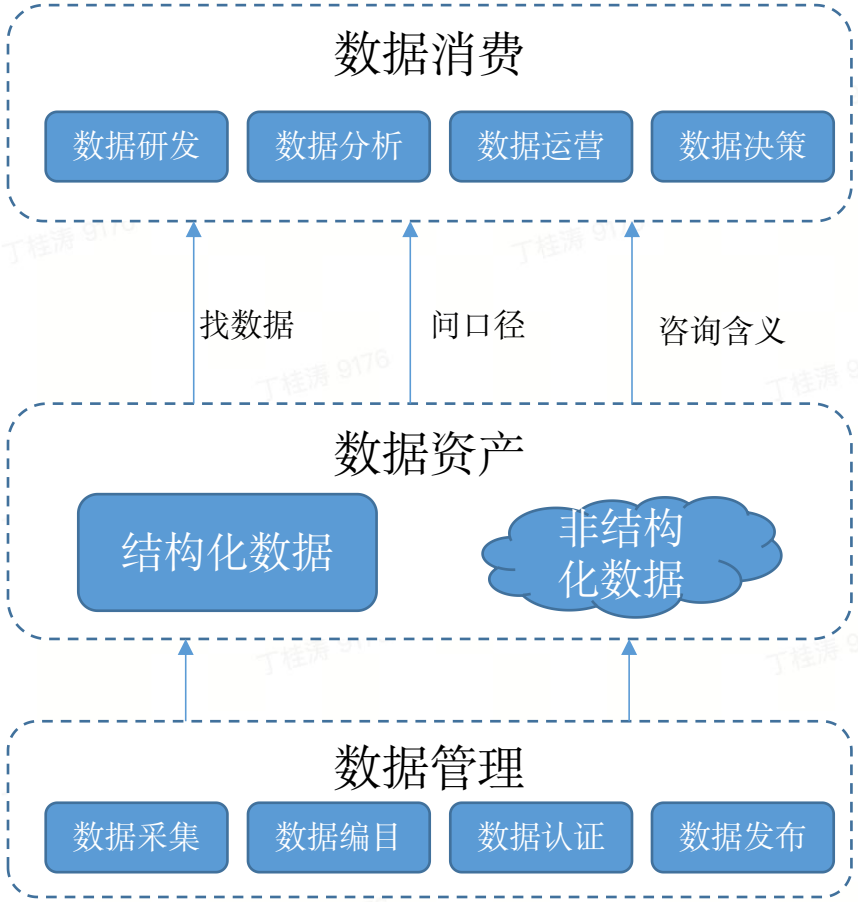
实现通过自然语言描述，自动生成代码，针对已有的代码可以自动实现自动生成、修复，优化、解释与注释等。对话式方式进行文档搜索、函数使用、代码示例等问题咨询。助力平台用户减少基础开发工作量、提升开发效率。

更便捷地生产数据、应用数据，实现更普惠的数据消费，为数字化提供现实基础。
打破了语言障碍，极大程度降低了数据开发的准入门槛，同时让专业数据研发人员更聚焦复杂场景的需求

AIGC 在 DataLeap 数据资产方向的实践 – 找数助手

数据消费问题

- 数据资产建设的核心目的是促进数据消费，使数据价值最大化
- 在海量数据场景下，如何准确、高效的找到数据是数据消费的前提
- 数据的查找和使用强依赖于**业务知识**的处理
- 结构化组织数据表达能力有限，在数据管理侧信息丢失
- 基于关键词的检索能力受限，在数据消费侧信息丢失



关键词检索

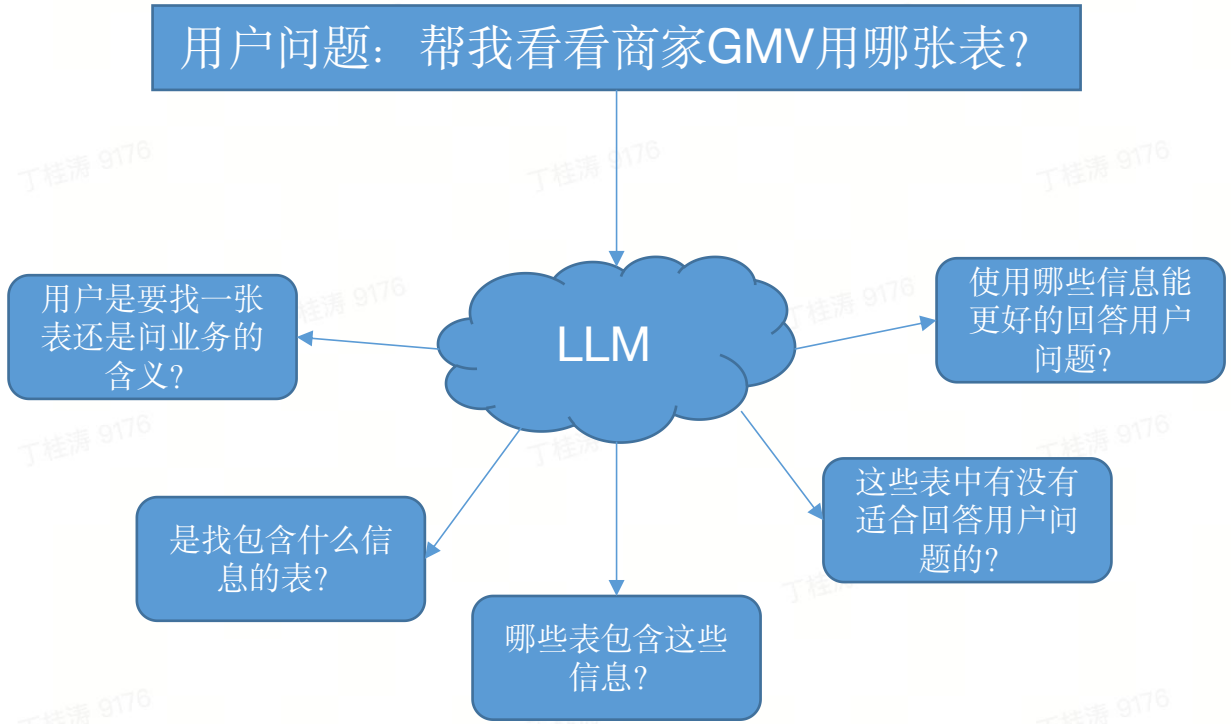
?

结构化组织

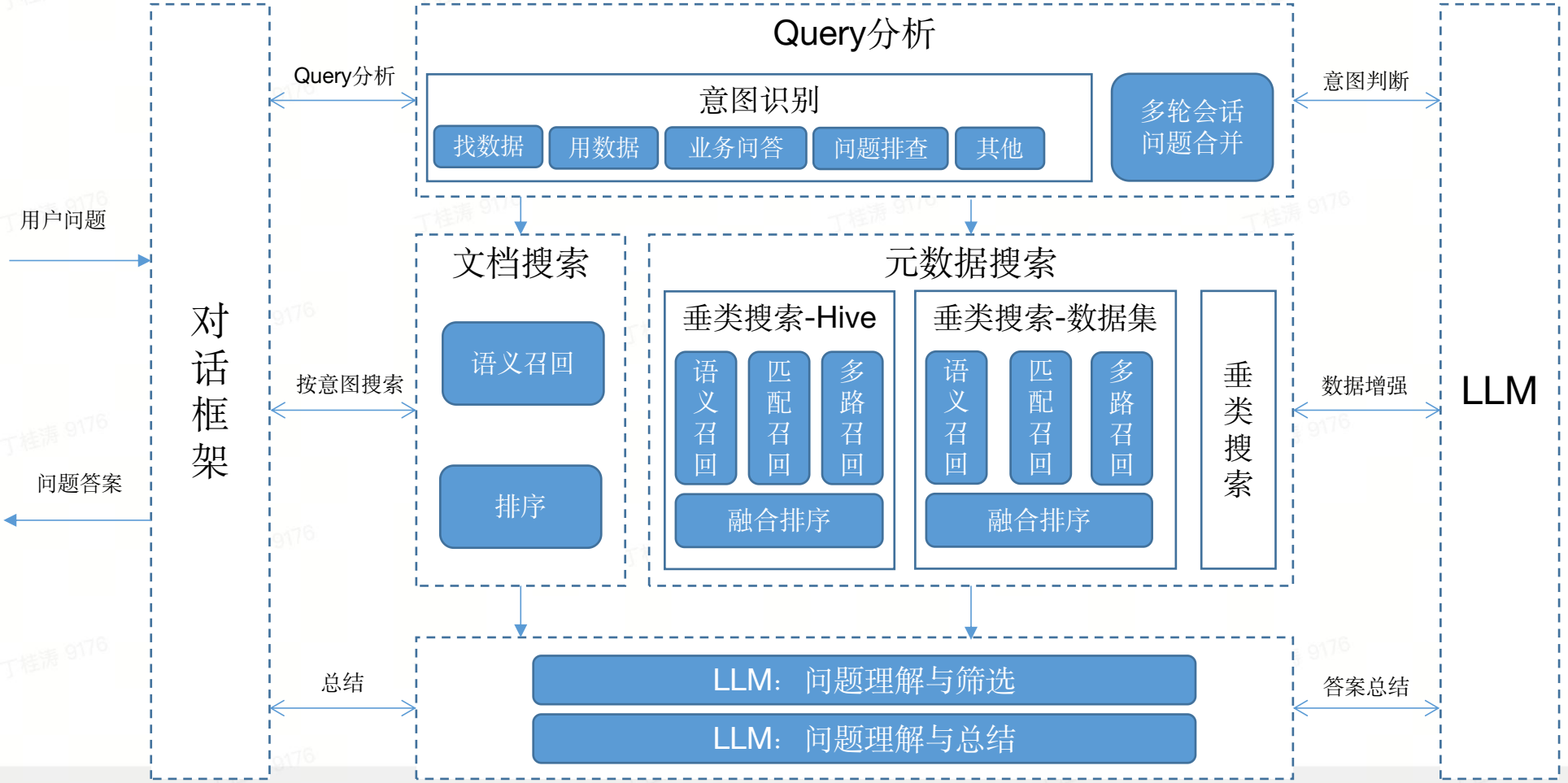
LLM在找数场景能做什么

- LLM的特性：理解、推断、生成

- 找数场景有如下复杂点：
  - 问题理解（理解）
  - 意图判断（推断）
  - 结构化元数据描述不足（生成）
  - 大量非结构化元数据沉淀于文档（生成）
  - 关键信息提取（生成）



DataLeap找数助手整体架构



问题理解

- **核心关键词提取**
  - 识别query中核心的term，提升找数准确性，提升用户体验

用户问题	关键词提取
shop_id和order_id的关系	字段/指标: shop_id 字段/指标: order_id
7日结算率	字段/指标：7日结算率
C_O率是什么意思	字段/指标：C_O率
xxx.a与yyy.b的区别是什么?	表名：xxx.a 表名：yyy.b
zzz.c的call_type有什么作用?	表名：zzz.c 字段/指标：call_type

- **多轮对话问题合并**
  - 判断用户新问题是否需要关联上一个问题信息
  - 合并多个问题为一个问题

上一轮问题	本轮问题	合并后问题
商家GMV用哪张表?	数据集呢?	商家GMV用哪个数据集?
如何查看直播间PV数据?	有没有带货粒度的数据?	有没有带货粒度的数据?
想看下DQC的报警实例用哪张表?	不要基线任务的	查看DQC的报警实例切不要基线任务的用哪张表?
数据安全等级分布用哪张表?	我想要标签粒度的	标签粒度的数据安全等级分布用哪张表?

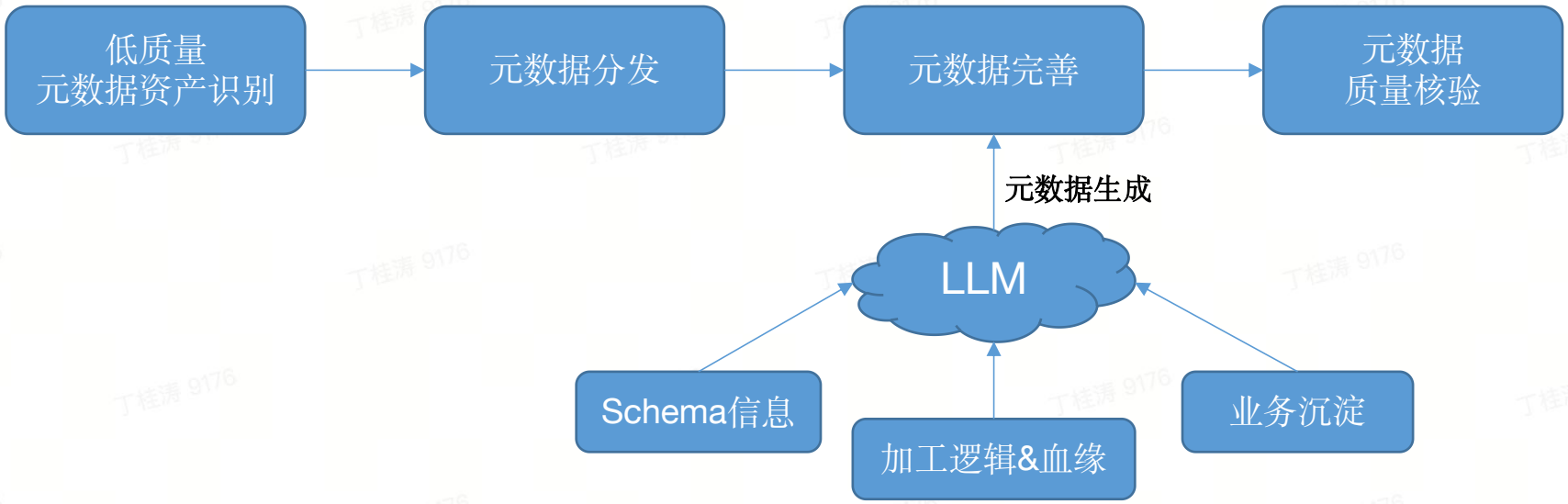
意图判断

- 依赖LLM的判断能力，判断用户的找数意图
- 基于业务真实找数场景调研，总结4大类意图

一级意图	二级意图	问题示例
找数据	找表、数据集	抖音是否有用户维度的消费视频表?
使用数据	问指标	xxx数据集中has_risk=0值的定义是什么?
	问口径	近7天直播间曝光次数口径是什么
	问区别	表xxx与表yyy中的user_id有什么区别?
业务咨询	-	什么是GMV?
问题排查	-	表xxx中的字段a为什么会有空的情况?
其他	闲聊	

- Prompt工程+模型精调

元数据生成



元数据质量衡量

- 信息填充度
- 信息丰富度

元数据治理分发

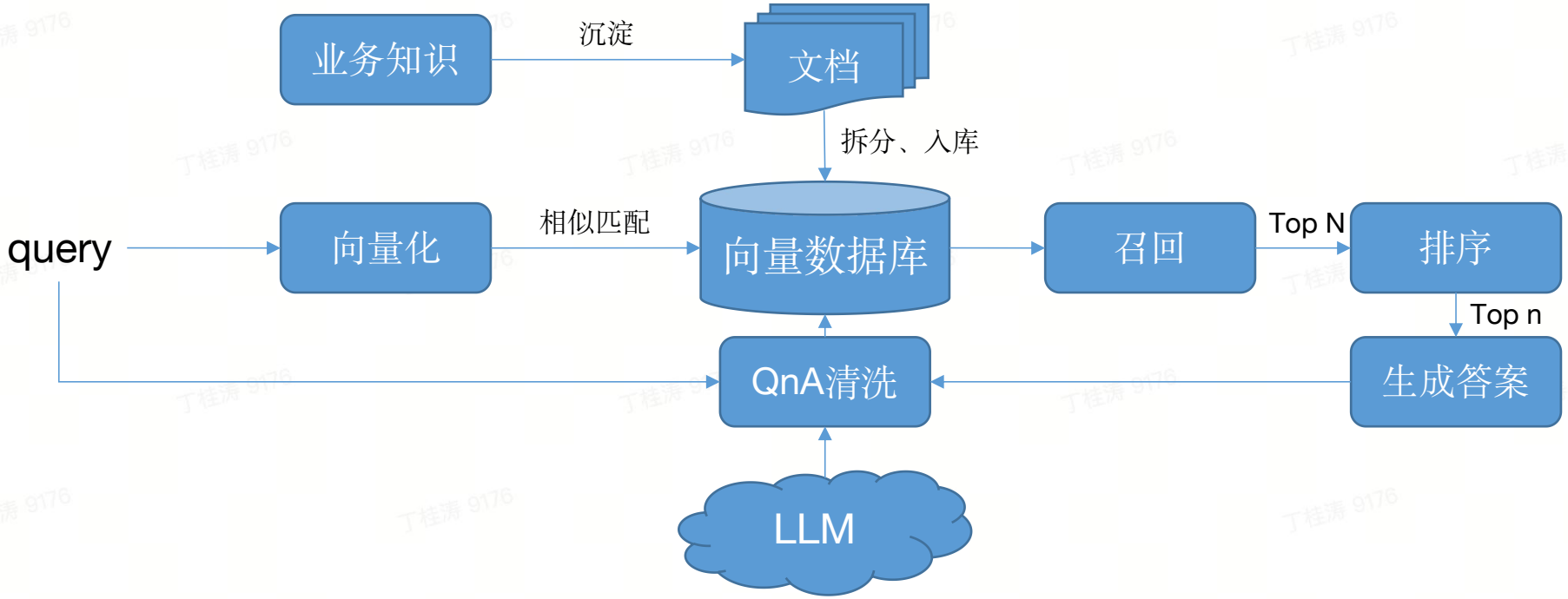
- 资产识别&圈选
- 任务分发
- 元数据完善

LLM：元数据生成提效

- Schema信息
- 生成逻辑代码+血缘信息
- 业务沉淀

业务知识沉淀与检索

- 文档模块化拆分，基于语义相似度匹配召回，解决非结构化业务知识问答问题
- 沉淀历史问答问题，以LLM进行清洗，丰富知识库



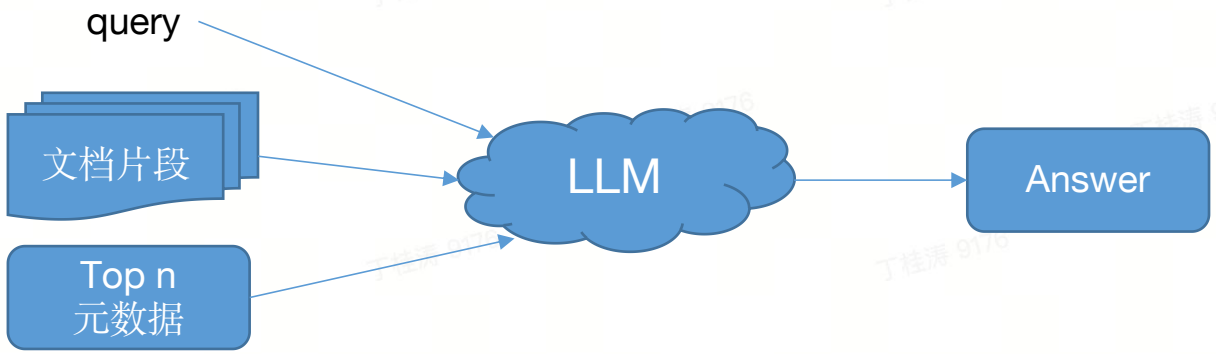
答案总结

结构化元数据总结

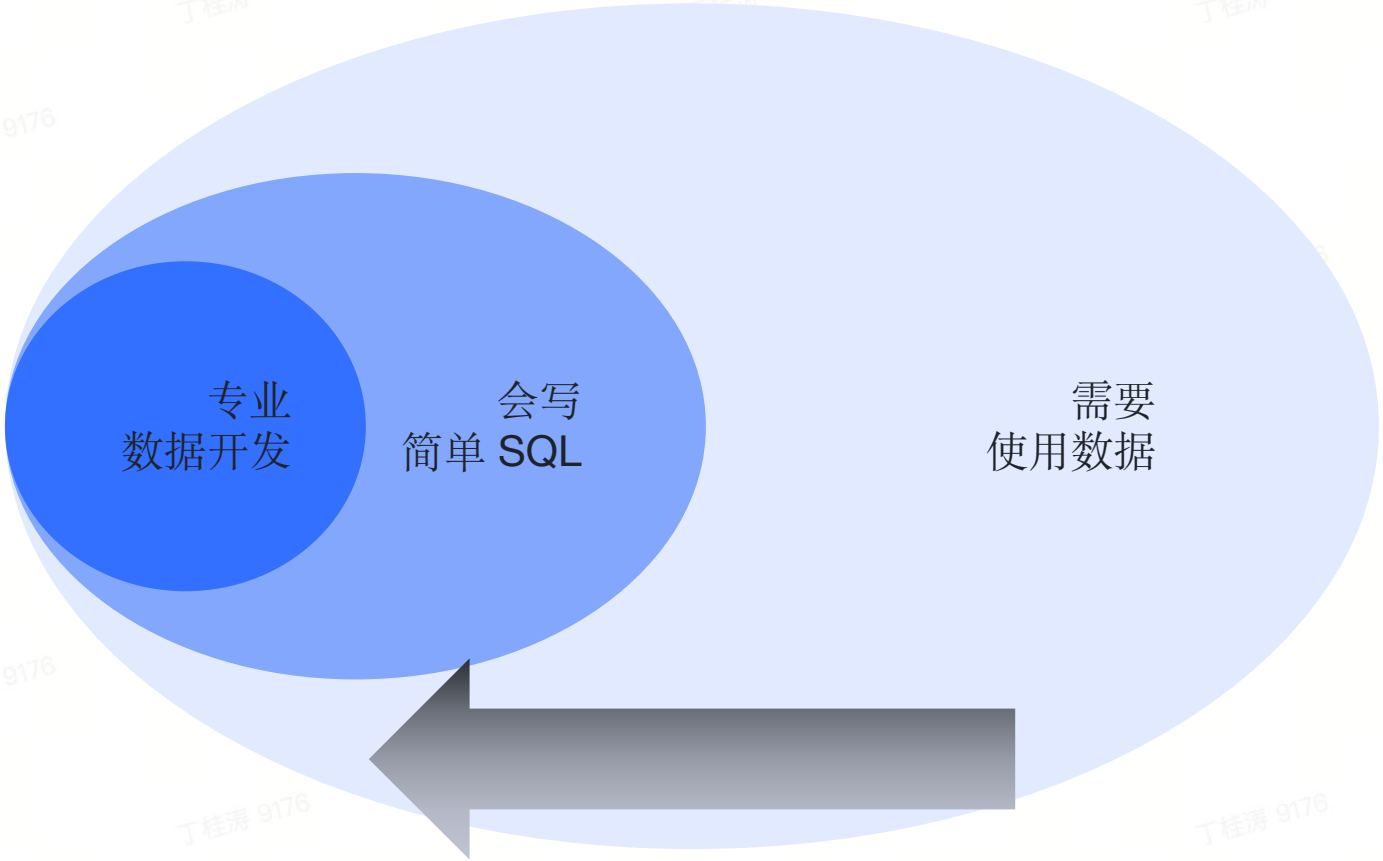
- 多轮召回+精排（Top m）
- Query + 表shcema + 增强信息
- prompt工程控制输出

非结构化文档总结

- 语义召回+相似度排序（Top n）
- Query + Top n
- prompt工程控制输出



AIGC 在 DataLeap 数据研发方向的实践 - 开发助手

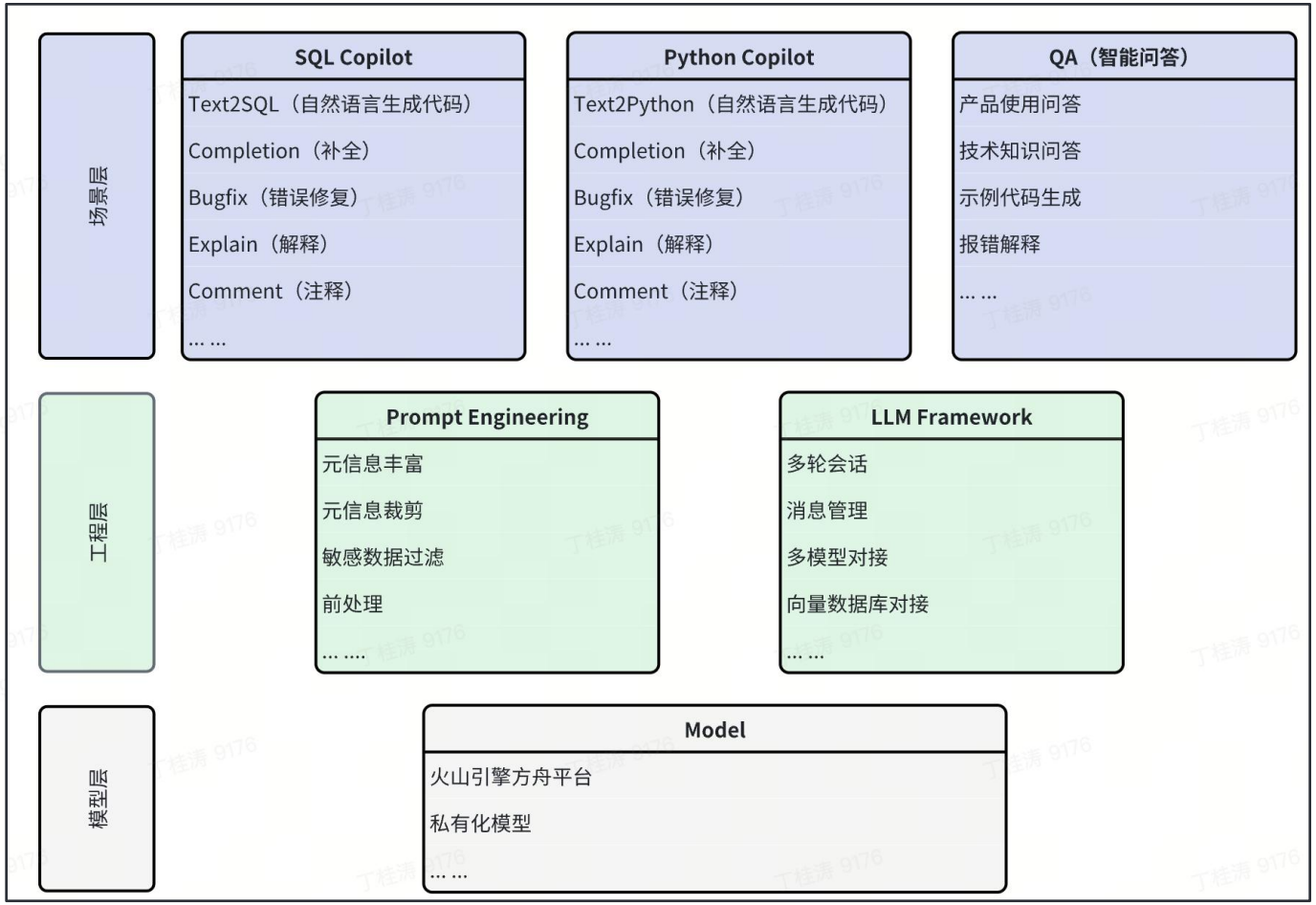


•数据平民化

•AIGC 可以降低数据开发的门槛，让需要使用数据的人离数据更近，同时也能提升专业数据开发的效率。

•数据平民化

- •提高模型准确率
- •降低 prompt 编写成本
- •减少多工具间的切换



•开发助手 - 产品架构

•场景：Coding Copilot、知识问答

•工程：Prompt Engineering、模型对接框架

•模型：支持 MaaS、私有化等多种模型

场景	交互形式	Prompt 成本	准确率要求	延迟容忍度
Text2SQL	主动提问	高	高	高
补全	被动提示	低	低	低
Bugfix	一键操作	低	高	中
问答	主动提问	中	高	中
… …				

•开发助手 - 场景设计

•需要根据场景的差异化要求，进行针对性设计、优化。

•开发助手 - Prompt Engineering

业务需求：“查询昨日销售额 top 1000 的商品信息”

用户输入：“一天”、“销量”、“1000”

```sql
WITH
-- 计算昨天的订单量
order_count AS (
    SELECT  product_id,
            count(*) AS order_count
    FROM    ods_order
    WHERE   date = '${DATE-1}'
    GROUP BY
            product_id
),
-- 对订单量进行排序并取出前1000个
top_products AS (
    SELECT  product_id,
            order_count,
            row_number() OVER(ORDER BY order_count DESC) AS RANK
    FROM    order_count
    LIMIT   1000
)
-- 连接商品表，获取商品信息
SELECT  tp.product_id,
        tp.order_count,
        tp.rank,
        p.product_name,
        p.product_description,
        p.product_price,
        p.product_stock,
        p.product_category,
        p.product_brand
FROM    top_products tp
JOIN    ods_product p
ON      tp.product_id = p.product_id
WHERE   p.date = '${DATE-1}'
```

•开发助手 - Prompt Engineering

•Prompt 模版

```
CREATE TABLE `ods_order`(
  `order_id` INT COMMENT '订单ID',
  `user_id` INT COMMENT '用户ID',
  `product_id` INT COMMENT '商品ID',
  `order_status` STRING COMMENT '订单状态',
  `order_amount` DOUBLE COMMENT '订单金额',
  `order_time` BIGINT COMMENT '订单时间')
```

•表结构填充

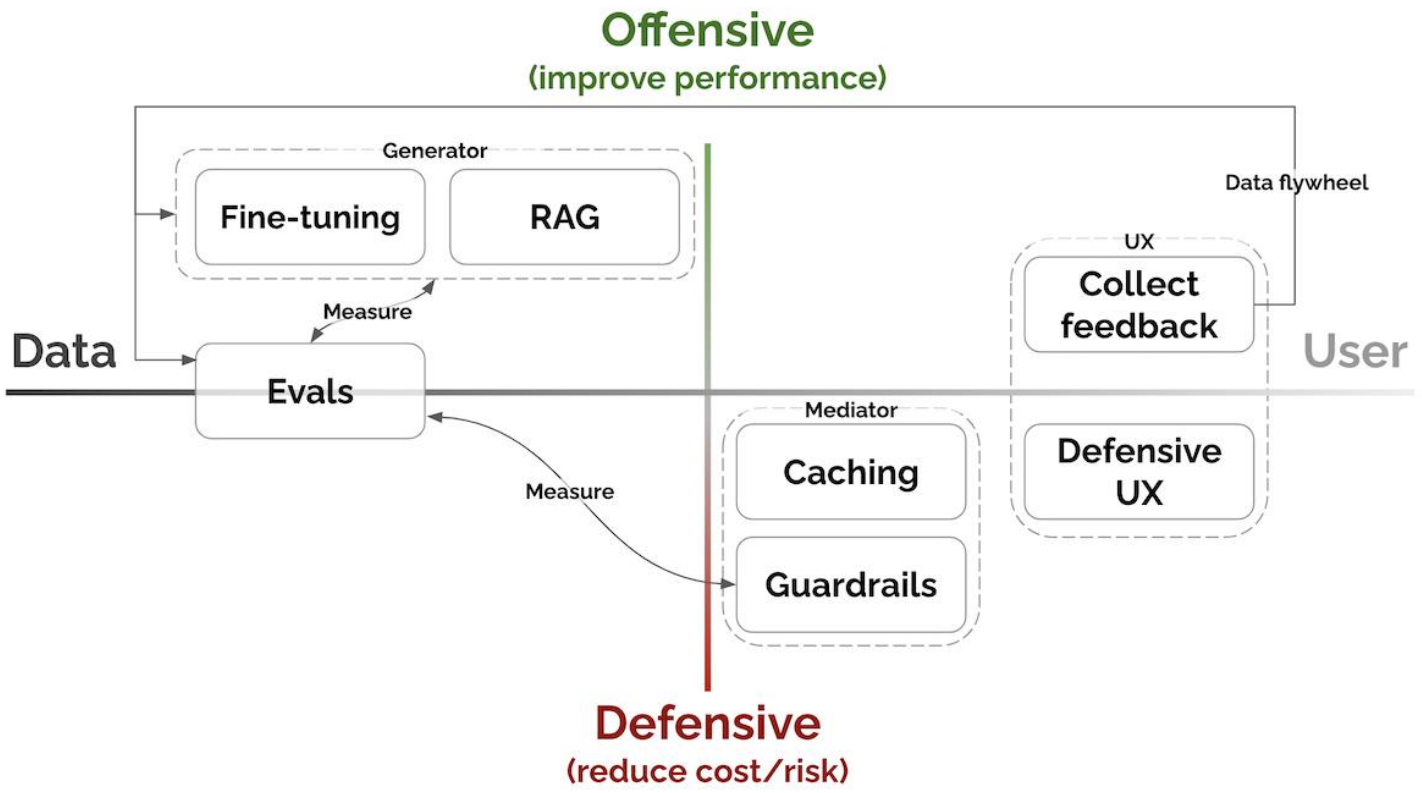
```
... ...

user: ...
assistant: ...
user: 商品信息需要包含商品描述
assistant ...
user: 查询 xxx 商品信息

... ...
```

•多轮上下文

```
  `product_id` INT COMMENT '商品ID',
  `product_name` STRING COMMENT '商品名称',
  `product_description` STRING COMMENT '商品描述',
  `product_price` DOUBLE COMMENT '商品价格',
--  `product_stock` INT COMMENT '商品库存',
--  `product_category` STRING COMMENT '商品类别',
--  `product_brand` STRING COMMENT '商品品牌',
--  `add_time` BIGINT COMMENT '添加时间'
```

•字段裁剪

•开发助手 - 准确率



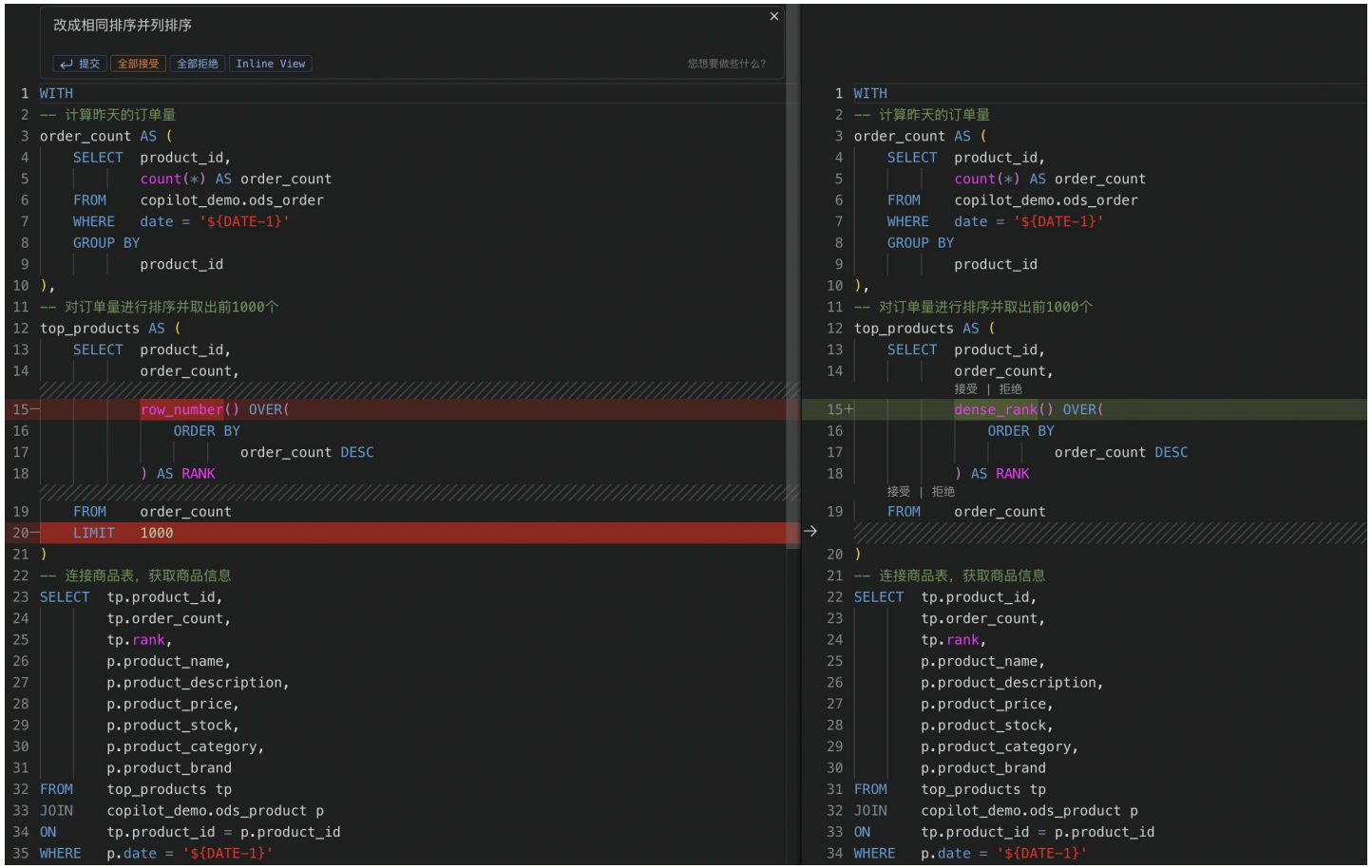
https://eugeneyan.com/writing/llm-patterns/

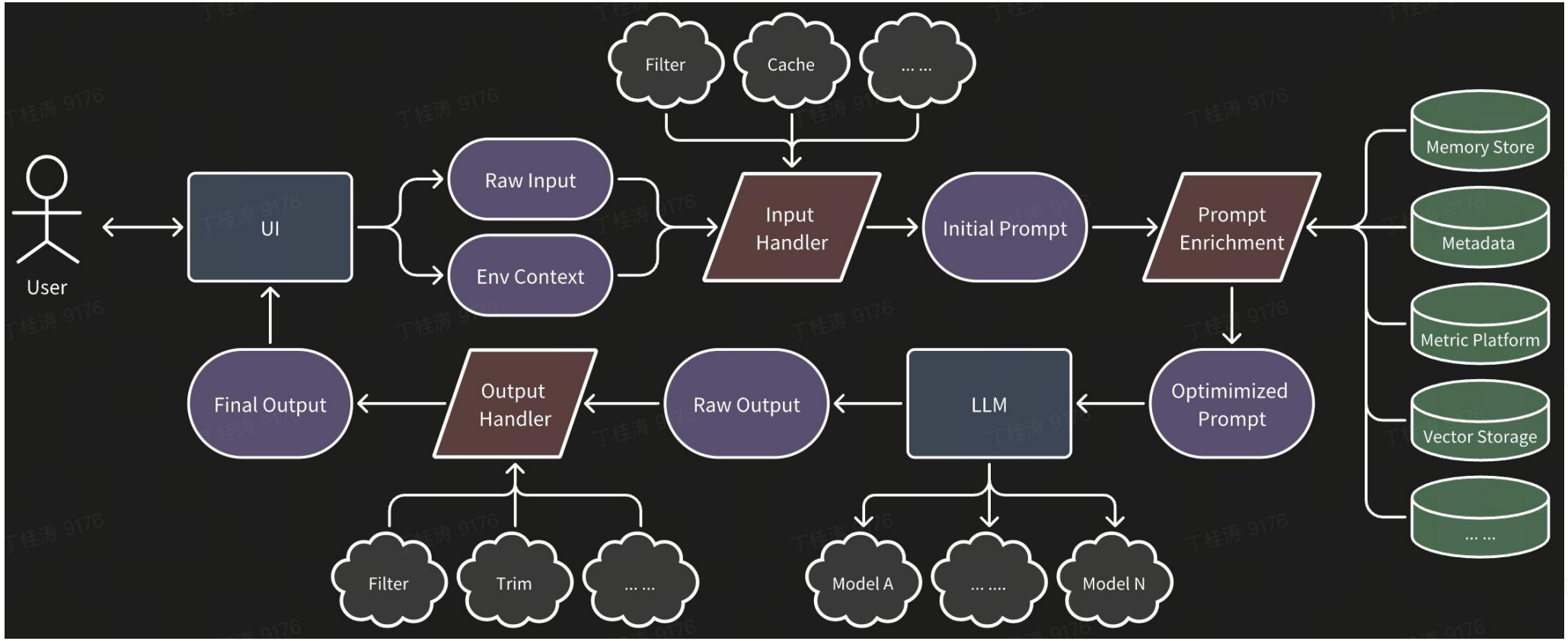
•开发助手 - 体验

•编辑器内交互，桌面级 IDE 体验

•关键链路延迟 < 秒级

•通过 A/B 实验优化模型策略




Filter
Cache
... ...
User
UI
Raw Input
Env Context
Input Handler
Initial Prompt
Prompt Enrichment
Memory Store
Metadata
Metric Platform
Vector Storage
... ...
Final Output
Output Handler
Raw Output
LLM
Optimimized Prompt
Filter
Trim
... ...
Model A
... .
Model N

•开发助手 - 技术架构

未来规划

- **数据生产与消费全流程建设**
  - 开发 – 测试 – 发布 – 运维
  - 找数 – 取数 – 分析
  - 业务信息沉淀

- **效果优化**
  - 对话理解 – 意图、关键词提取
  - 召回准确率 – 语义、相似性、排序策略
  - 大模型总结能力、避免“幻觉”、FineTune

- **内部 -> 对外**
  - 欢迎试用 DataLeap AI 助手 & 大数据研发套件

联系我们

火山引擎DataLeap找数&研发助手
咨询及试用申请

获取更多技术干货、活动信息
进入官方交流群

THANKS

软件正在重新定义世界

Software Is Redefining The World