

# 金融领域大模型数据集管理 与应用

演讲人：史鑫鑫

蚂蚁集团 / 大安全 数据研发专家

**AiCon**  
全球人工智能开发与应用大会

# 目录

01 引言与背景

02 智能数据管理

03 CoT思维链数据工厂

04 多模态数据管理与应用

05 业务场景应用案例分享

06 总结与展望

📍 北京

**QCon**

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

**AiCon**

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

**QCon**

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

**AiCon**

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

**AiCon**

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

**AiCon**

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOPs
- 具身智能

# 极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



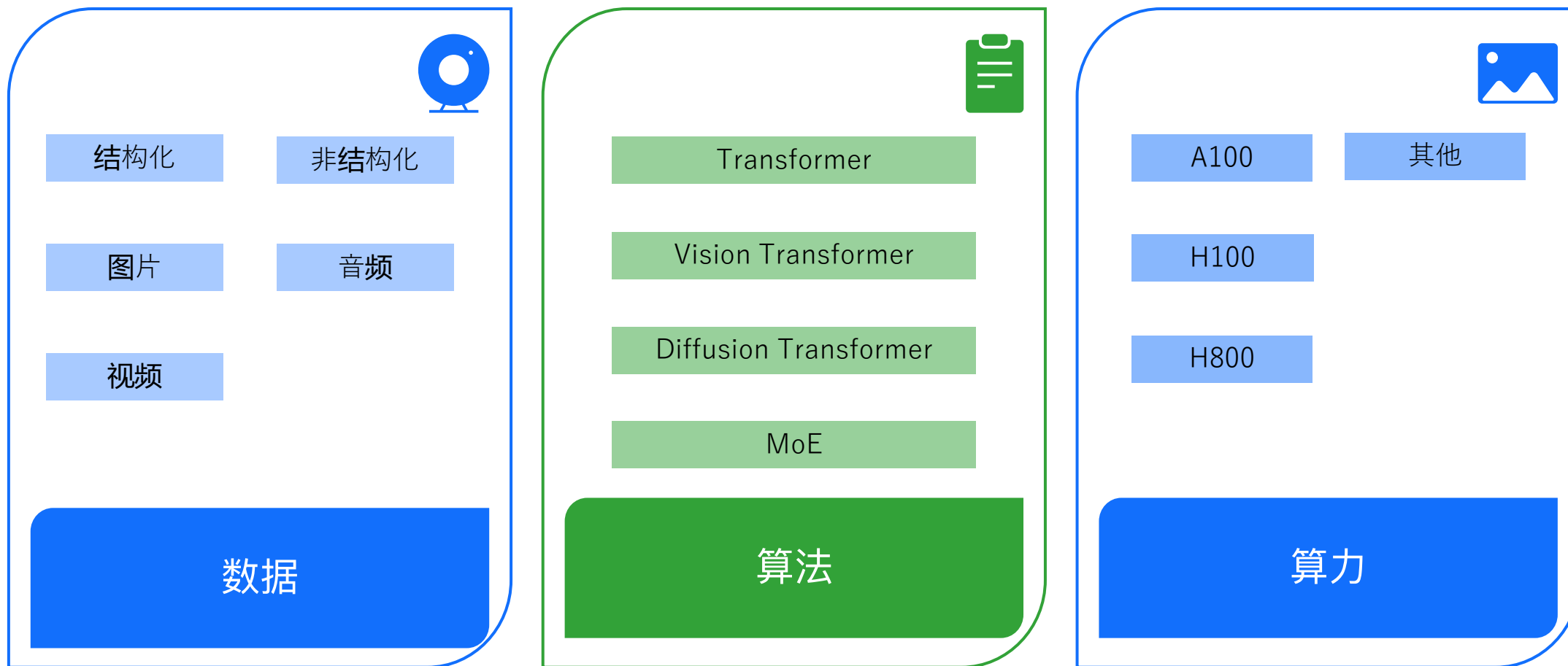
参会咨询



查看会议

# 01 引言与背景

# 人工智能的三架马车



# 小模型到大模型的变迁

小模型

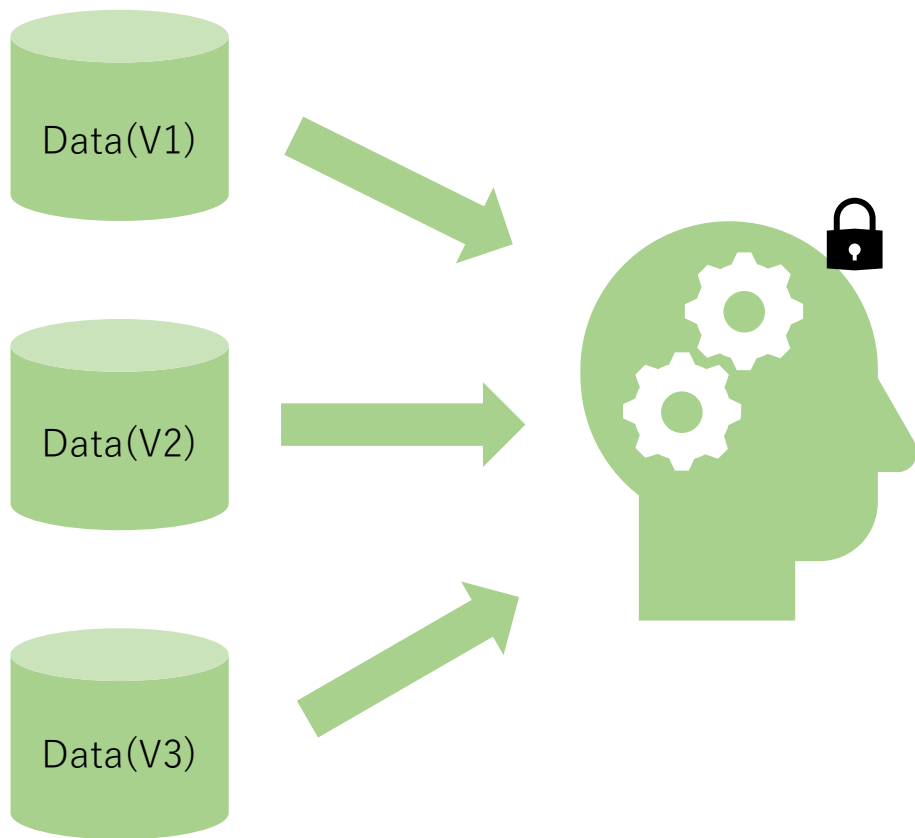


大模型



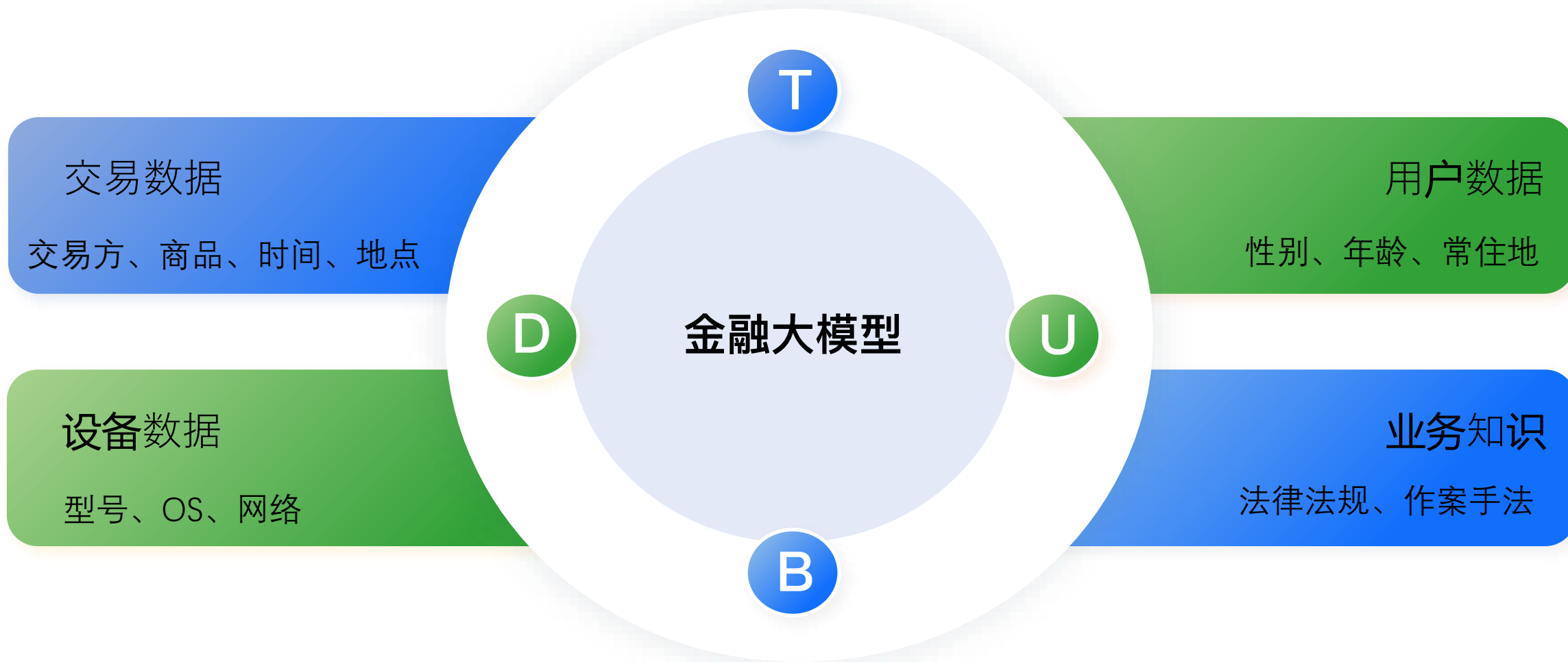


# Data-Centric AI



DCAI (Data-Centric AI) 是一种大模型研发范式，其以高质量数据集为核心迭代大模型的性能。

# 金融大模型数据





# 业务痛点

- 业务复杂、数据分散在不同的业务场景、数据看不全、看不清
- 数据存在孤岛、重复使用困难
- 数据存储和计算与大模型训练平台之间链路长、应用效率低
- 在重点业务场景，数据精细化运营困难

# 02 智能数据集管理

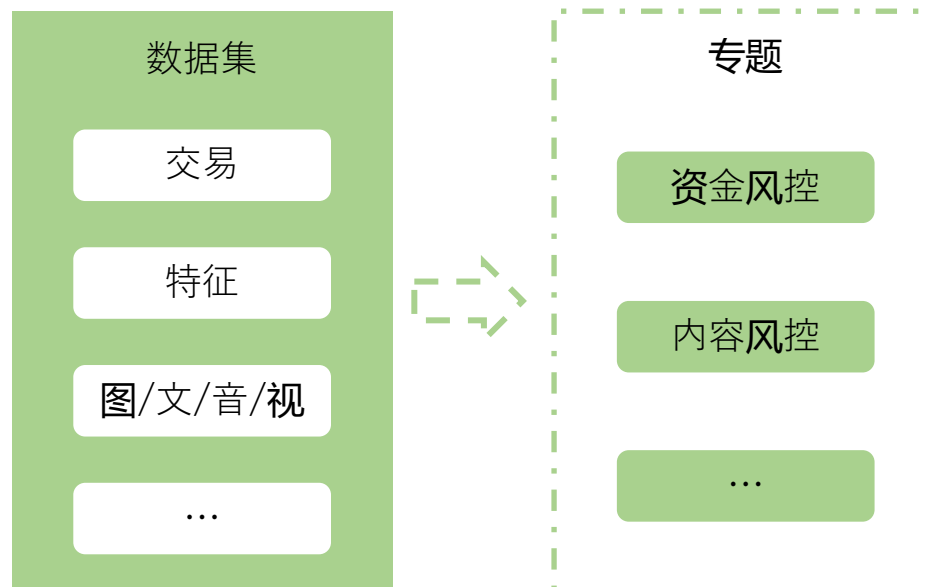
# 关键问题

- 看清、看全**业务**数据
- 打破数据孤岛，加速数据流通
- 提升数据**应用**效率
- **标签**化管理的**精细化**数据运营

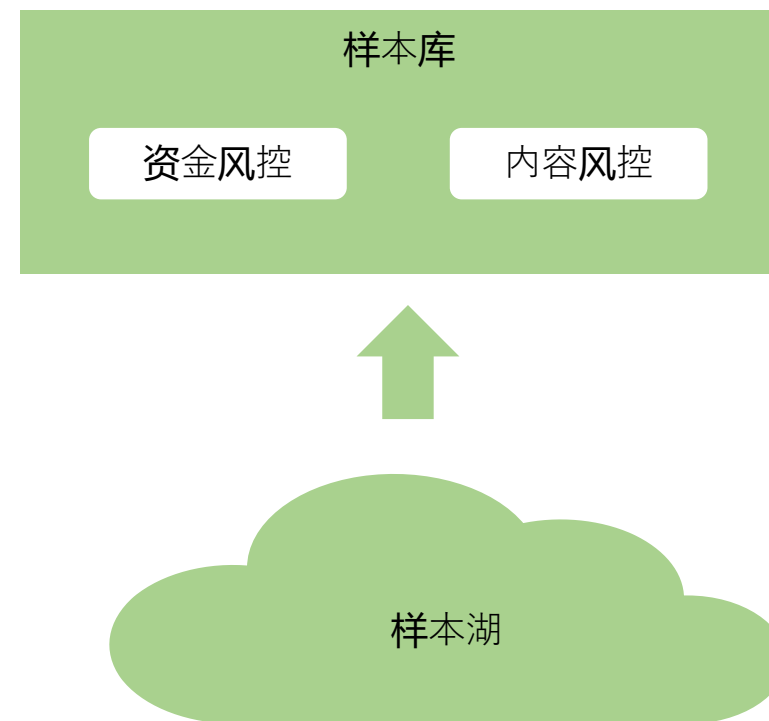
- ✓ 以数据集、数据**专题**的数据管理能力
- ✓ **统一**数据管理平台
- ✓ 打通数据**应用**上下游**关键**链路，让数据**应用**走上“高速路”
- ✓ 重点**业务**场景数据，**精细化**打标，实现数据**标签**化运营

# 数据管理体系

点



面



# 数据集

## ◆ 看清安全域数据集/样本库/知识库全貌



① 核心板块：列表页，详情页（简介、详情、下载）；

② 功能列表：数据集注册，编辑，标签级连，预览，权限管理，高敏内容处理，核心数据更新（浏览、下载等），样本量级；

### 数据类型

文本(35) 图片(160) 音频(92)  
视频(12) 书籍(1) 结构化(11) COT(3)

### 来源类型

采购(14) 标注(32)  
业务(50) 生成式(8)

### 业务类型

加油站(13) 数天蜜(14)  
流媒体库(96) 万象(45)  
安全智造(0)

### 安全等级

公开(123) 私有(42) 高敏(76)

### 数据规模 (rows)

<10K(2) 10K-1M(0) 1M-10M(0)  
>10M(9)

### 应用方式

SDK(0) NAS(4) ODPS(7) FILE(0)

	<div>文本 图片 结构化 COT 采集 自采 标注 公开 审核智能化 &lt;10K NAS</div> <div>入库时间:2025-05-21 16:42:02 安全等级:公开 数据品质:★★★★</div>	0 0 0
	<div>直播评测原始数据集</div> <div>文本 图片 音频 视频 业务 私有 直播智能化 数据源 &gt;10M ODPS</div> <div>入库时间:2025-05-15 19:49:09 安全等级:私有 数据品质:★★★</div>	79 4 0
	<div>直播训练原始数据集</div> <div>文本 图片 音频 视频 业务 私有 直播智能化 数据源 &gt;10M ODPS</div> <div>入库时间:2025-05-15 19:43:21 安全等级:私有 数据品质:★★★</div>	60 8 0
	<div>直播全量原始数据集</div> <div>文本 图片 音频 视频 业务 私有 直播智能化 数据源 &gt;10M ODPS</div> <div>入库时间:2025-05-15 19:37:17 安全等级:私有 数据品质:★★★</div>	65 5 0
	<div>MegaPairs-通用多模态检索的海量数据合成数据集</div> <div>图片 公开 &gt;10M NAS</div> <div>入库时间:2025-05-13 16:08:10 安全等级:公开 数据品质:★★★★</div>	45 1 0

# 样本库

## ◆ 聚焦核心数据集应用：秒级样本精细化标签体系分析与策略运营，一键对接各业务系统



- ① 能力建设：标签体系自定义；样本列表模板化；列表样式模板化；功能标准化（增/删/查/改/接口/上传）；内嵌页模板化；数据回流标准化；
- ② 功能列表：样本上传，编辑，标签管理，高敏内容处理，系统对接，帮助文档。

基础筛选

Basic Labels

数据类型

全部(34441) 文本(34441) 图片(0) 音频(0) 视频(0)

来源类型

业务(0) 生成(34441) 域外(0)

安全等级

公开(34441) 私有(0) 高敏(0)

数据质量

S(9144) A(25297) B(0) 未知(0)

Special Labels

样本批次: 全部(34441) ... 展开

样本类别: 全部(34441) 训练集(21719) 自测集(12722) ... 展开

黑白标签: 全部(34441) 通讯(6441)

风险类型: 全部(28983)

场景: 全部(28983)

账户类型: 全部(28983) 个人(28983) 企业(0)

商户资质: 全部(24800)

经营属性: 全部(24800)

基础模型: 全部(12722)

训练方法: 全部(12722) LORA(2302) SFT(10420)

事件属性: 全部(64130) ... 展开

特征分类: 全部(57924)

是否含特征回溯: 全部(28983)

已选中 0 条数据 全选 反选 下载 应用 时间排序

Instruction:

Input: # 角色 你现在是一名经验丰富的欺诈风控分析专家。

Reasoning:

Answer: 好的，我现在需要根据提供的输入信息，

# ■ 标签化数据管理

## 业务标签

主体

- 性别
- 年龄
- 省份

设备

- 型号
- 系统版本
- 可信

交易

- 金额
- 时间
- 双方关系

## 基础标签

文本

- Token数
- 语种

图片

- 分辨率
- 大小

视频

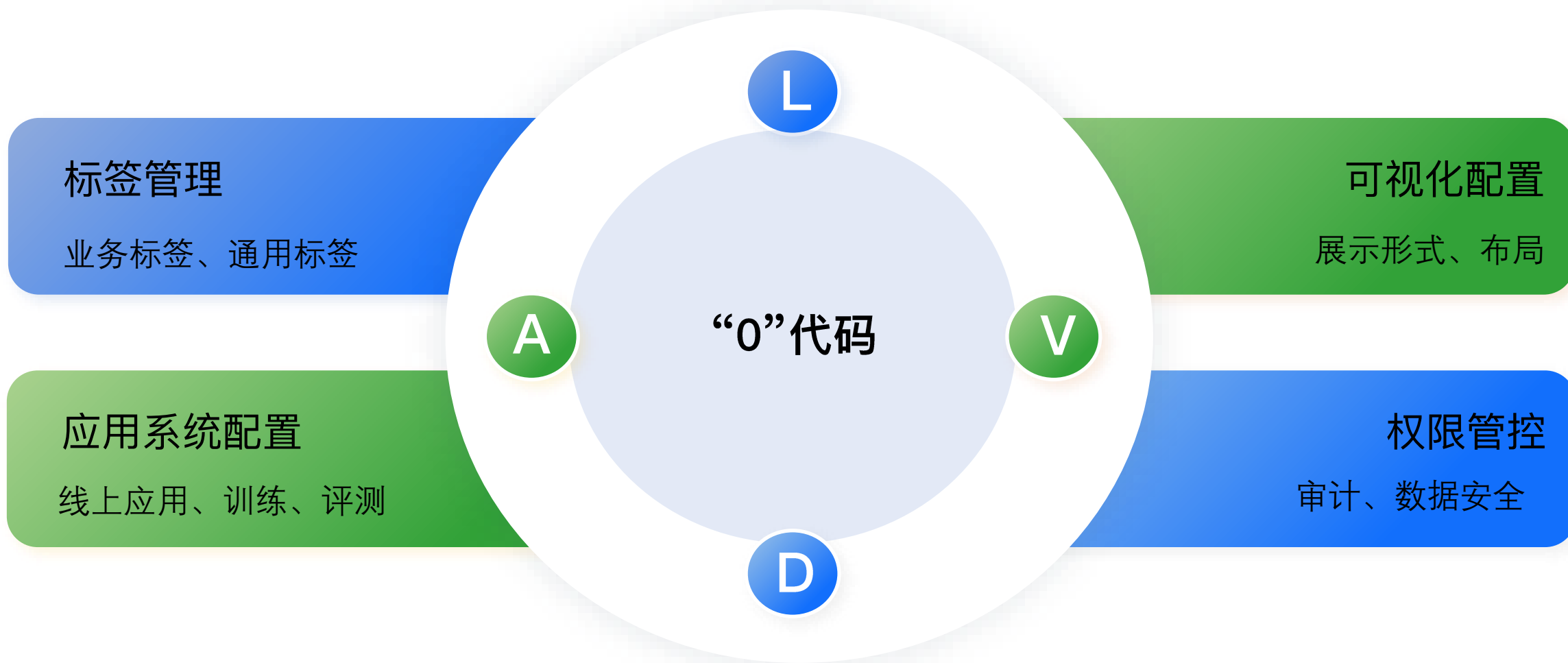
- 分辨率
- 时长

音频

- 码率
- 比特率

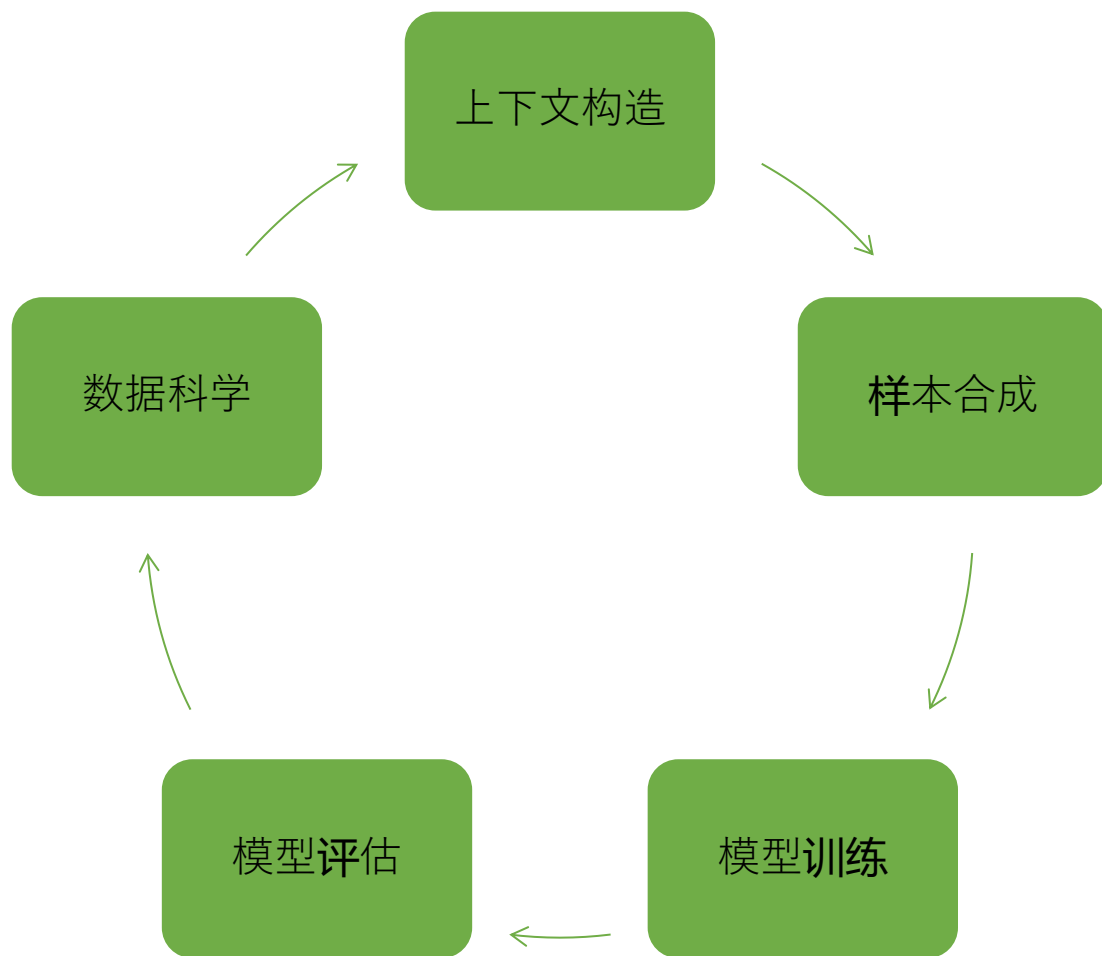


# “0”代码数据管理产品化



# 03 CoT思维链数据工厂

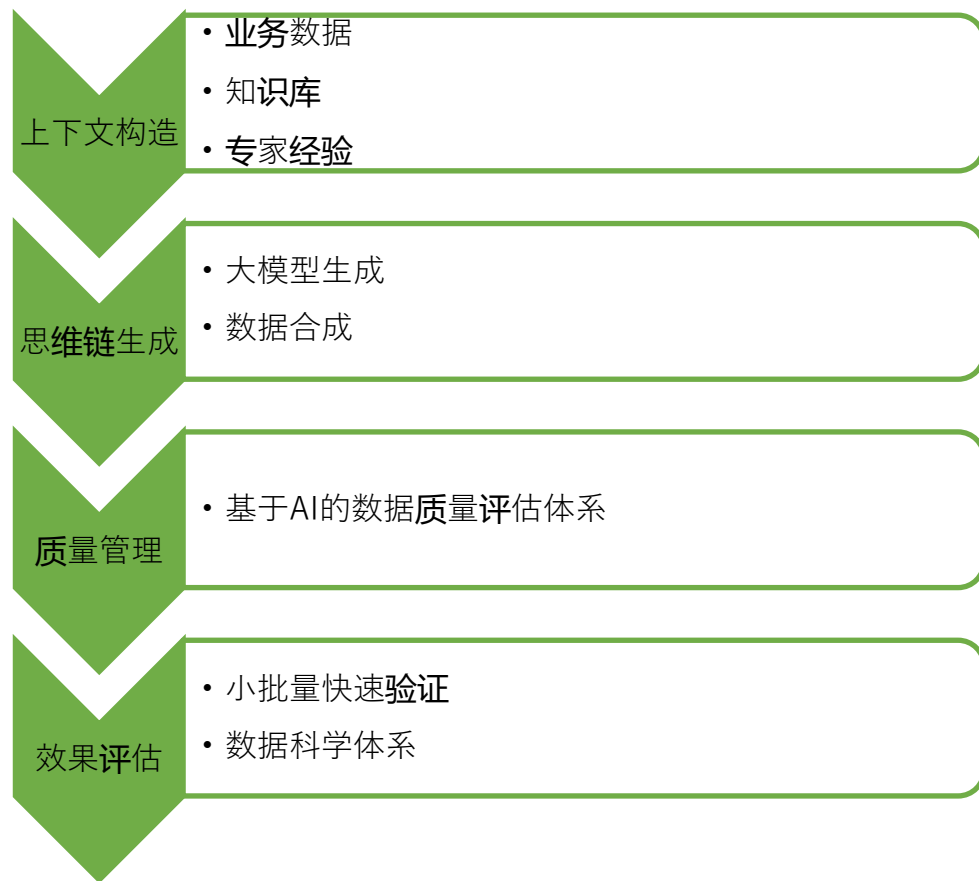
# 设计理念



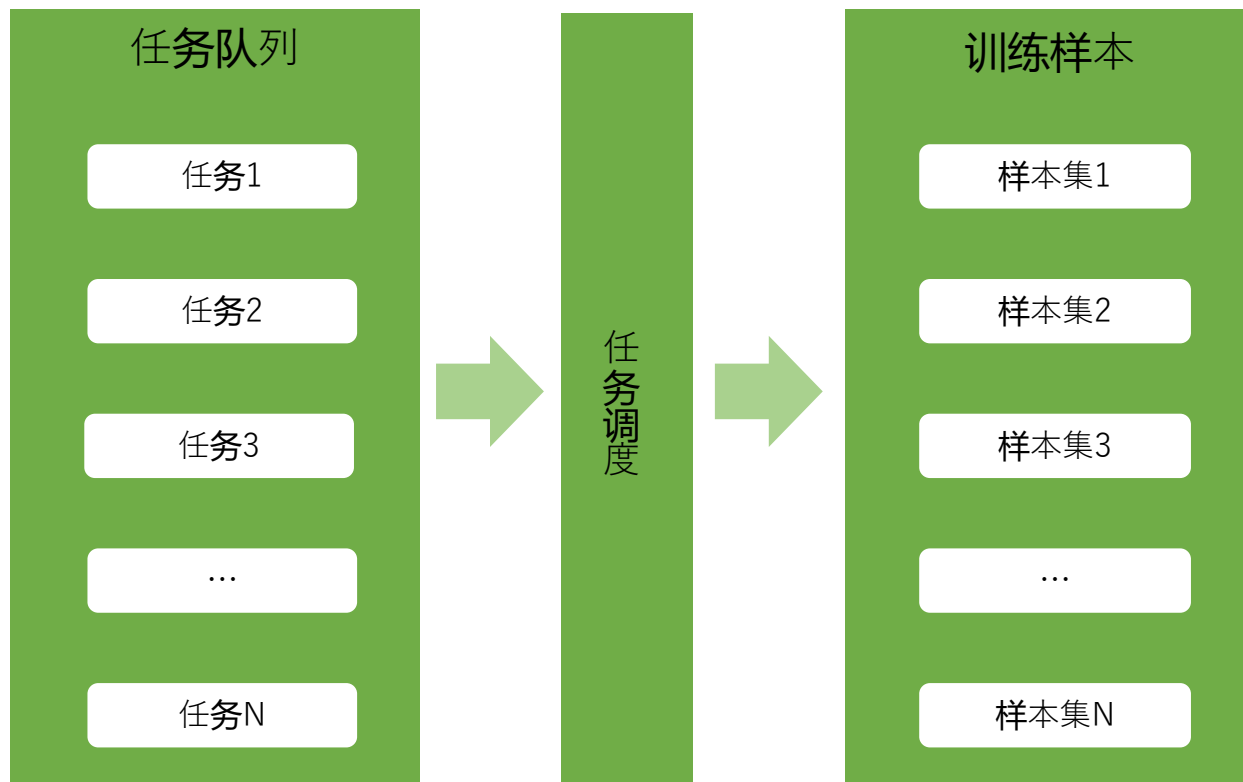
“飞轮”驱动的数据研发，加速大模型应用落地

# CoT构建车间

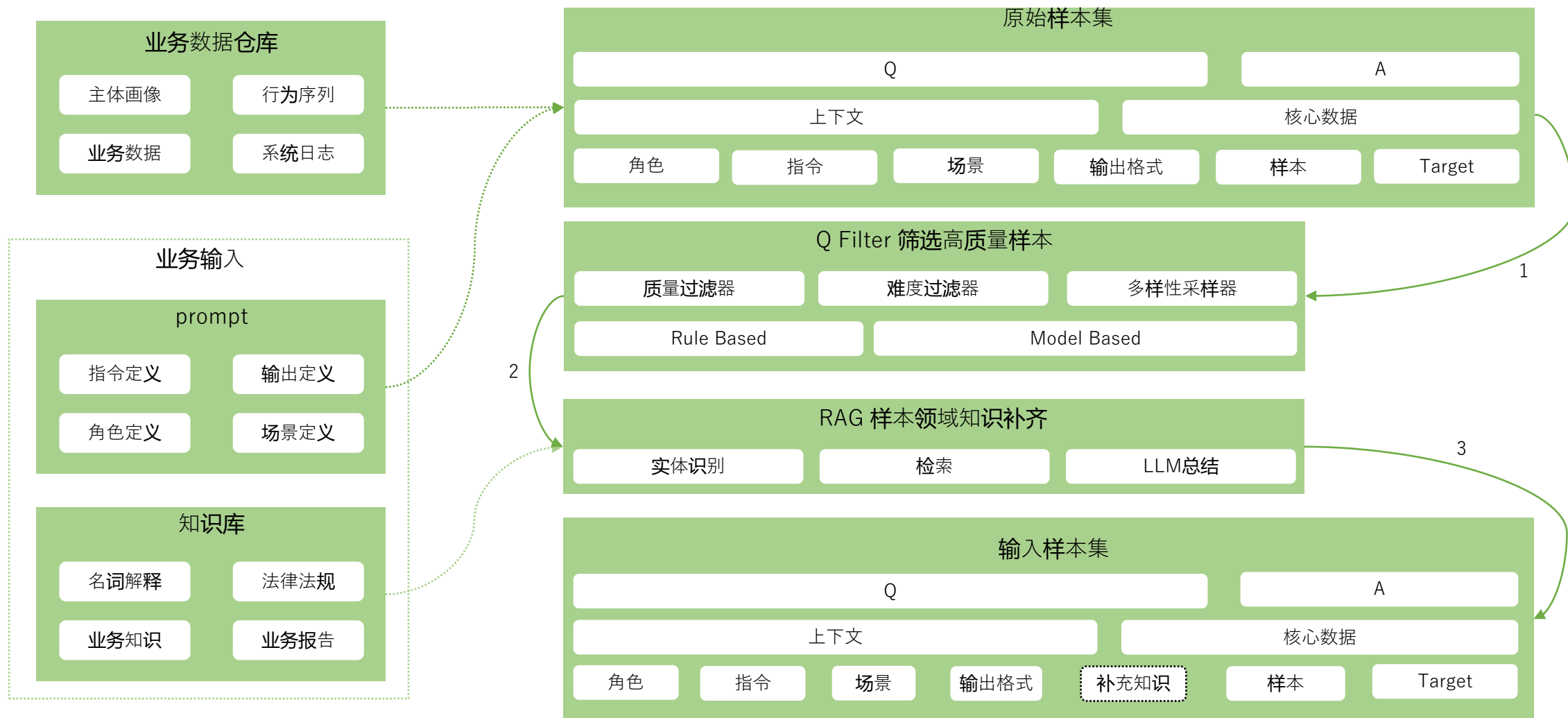
## 流程编排



## 批量生产



# 长上下文构造

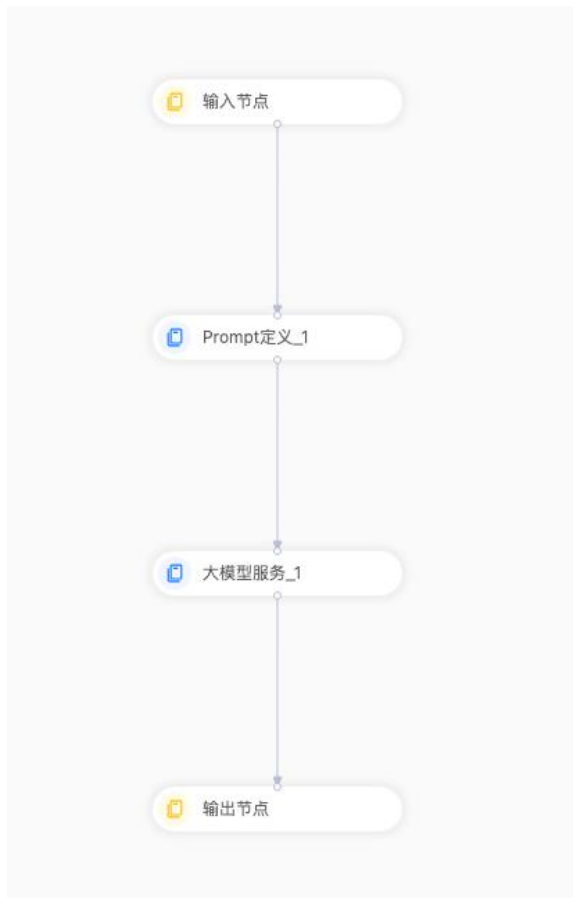


# CoT样本质量管理



# CoT样本构建

简单模版



复杂模版



- “手搓”CoT进化到自动化、流程化
- 高质量CoT样本构建进入日产十万+, 解决业务CoT样本慌的问题
- 样本产出到模型训练进入流程化



# 04 多模态数据管理与应用

# ■ 多模态数据管理-存储系统

缓存

- 表格
- JSON
- CSV
- Excel
- 富文本

结构化/半结构化

- 长图
- 横图
- JPG、PNG、WebP

图片

- 长视频
- 短视频
- MP4、MOV、AVI、WebM

视频

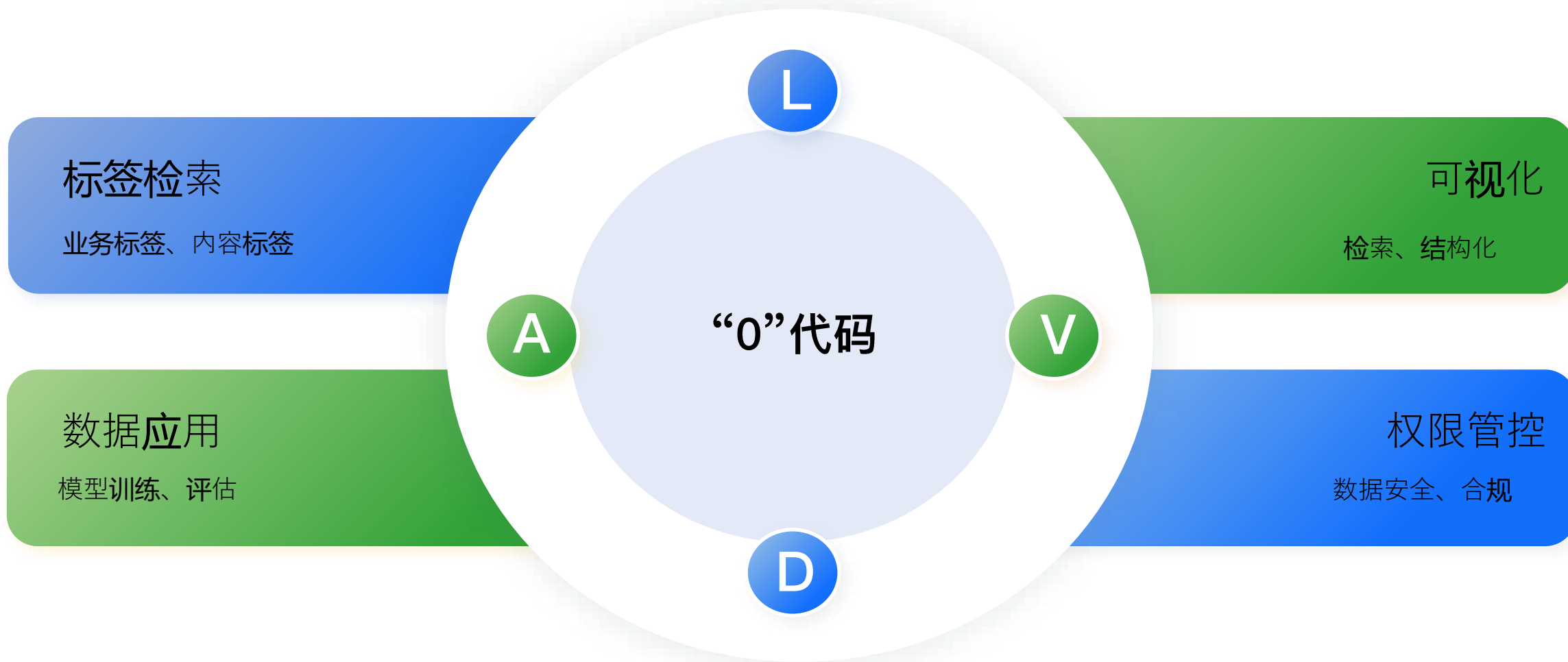
- 长音频、短音频
- 纯音乐
- 单人、多人

音频

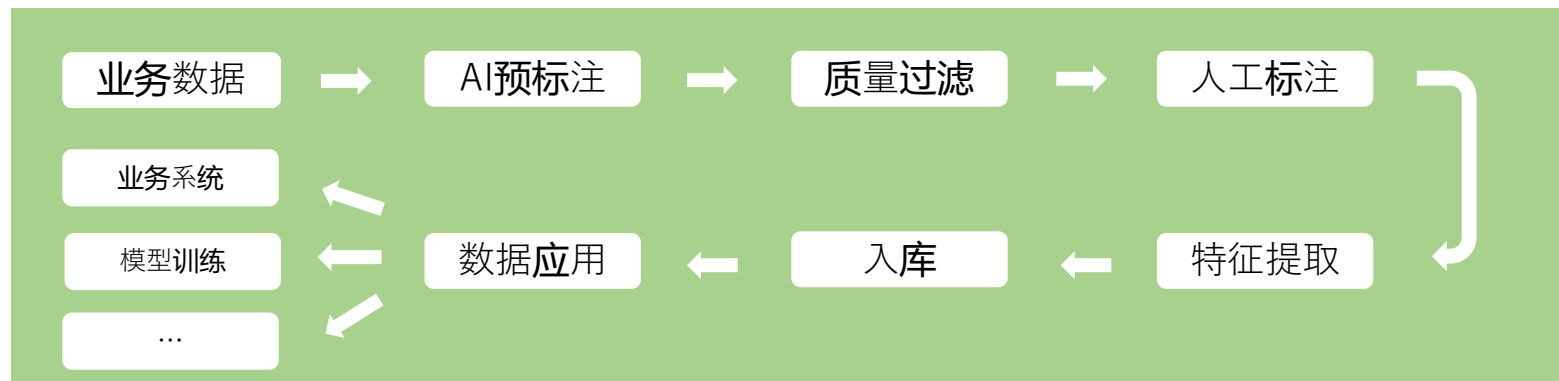
数据湖

OSS

# ■ 多模态数据管理-可视化



# 多模态数据管理-可视化



筛选条件

更新时间: 请输入开始日... → 请输入结束日... 上传样本

内容类型: 全部 文本 图片 音频 视频 直播

风险模式: 全部 纯图片 图片OCR 纯音频 音频ASR

来源渠道: 全部 AIGC(37,337)

防控难度: 全部 简单

风险类型: 违法违禁 公序良俗 业务体验 直播质量 音频质量(1)

全部 违反广告法(58,016)

样本列表

共 108,182 条样本, 已选中0条样本

文本 图片 音频 视频 直播

全选 反选 共 108,182 条样本, 已选中0条样本

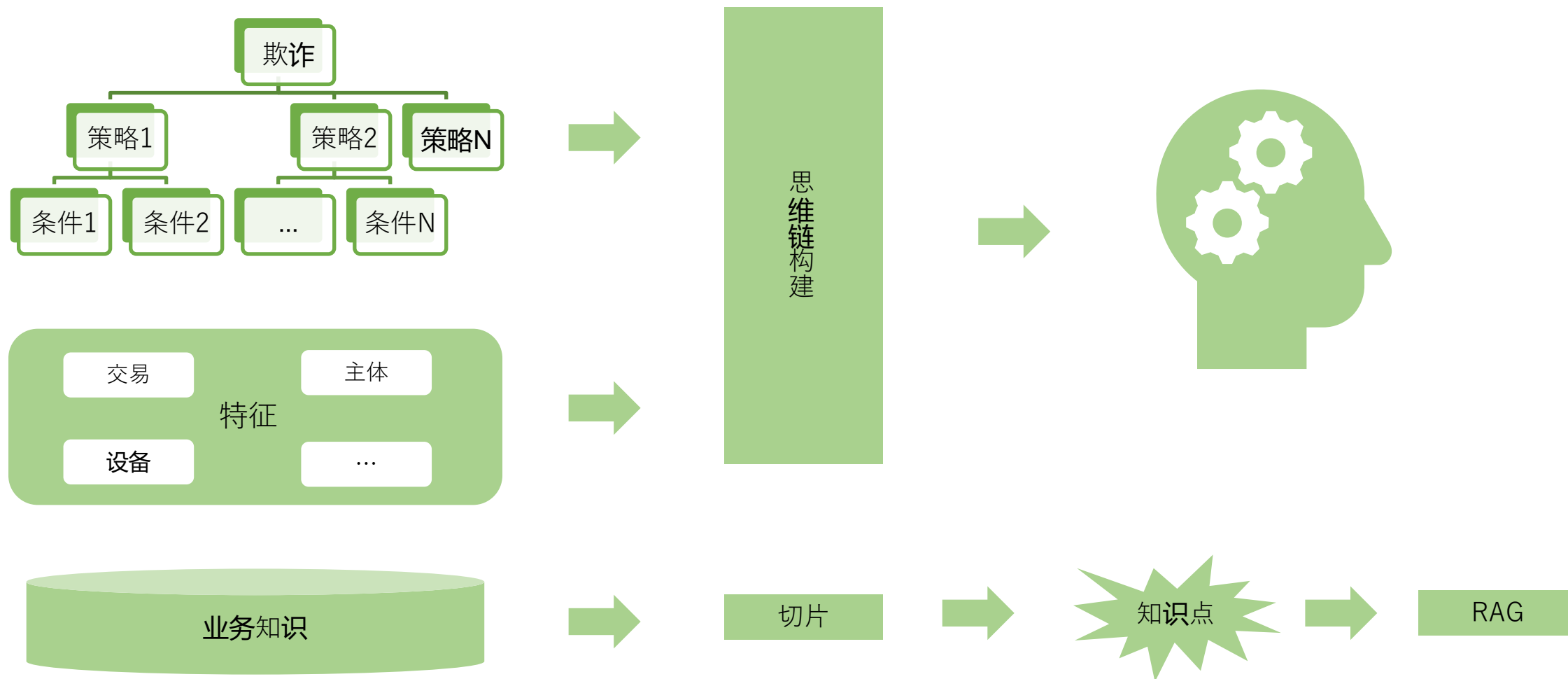
风险模式: 纯图片 来源渠道: 更新: 2025-04-07 18:30 查看关键词 查看音频 查看ASR 监管指令数据集

风险模式: 纯图片 来源渠道: 更新: 2025-04-04 10:24 查看关键词 查看音频 查看ASR 监管指令数据集

风险模式: 纯图片 来源渠道: 更新: 2025-04-03 17:54 查看关键词 查看音频 查看ASR 监管指令数据集

# 05 业务场景案例分享

# ■ 资金风控场景大模型数据管理



# 资金风控场景CoT构造案例

- 风险管控策略语义化，将专家经验及业务知识转化为文本内容，制定模型思考框架
- 业务特征语义化，数字化特征语义化，转化为模型可理解的内容，为模型提供关键判断依据

```
# 角色
你是一名经验丰富的欺诈风控分析专家。

# 风险判断逻辑
根据输入的客户基本信息、交易行为特征、设备特征并依据知识库中的人机交互规则，判断是否存在风险，并给出判定依据：
1、客户可信：
2、关系可信：
3、风险判断：
4、综合判断：

# 可信判断增强
- 根据输入的客户基本信息，判断客户是否为新客户。
- 根据输入的交易行为特征，判断交易是否异常。
- 根据输入的设备特征，判断设备是否安全。

# 风险判断增强
- 根据输入的客户基本信息，判断客户是否属于高风险人群。
- 根据输入的交易行为特征，判断交易是否属于高风险交易。
- 根据输入的设备特征，判断设备是否属于高风险设备。

## 风险分析约束
- 风险分析必须基于输入的特征进行。
- 风险分析必须基于知识库中的规则进行。

# 输出要求
## 输出的格式
输出格式需要满足以下要求：
<think>
</think>
<answer>
<hasRisk>{是否有风险}</hasRisk>
<riskInfo>
  <风险对象>{风险涉及对象}</风险对象>
  <风险类型>{风险类型}</风险类型>
  <风险等级>{风险严重程度}</风险等级>
</riskInfo>
```

```
Answer
好的，我需要根据提供的输入信息，首先，我需要仔细阅读输入的特征信息，然后按照步骤逐一分析各个部分。

首先，我需要分析客户基本信息。客户信息显示为：客户ID为123456，客户名为张三，客户年龄为35岁，客户性别为男。

接下来是商户信息。商户信息显示为：商户ID为789012，商户名为李四，商户类型为餐饮。

双方关系核销方面，根据输入的信息，双方关系核销结果为：双方关系核销通过。

经营健康度方面，根据输入的信息，经营健康度评估结果为：经营健康度良好。

主动方面画像分析中，根据输入的信息，主动方面画像分析结果为：主动方面画像分析通过。

被动方面画像方面，根据输入的信息，被动方面画像分析结果为：被动方面画像分析通过。

交易画像中，根据输入的信息，交易画像分析结果为：交易画像分析通过。

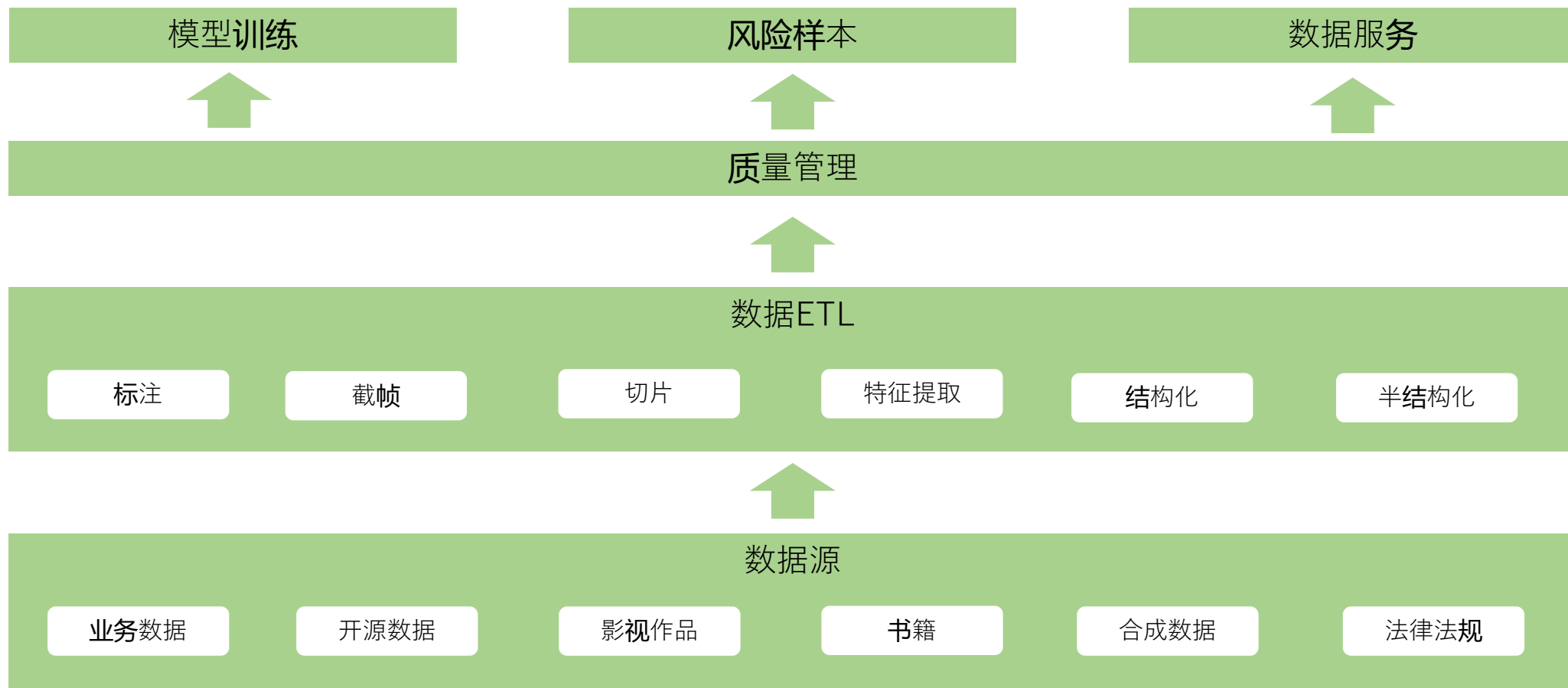
综合判断：根据以上分析结果，综合判断该客户是否存在风险。

</think>

<answer>
<hasRisk>N</hasRisk>
<infoCodeList>
  <infoCode>1</infoCode>
  <infoCode>2</infoCode>
  <infoCode>3</infoCode>
  <infoCode>4</infoCode>
  <infoCode>无高风险特征</infoCode>
</infoCodeList>
<merchantProfile>
  <运营模式>{运营模式}</运营模式>
  <所属行业>{所属行业}</所属行业>
  <商品类型>{商品类型}</商品类型>
</merchantProfile>
</answer>
```



# 内容风控场景多模态数据管理



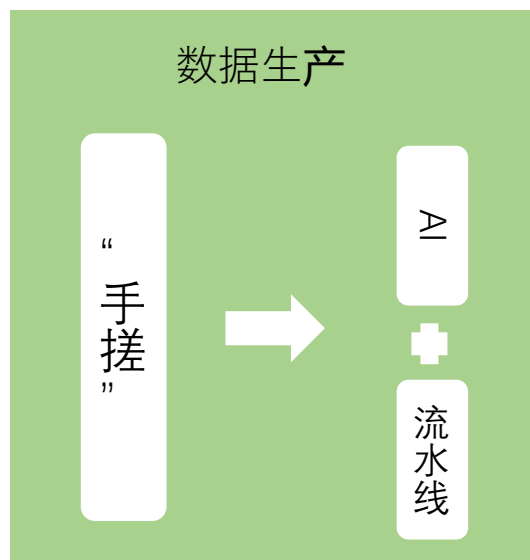
# 06 总结与展望

# 总结与展望

- 数据管理：从「**杂乱无章**」到「**整齐划一**」的提升
- 数据生产革命：从「**人工标注**」到「**AI增强闭环**」，从「**纯手工**」到「**工业化**」
- 数据**质量**评估体系：从「**准确**」到「**高质量**」
- 数据**应用**效率：从「**天**」到「**秒**」的提升

# 希望您能收获

- 大模型数据进行数据集、专题、样本库分层数据管理理念
- 标签化数据管理方式为业务带来数据应用带来的效率提升
- AI+的方式提升大模型数据生产效率



📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOPs
- 具身智能

# 极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

# THANKS

探索 AI 应用边界

Explore the limits of AI applications

## AiCon

全球人工智能开发与应用大会