InfoQ 极客传媒

人机协同构建10x组织效能和内容安全新范式

演讲人：王东旭

快手 / 磁力引擎风控技术负责人

AiCon

全球人工智能开发与应用大会

AiCon
全球人工智能开发与应用大会

目录

极客邦科技 2026 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

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

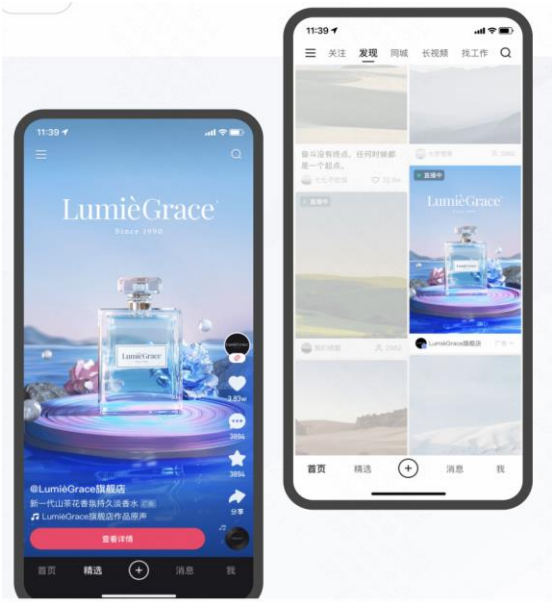
InfoQ 极客传媒

01

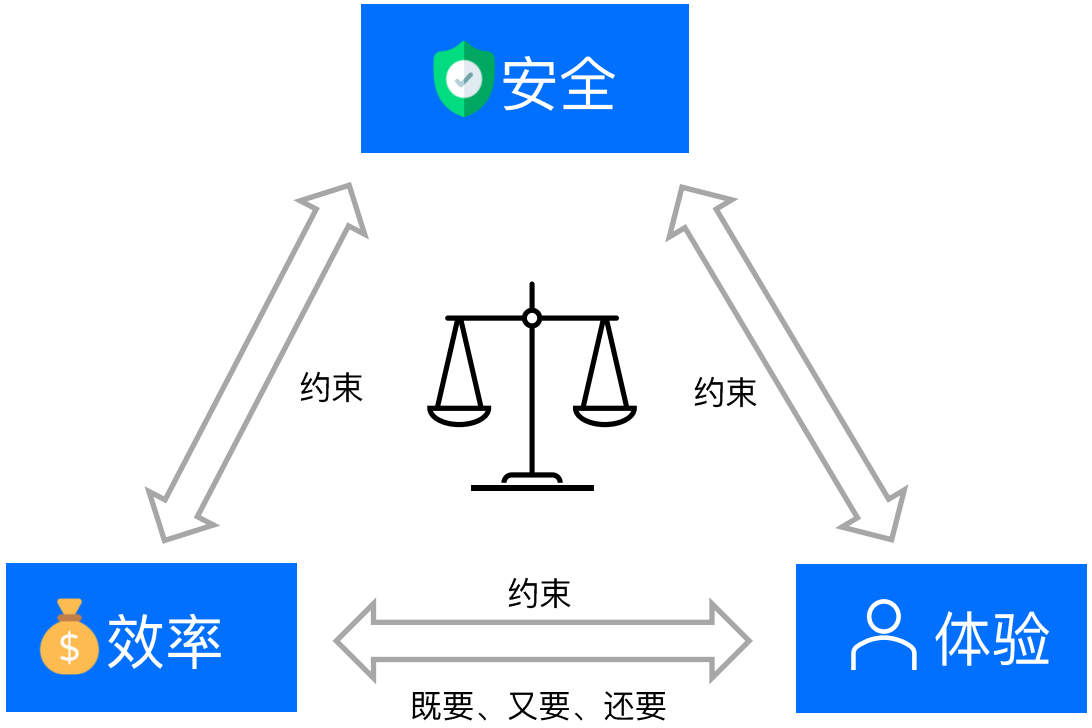
LLM带来的生产力变革

内容安全的新挑战

内容安全的业务问题和挑战

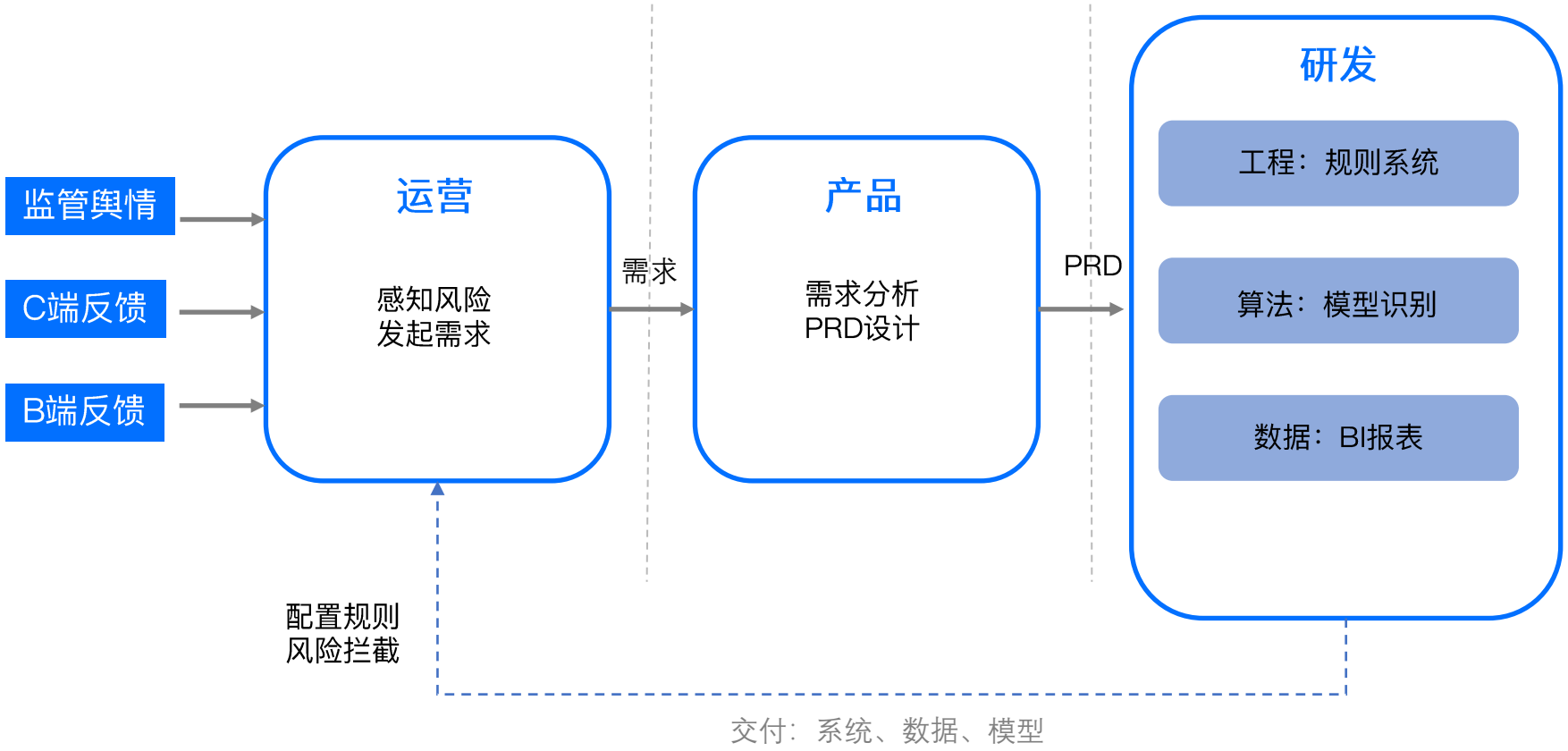


短视频广告内容审核业务形态



安全、体验、效率的“不可能三角”

内容安全的传统“固态组织”



固态组织　分工明确　边界清晰

过去10年，发生了什么变化？

人工经验规则驱动 | 机器学习开启智能审核 | DeepLearning时代 | ChatGPT引爆生产力革命 | 内容安全生产关系变革

1 2014的内容安全

- 基于敏感词库的文本过滤系统，覆盖政治、色情、暴力等类别
- 建立数千人规模的内容审核团队，实行三班倒7×24小时监控
- 用户举报–人工复审的双重机制，形成社区共治雏形
- 初代图像识别技术应用于色情、暴力图片的批量筛查

2015

2 2017的内容安全

- 卷积神经网络实现图像多分类识别，准确率>90%
- LSTM、BERT等模型用于文本语义分析，识别隐晦违规表达

2017

- 智能分发系统将高危内容优先推人，审核效率提升
- 音视频指纹技术实现已知违规内容的秒级拦截

3 2020的内容安全

- 图文音视频统一理解框架，跨模态语义关联分析

2020

- OCR+NLP组合识别图片内嵌文字，封堵"图片藏字"漏洞
- 实时流媒体审核技术支持Live直播场景毫秒级响应
- 深度学习模型成为内容安全审核的标准模式

4 2022的内容安全

2022

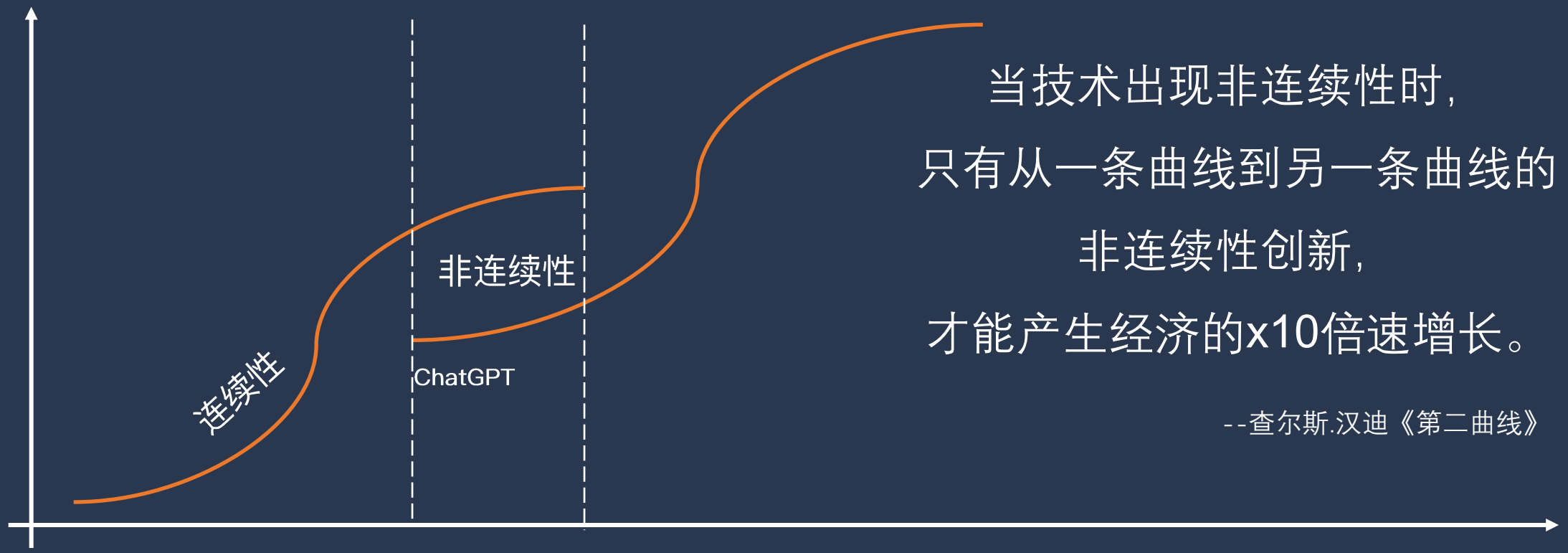
- ChatGPT带来文本内容的井喷式爆发，审核压力骤增
- 大模型幻觉、大模型内生安全、安全围栏
- 低成本Prompt对抗，内容创造成本大幅降低

5 2025的内容安全

2025

- AIGC多模态内容审核的效果和效率平衡
- 从单纯的内容理解，加速迈向内容理解+生成统一范式
- AI带来的内容安全团队各角色职能重塑

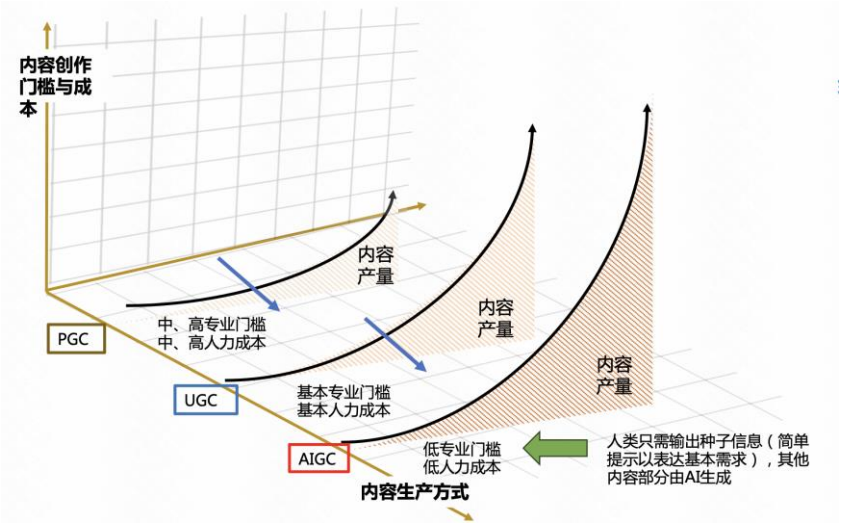
技术出现“不连续性”

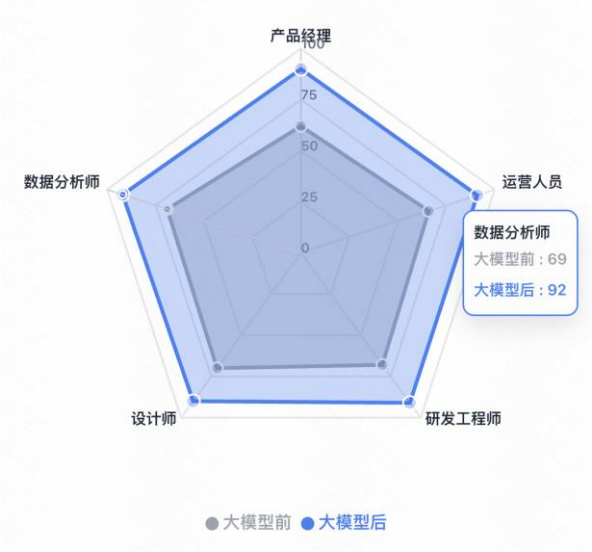


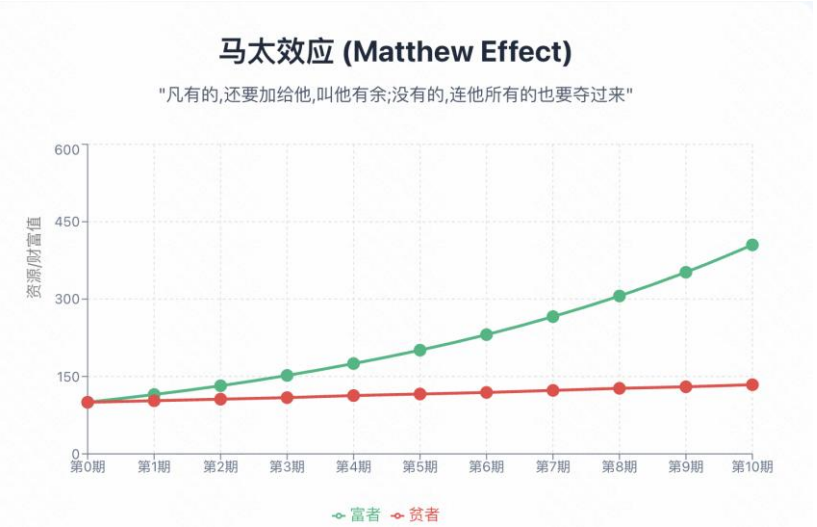
当技术出现非连续性时，
只有从一条曲线到另一条曲线的
非连续性创新，
才能产生经济的x10倍速增长。

--查尔斯.汉迪《第二曲线》

大模型技术发展带来的挑战







内容产量井喷爆发	技术门槛降低	马太效应凸显
审核成本飙升	能力差距缩小	强者愈强，弱者愈弱
低门槛Prompt对抗	模型平权	重复劳动被AI取代

这是最好的时代，

也是最坏的时代。

-- 狄更斯

02

AI驱动组织架构重构

AI驱动产运研角色持续向价值链上游升级

AI时代风控产运研生产关系变革

产品经理面临的挑战

未来的产品，长什么样?
不懂LLM，怎么设计出AI-Native的产品?

产品工程师：从PRD设计到“P2P”设计

Figma　Lovable　Bolt

新的产品设计模式（P2P）：prompt to product

Prompt驱动的产品原型设计

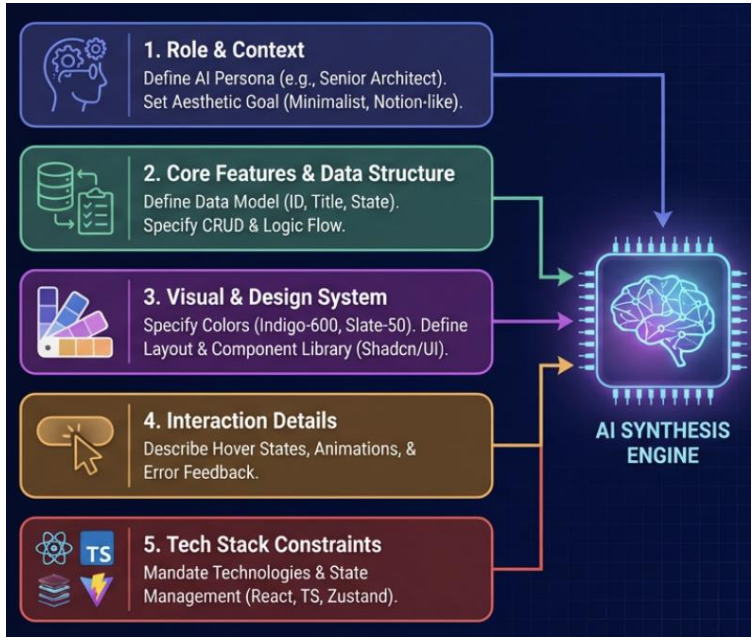
对话式产品原型设计

1. Prompt驱动的产品原型设计，直接和大模型对话，给出原型

2、产品原型达到可跳转、可执行的水准，告别“一句话PRD”

3、产品经理不再会被研发吐糟不懂代码，只能提需求

什么是好的产品原型设计Prompt?



Prompt → Product

AI 时代产品经理能力进阶指南

P 完美的 Prompt 结构

- **角色与上下文**
  你是谁？(资深架构师) 你在做什么？(极简To-Do)
- **核心数据结构**
  定义字段 (ID, Title, Status) 与 CRUD 逻辑。
- **视觉与设计系统**
  指定 Tailwind 颜色、Shadcn 组件库、布局方式。
- **交互细节**
  Hover 效果、空状态提示、动画过渡。

S PM 能力新要求

- **模块化思维**
  像搭积木一样拆解页面组件，而非画平面图。
- **技术理解力 (Tech Literacy)**
  懂布局原理、状态管理，能用“技术方言”对话。
- **逻辑闭环能力**
  消除歧义，像写法律文书一样写需求。
- **审美与调试**
  从“翻译官”升级为“设计总监”与“指挥官”。

策略运营面临的挑战

只会简单的规则运营，
我的工作会不会被LLM取代?

风险PE运营：从写规则到Prompt Engineer

角色变化

1. 不再是简单的审核规则表达式（左右变量、操作符）配置

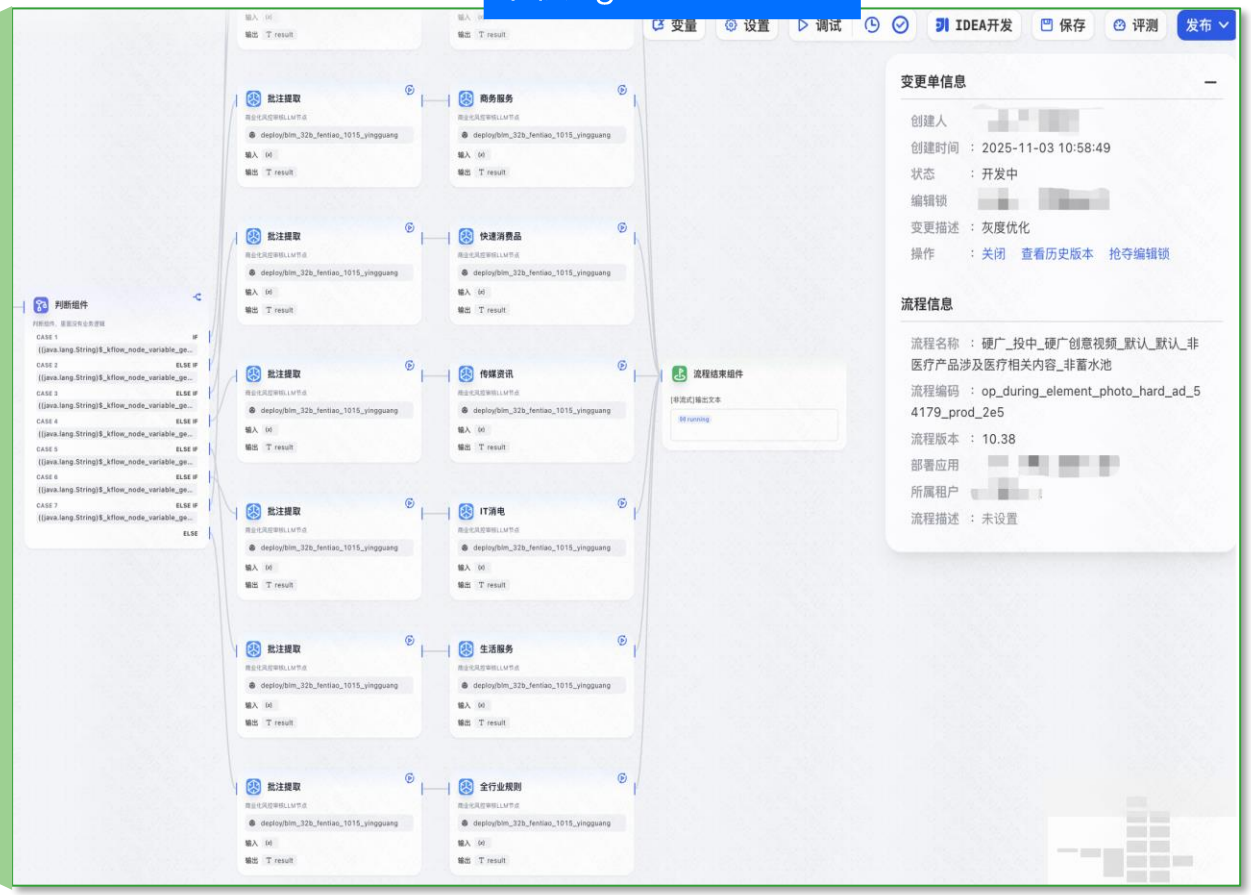
2、具备面向MLLM的Prompt编写、优化能力，擅长长CoT的推理链设计和拆解

3、RAG知识库的设计、迭代和维护，风险错题集沉淀并注入到多模态大模型

风险运营Prompt

#角色定义
你是一个短视频广告审核人员，负责……，对待审素材信息进行判断。你的任务是……的标准。
#审核流程
1.分析素材信息
–仔细查看抽帧图片中的视觉……
2. ……
#注意事项和限制条件
–必须完全依赖抽帧图……
–不得通过推测……
……
#核心判断点
首先明确人物为受益人才……
受益人定义：人物使用了产品……
#豁免场景
推广素材中仅提及成为自己的老板、……继续招募合作商，无需拒绝
#回复格式
<think>推理过程CoT</think>
<answer>违规/不违规</answer>
#待审素材信息
抽帧图地址集合:{frame}
ASR输入:{ASR}
……

审核Agent Workflow

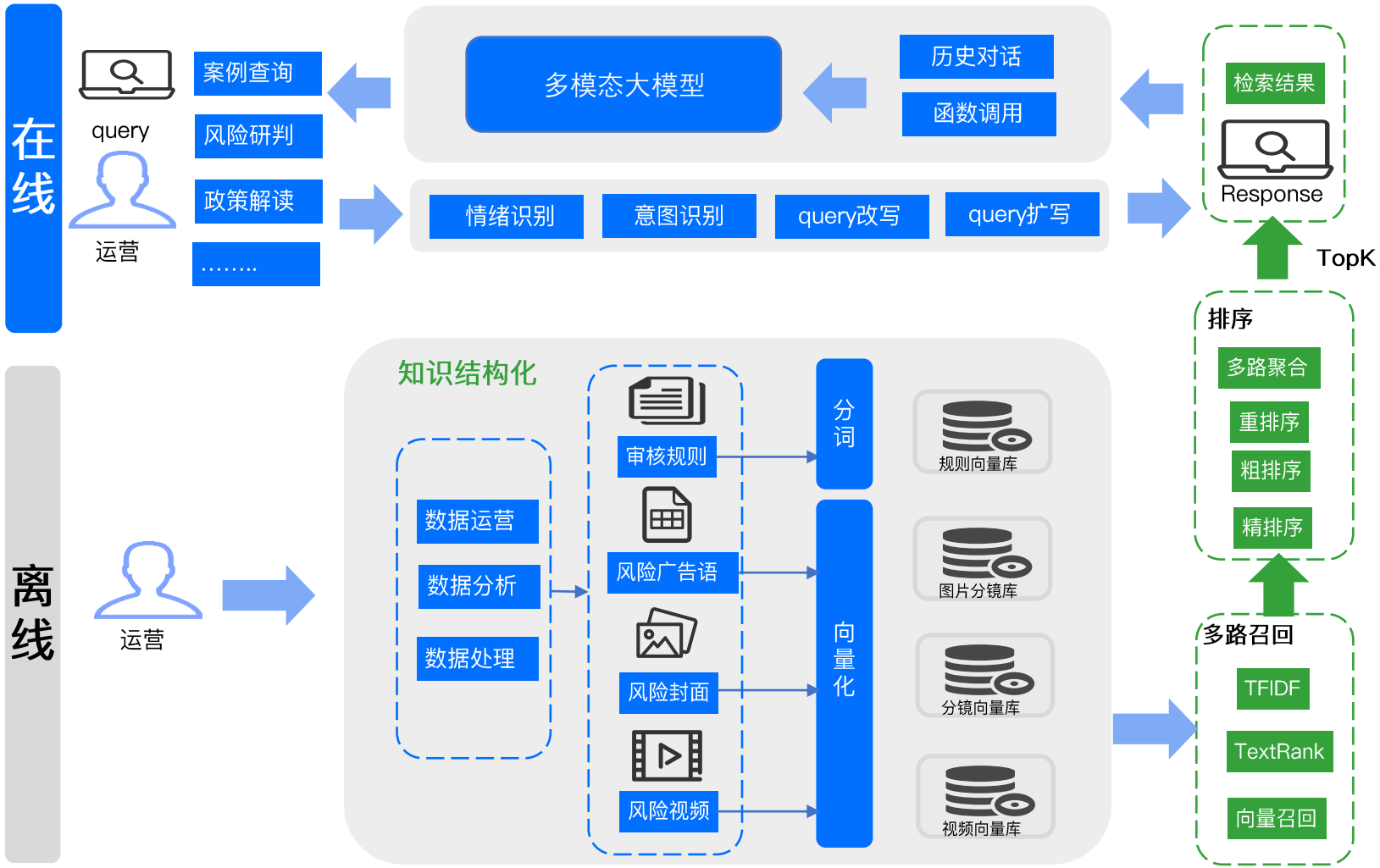


不但要会PE运营，还要RAG运营

知识库管理
模式变化

1、风险案例资产化，避免分散维护于各业务方的离线文档中，并支持将案例沉淀有效注入多模态大模型，提升知识复用能力。

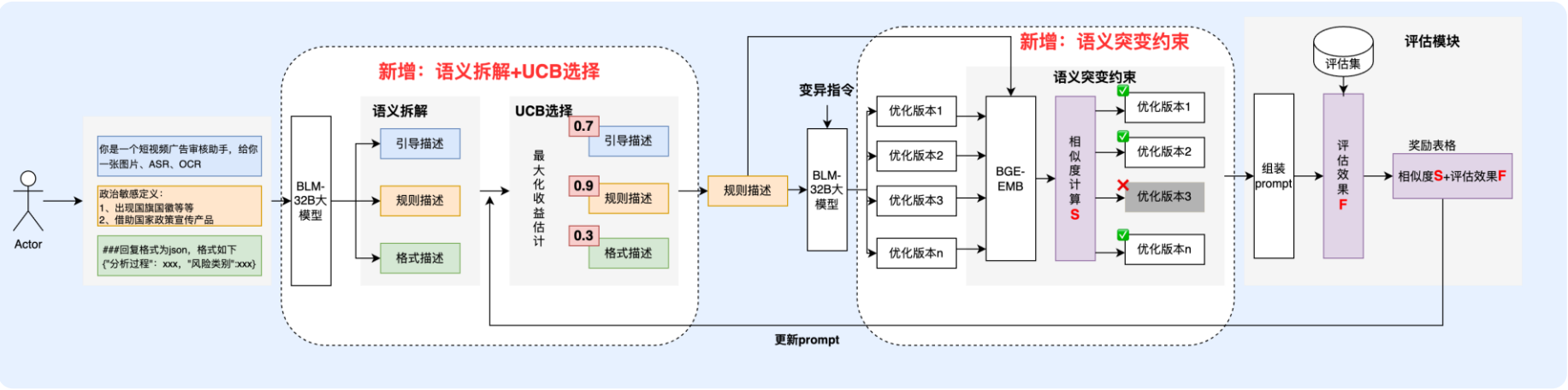
2、风险研判智能化，在风险漏放与误伤优化场景中，可直接输出研判结果，减少对QI研判环节的依赖，提升处理效率与响应速度。



运营也可以“微调”模型，做模型的规则“教练”

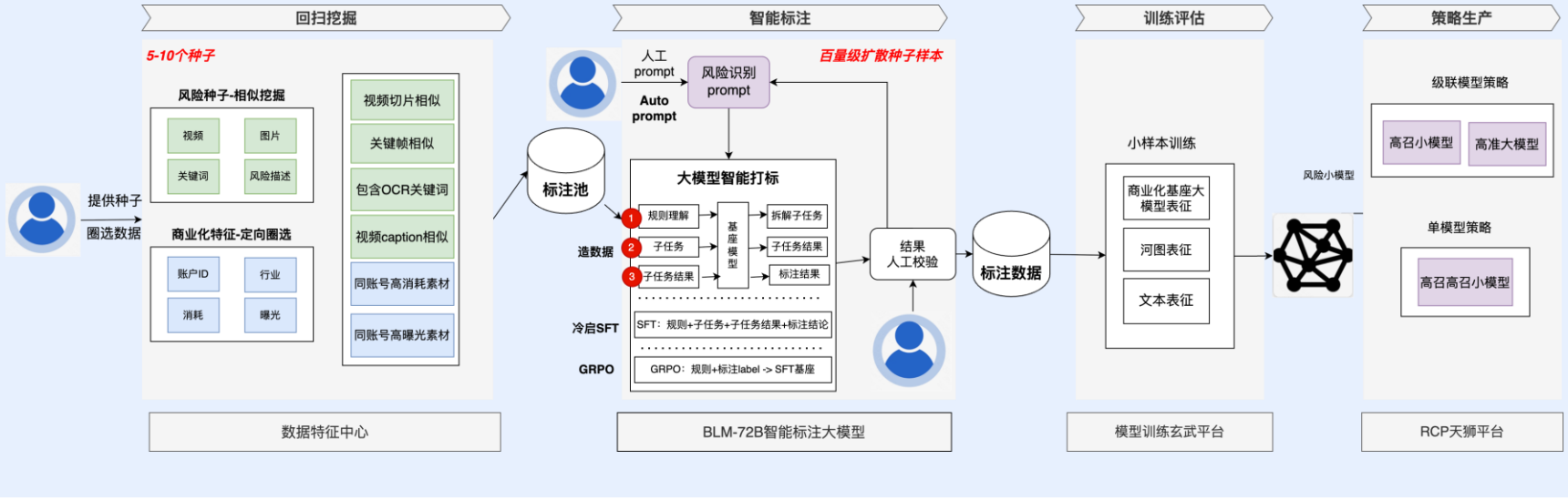
APE

运营角色，初始化风险Prompt种子，自动语义拆解+UCB选择，探索最优风控Prompt效果



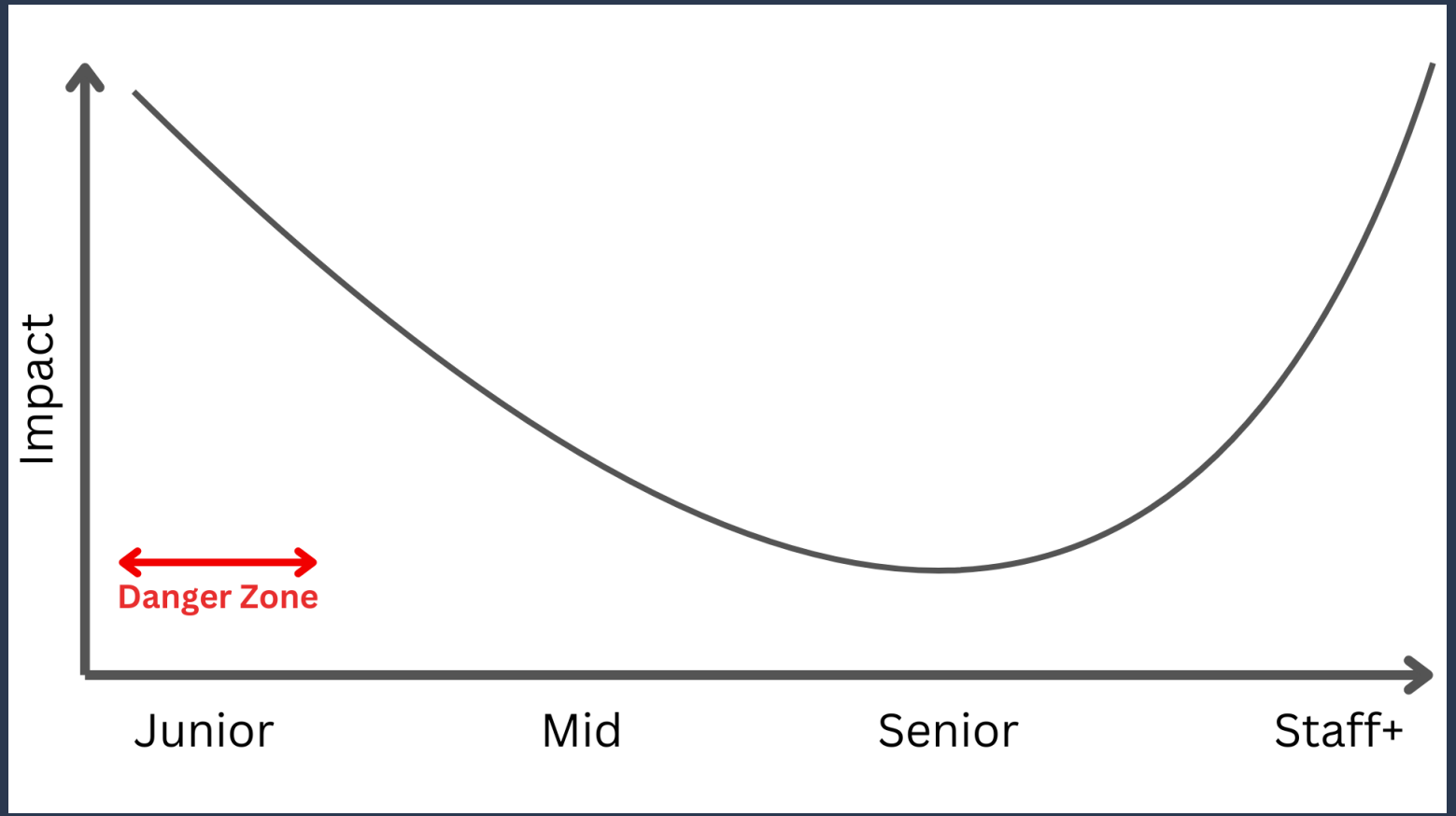
AML

运营角色，初始化风险内容种子，自动化挖掘、标注、扩散，并完成模型迭代



研发工程师面临的挑战

如何逃离
LLM带来的
Danger Zone?



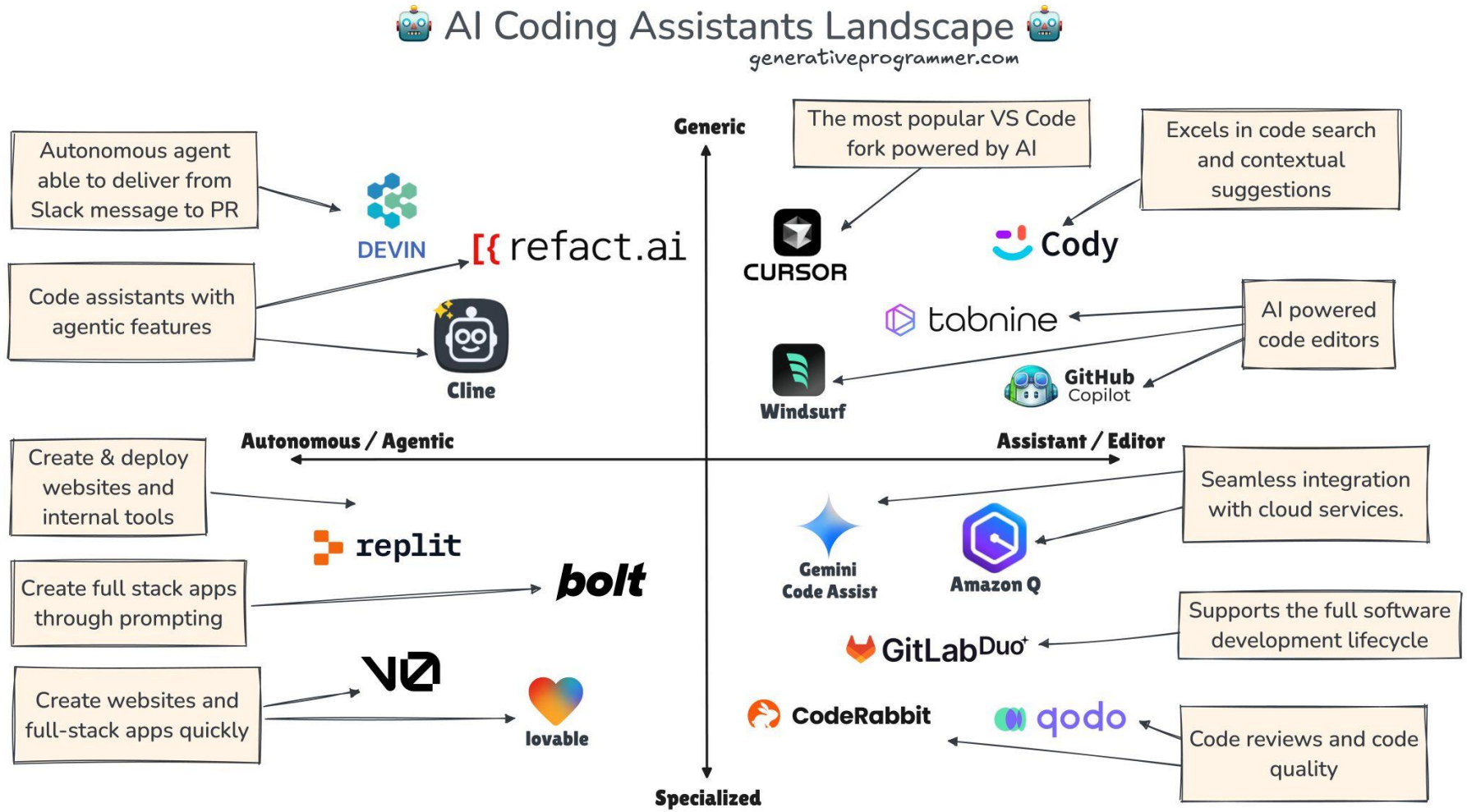
AI辅助编码：利用AI提效

AI大幅提升编码效率

1、重复性工作大规模通过AI实现自动化，工程师职能向高价值领域上移

3、好的程序员，不再是编码最快的，而是最善于编排利用AI的

2、研发工程师，也可以自己就是一个“团队”，团队员工是各种AI Coding Agent



构建面向LLM的风控系统

风控 LLMOps

风控大模型训练推理平台

Deepseek Qwen ...

LLM Selection

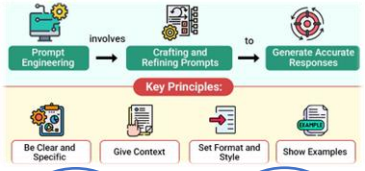
Data Integration

Fine–Tunning

In–Context–Learning & Embeddings

RLHF

Prompt运营



Prompt Engineering

Context Mgmt.

Model Deployment

模型性能监测

Model Performance /Monitoring

Model Governance

Online Inference

模型推理加速

模型量化、压缩

KV Cache

ViT拆分

大模型审核系统

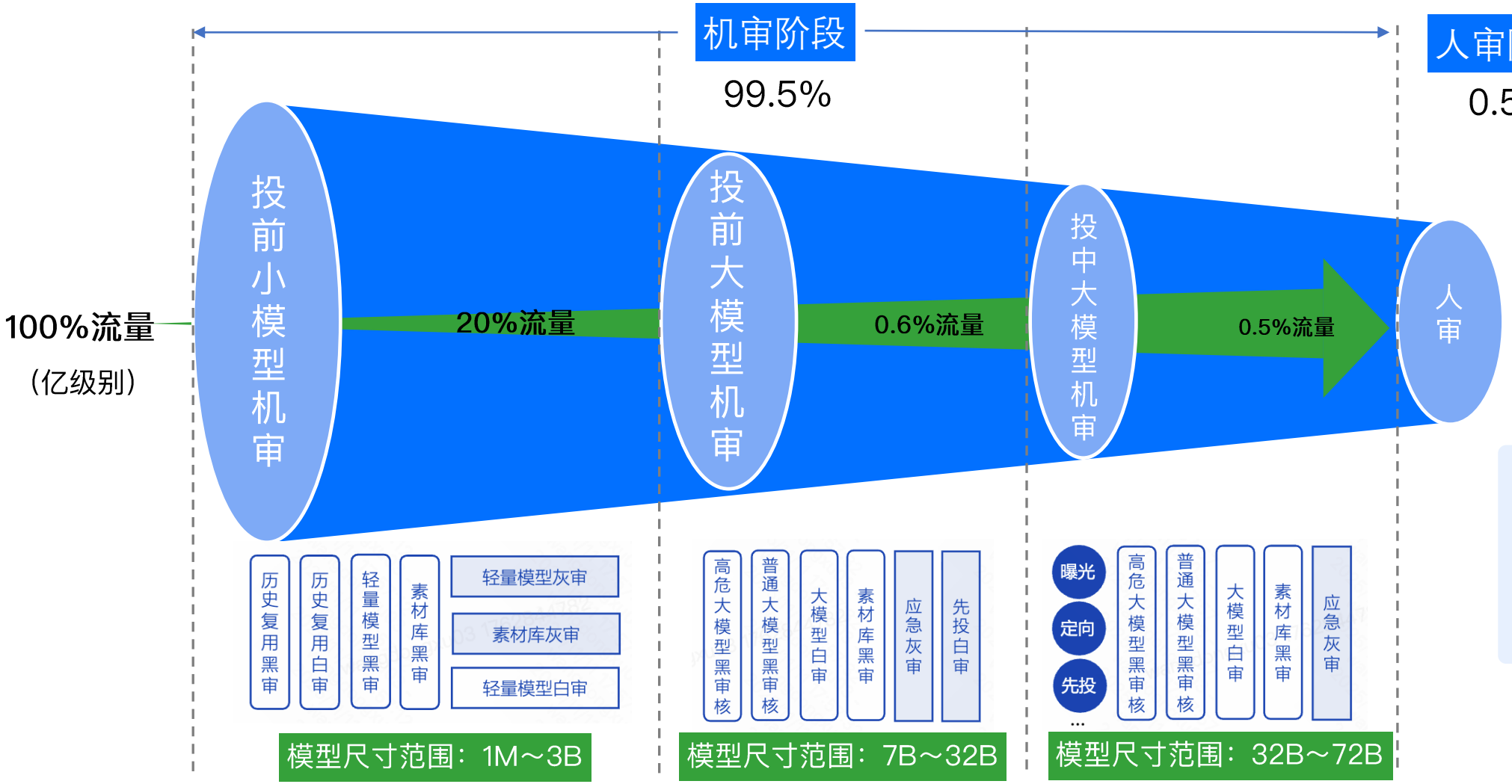
特征计算

请求路由

LLM 异常检测

多路合并

大模型审核系统



架构特点

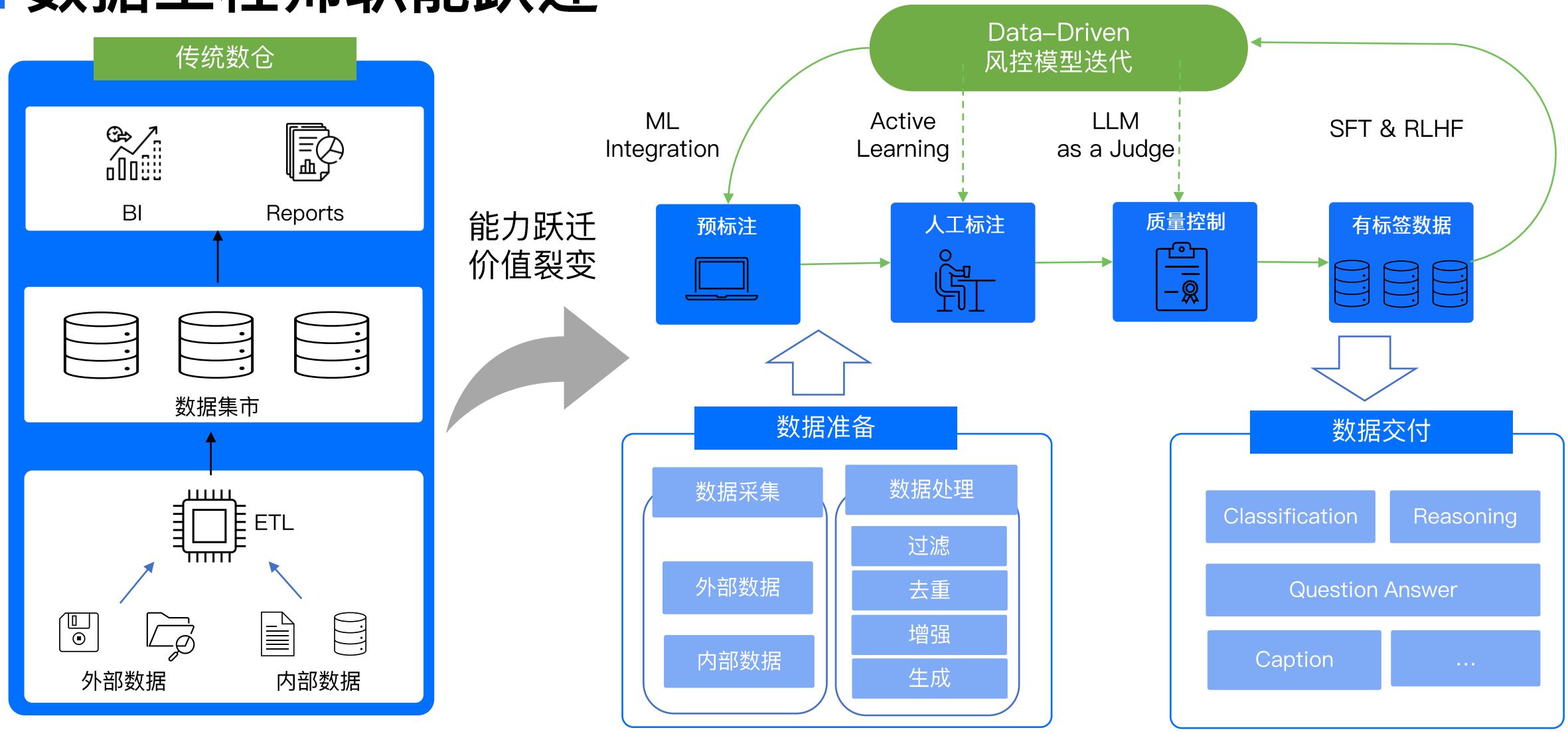
1、漏斗式审核架构，逐层过滤审核流量

2、机审覆盖>99.5%

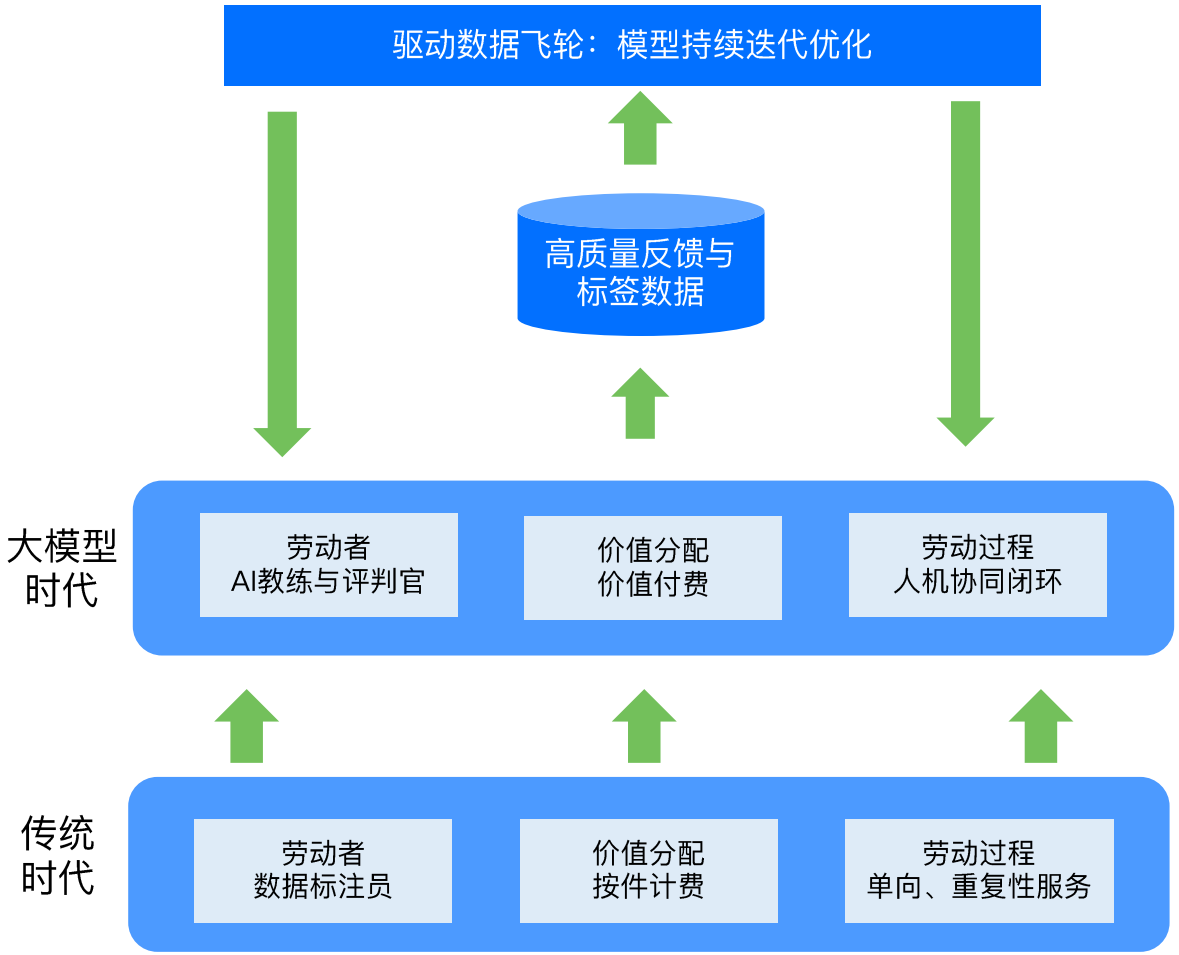
数据工程师面临的挑战

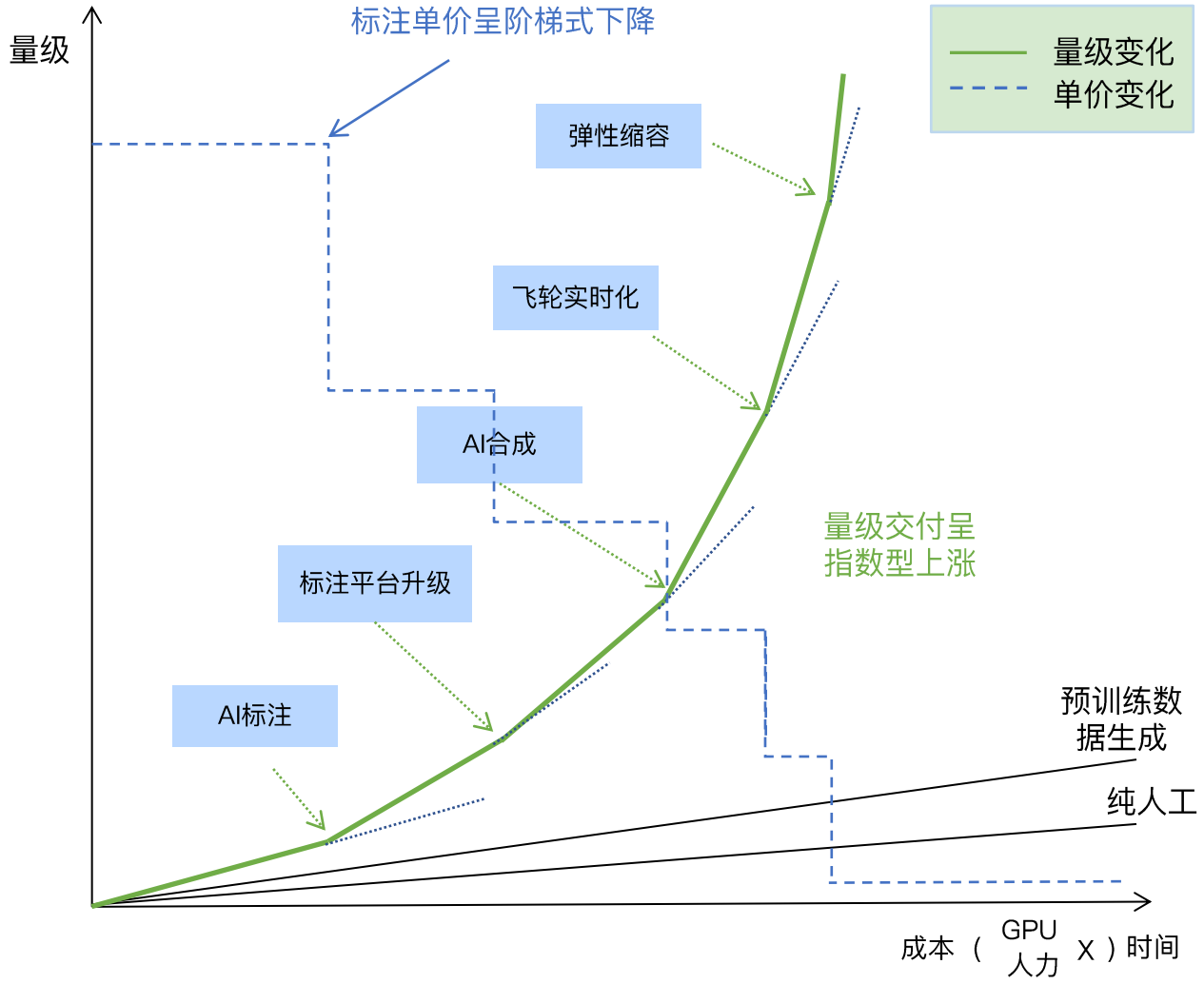
作为Data Engineer
只做BI报表，
怎么结合AI，
让自己升级为Data Scientist?

数据工程师职能跃迁


传统数仓
BI
Reports
数据集市
ETL
外部数据
内部数据
能力跃迁
价值裂变
Data-Driven
风控模型迭代
ML
Integration
Active
Learning
LLM
as a Judge
SFT & RLHF
预标注
人工标注
质量控制
有标签数据
数据准备
数据采集
外部数据
内部数据
数据处理
过滤
去重
增强
生成
数据交付
Classification
Reasoning
Question Answer
Caption
...

数据价值生产逻辑的变化





算法工程师面临的挑战

大模型平权，技术门槛降低，
人人都是AI原住民
算法的核心竞争力是什么？

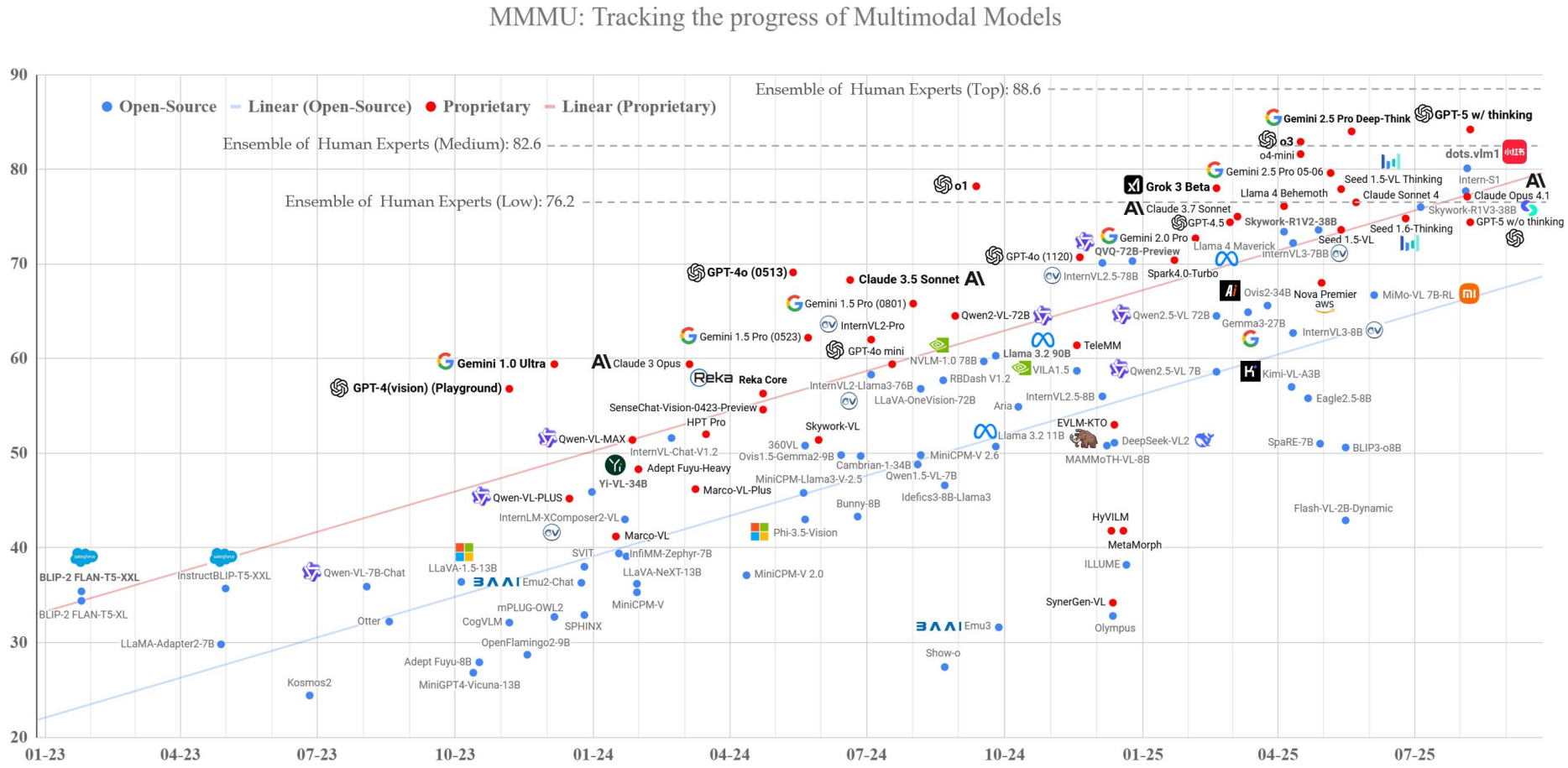
多模态大模型的发展进程

MLLM技术已经成熟

1、2024年9月，O1模型首次超越低等水平人类专家（76.2分）

2、2025年5月，Gemini 2.5 Pro Deep–Think模型首次超越中等水平人类专家（82.6分）

3、MLLM发展的趋势仍在继续

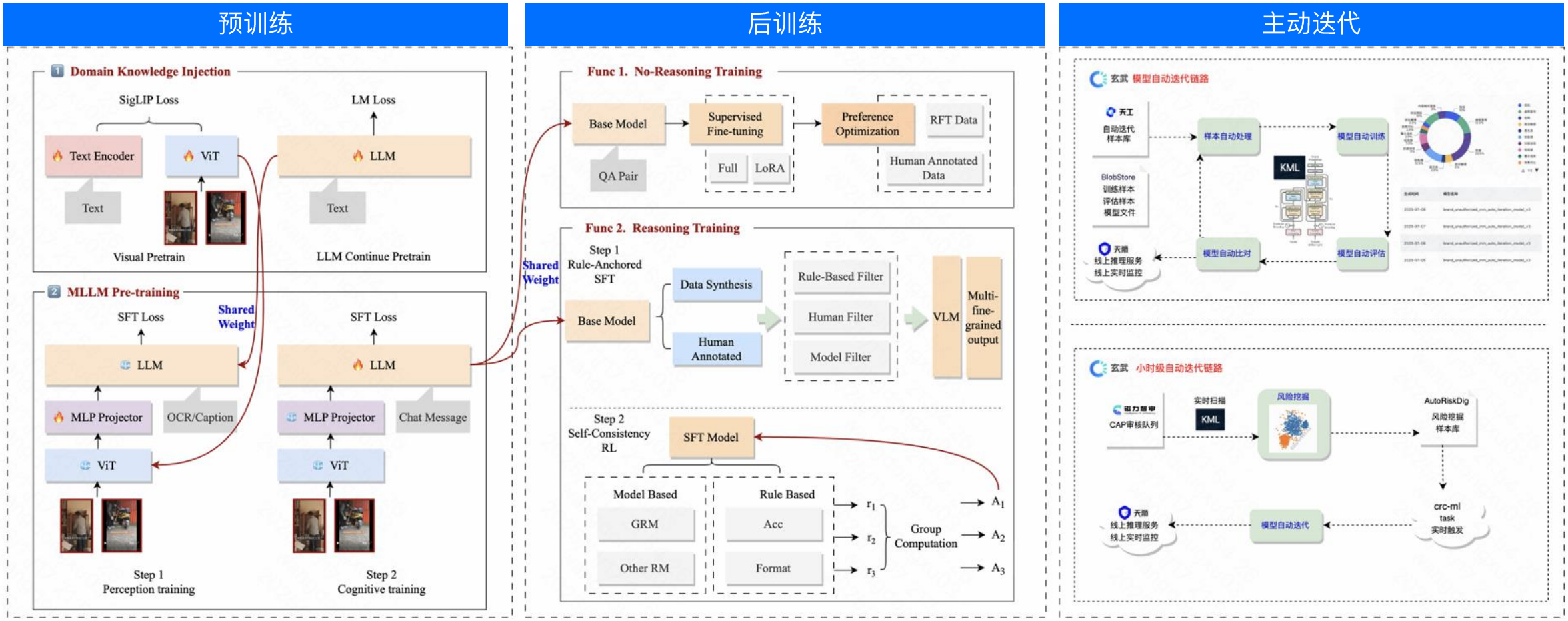


https://mmmu-benchmark.github.io/

算法：向上游业务突破，大模型审核

25Q1: 大模型审核初探	25Q2: 素材AI修复	25Q3:智能运营模式	25Q4: 规模化业务扩张
早期探索期	**技术创新期**	**生产关系转换期**	**成熟期**
• 从广告语开始探索文本大模型审核 • 基于模型微调方案落地试点 • 大模型训练推理平台初步构建成型	• 不单是创意审核，还要对创意修复 • 不限于内容理解模型，理解+生成 • 识别+定位一体化，精细化刻画风险	• 从算法主导到运营主导智能审核 • Workflow、RAG、Agent配套 • 自动化PE能力简化Prompt运营	• 覆盖文本、图像、短视频场景 • 围绕KwaiBLM大模型形成对抗体系 • 从MLLM迈向Multi-Agent架构
算法策略			
LoRA 模型微调 训推能力	风险排序 Policy召回 Router过滤	风险识别 风险定位 AI生成修复	识别定位一体化 分镜替换、打码 创意过审	自动化PE 拒绝理由精细化 MCP拓展	RAG规则库 CoT蒸馏 WorkFlow	全场景覆盖 Stella数据飞轮 Multi Agent	锐鉴SharpSight RiskScan智能回扫 A2A协同
平台沉淀			
KwaiBLM风控预训练大模型	AhaEdit一键修复	OpHub智能运营	RiskAgent Family

算法：向下深度突破，垂域预训练大模型BLM

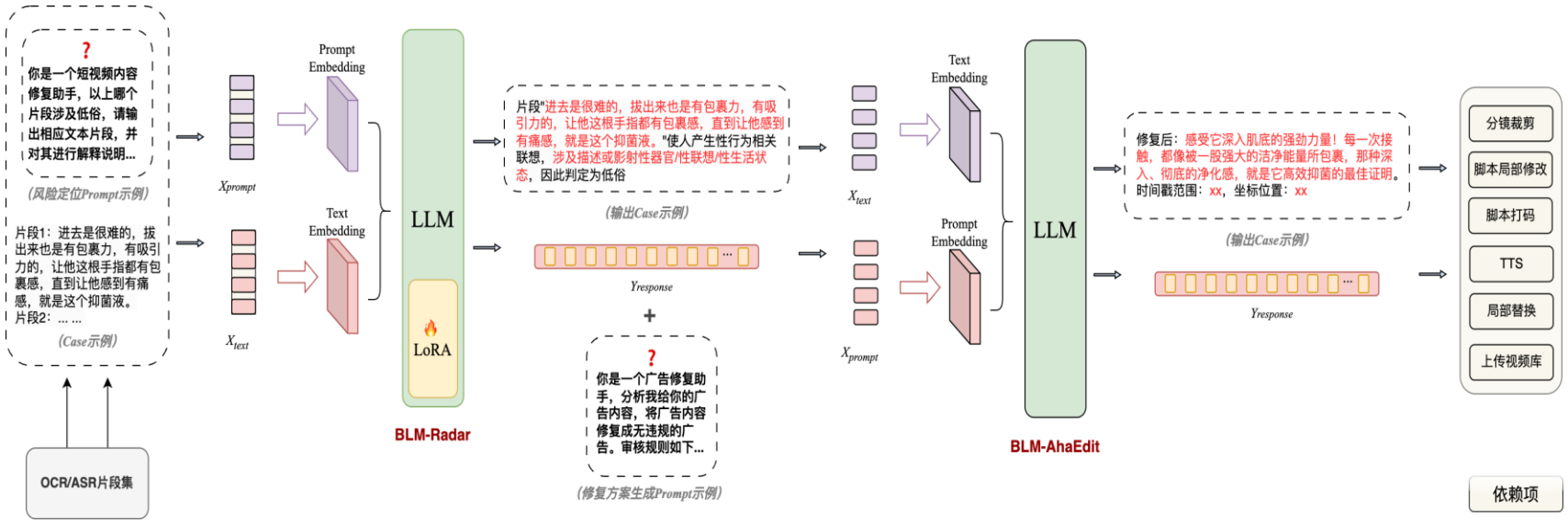


算法：向前技术创新，理解+生成统一新范式

KwaiBLM-AhaEdit

1、不再止步于过去单一的内容理解，而是构建了内容理解与生成的统一能力，即在识别风险的同时，提供智能化的修复方案

2、对中小自助客户意义重大：有效解决了广告主不理解审核规则、难以精准定位问题的痛点，助力广告主长效经营



风险识别&定位
(BLM-Radar)
200+拒绝理由规则注入

风险修复
(BLM-AhaEdit)
效果成本约束下尽可能做最好的选择

素材生产

组织职能重塑和考核指标变化

产品经理 (PM)

职能重塑

- 手写PRD –> Vibe Coding原型设计
- 通过AI辅助PRD质量

AI考核指标

- Vibe Code Design比例
- AI评审PRD通过率
- AI产品系统设计数量

原型设计产能 +100%

策略运营 (OP)

职能重塑

- 写规则 –> PE运营
- Zero-shot，T+H布防
- AI的规则“教练”

AI考核指标

- PE/Workflow/RAG数量
- PE业务接入率
- 驱动模型SFT管道数量

对抗周期 周–>小时

工程研发 (RD)

职能重塑

- AI辅助编码
- Vibe Coding
- LLM架构“设计者”

AI考核指标

- AI代码入库率
- AI算法工程Repo贡献
- 单人研发吞吐量 (SP)

Coding效率 +30%

数据工程(DE)

职能重塑

- BI –> BI + AI
- 大模型自动标注
- 提供模型迭代的“燃料”

AI考核指标

- 标注自动化率 > 70%
- Chat BI找数成本
- 多模态样本资产量级

标注自动化率 > 70%

算法 (Algo)

职能重塑

- 判别式–>理解+生成统一
- 垂直领域预训练
- 生产关系重塑的“发起者”

AI考核指标

- 传统算法迁移率
- Agent覆盖率
- Pre-Training/RL增益

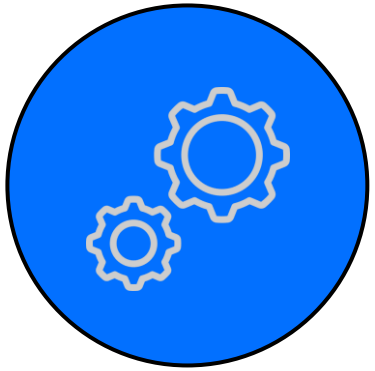
垂直领域基座覆盖>80%

03

AI驱动协同模式升级

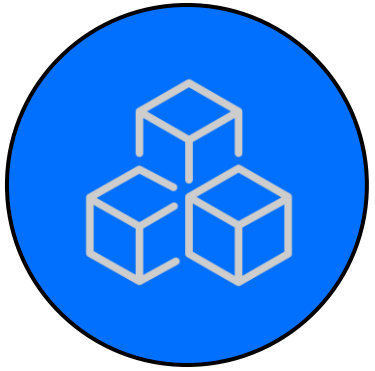
构建“人机协同”的AI增强型安全系统

三种协同模式



大小模型协同

- 小模型实时拦截，大模型深度研判
- 资源智能调度，精准与效率兼顾



Multi Agent协同

- 多Agent分级防控，风险立体覆盖
- 动态协同决策，提升系统鲁棒性

HITL人机混合协同

- AI预审高可疑，人工复核保精准
- 人机优势互补，降低误判漏报率

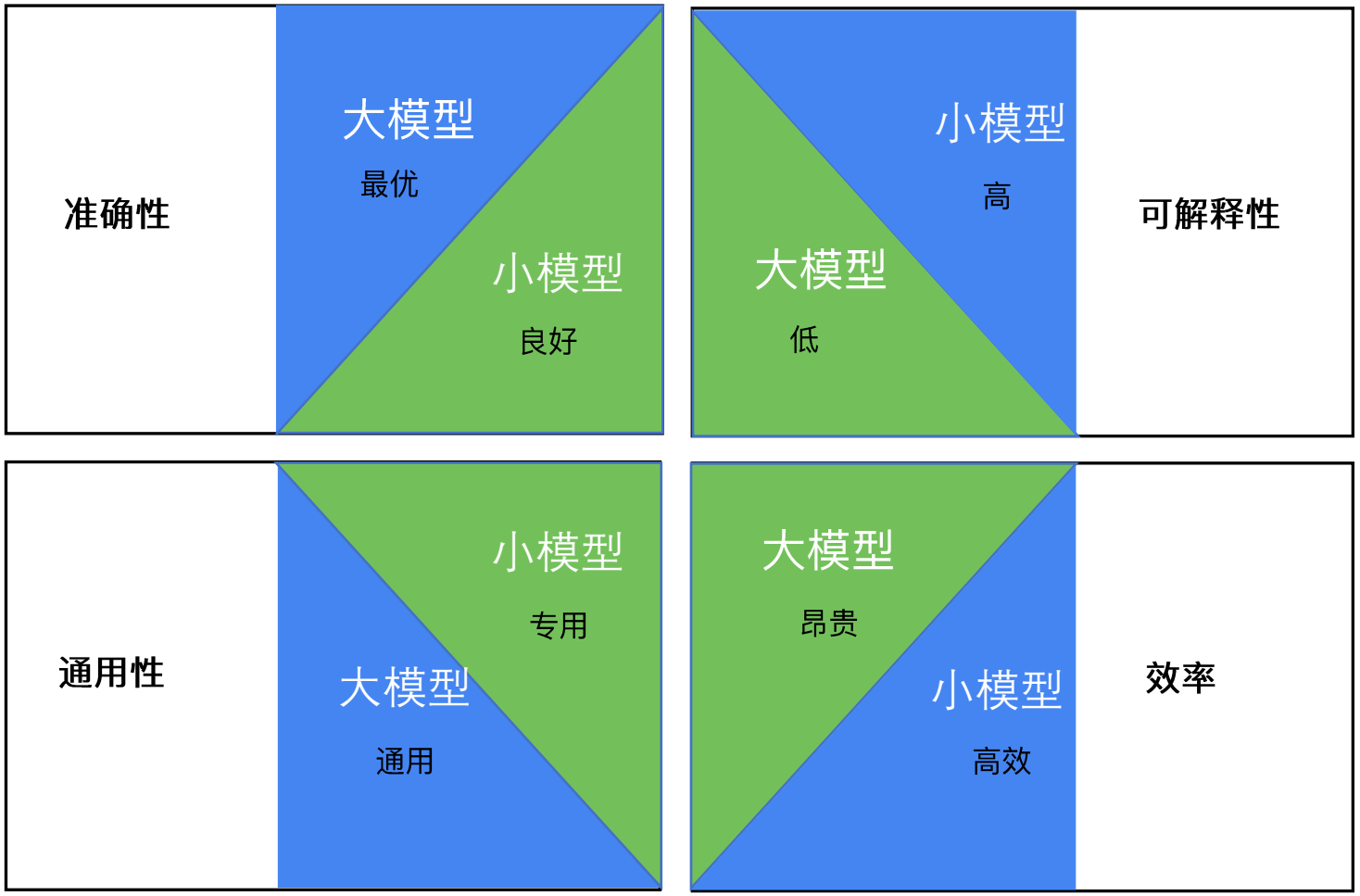
协同模式1：大小模型协同

大模型增强小模型

- 大模型知识蒸馏，强化小模型识别能力
- 大模型生成训练数据或者打标，小模型获取高质量标注样本

小模型增强大模型

- 通过小模型Routing请求，先做粗筛选
- 过滤掉大部分不相关请求，然后再通过二阶段过大模型精判风险



协同模式2: Multi Agent协同

风险环境

Risk Agent 门户

规划Agent

- 风控任务拆解
- 执行路径选择
- 路由调度
- A2A编排
- 协调分歧
- 结果聚合
- 风险终判

感知Agent — MCP | LLM | 记忆

A2A协议

识别Agent — MCP | LLM | 记忆

A2A协议

反思Agent — MCP | LLM | 记忆

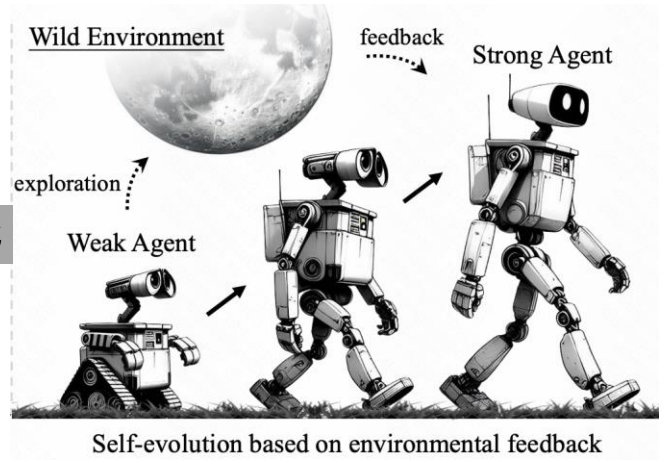
A2A协议

仲裁Agent — MCP | LLM | 记忆

A2A协议

反馈Agent — MCP | LLM | 记忆

反馈迭代



Self-evolution based on environmental feedback

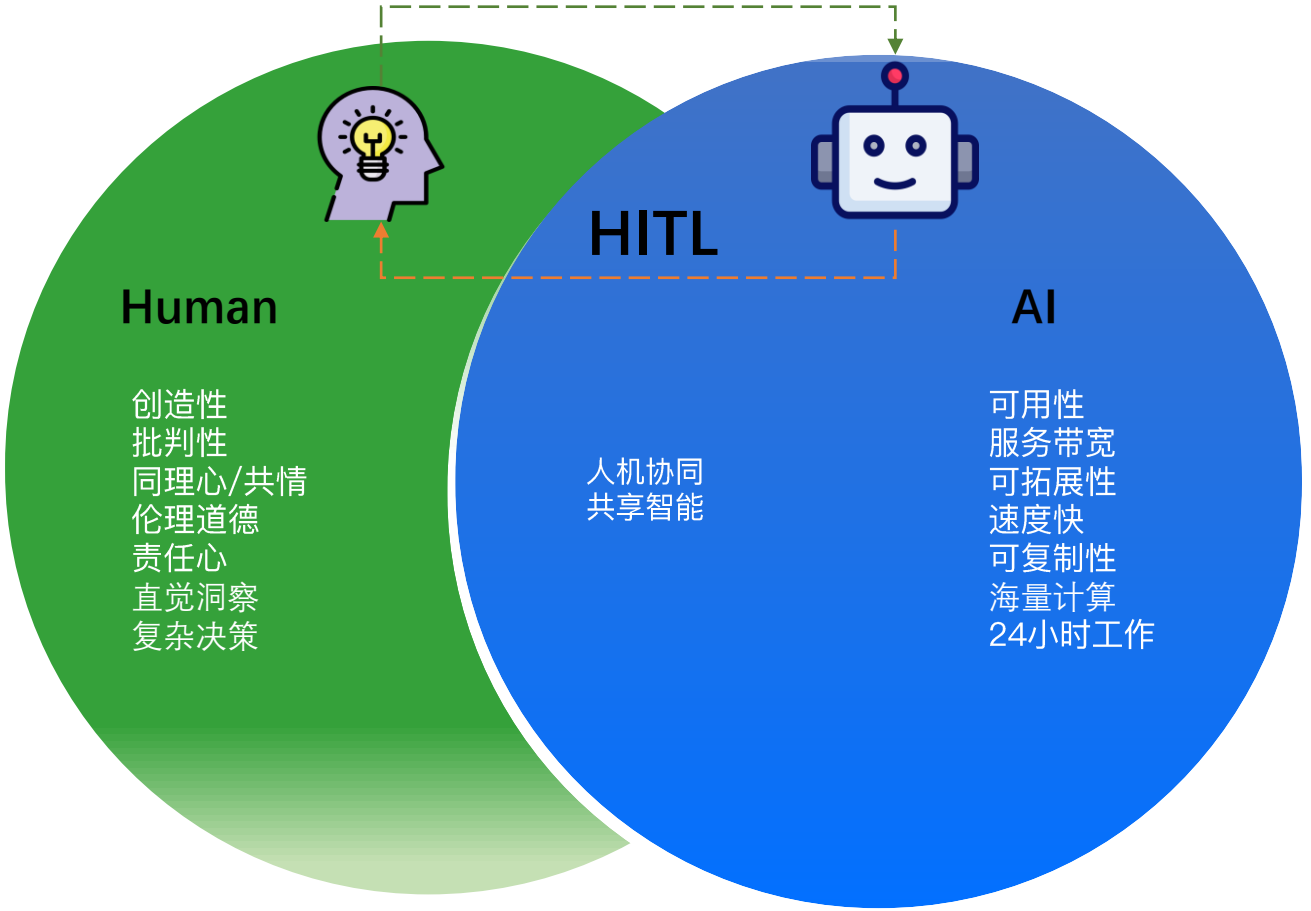
协同模式3: HITL人机混合协同

AI辅导、增强人的能力

- 批量内容预审，降低人工工作量
- 智能风险评分，优先处理高危内容
- 多语言实时翻译，跨语言审核支持
- 历史案例学习，辅助决策参考

人辅导、增强AI的能力

- 纠正AI误判，持续优化模型准确率
- 标注边界案例，扩展AI理解能力
- 更新审核规则，适应政策变化
- 提供文化背景，增强语境理解



04 未来展望

打造AI-Native型安全组织

AI-Native内容安全蓝图

以人为主
人类主导审核和治理，机器作为辅助。
审核成本较高，适用于中小体量的业务规模。

人机协作
机器主导审核，人类和机器协同治理监督。
审核成本大幅度降低，人类从执行者变成指挥官。

以AI为主
机器主导审核和治理，形成高度
自治的闭环，人类仅承担审计职责。

Full Human Review (L1)	Hybrid Review (L2)	Full Machine Review (L3)	Human in the Loop (L4)	AGI Governance (L5)
• 审核全部由人类完成，治理全部由人类完成。 • 可支撑业务体量小，处于初步构建好业务**审核链路**的阶段。	• 大部分审核由机器完成、少部分高难度审核由人类完成，治理全部由人类完成。 • 可支撑业务体量较大，应用**Xgboost**、**CNN**、**NeXtVLAD**、 **BERT**和**YOLO**等**机器学习**算法来进行机器审核。	• 审核全部由机器完成，治理全部由人类完成。 • 可支撑业务体量大，应用了**LLM**和**VLM**等**大模型技术**替代了人类完成高难度样本的审核，大幅度降低了审核成本。	• 审核全部由机器完成，大部分治理由机器完成、少部分治理由人类完成。 • 可支撑业务体量大，应用**Agentic AI**技术自动完成离线治理，人类只提供政策规则和提供少量的高质量反馈信号。	• 审核全部由机器完成、治理全部由机器完成。 • 可支撑业务体量大，**AGI**在感知、规划和行动方面全面超越人类，人类只需要下达命令，AGI高质量地**自主完成**所有审核和治理任务。

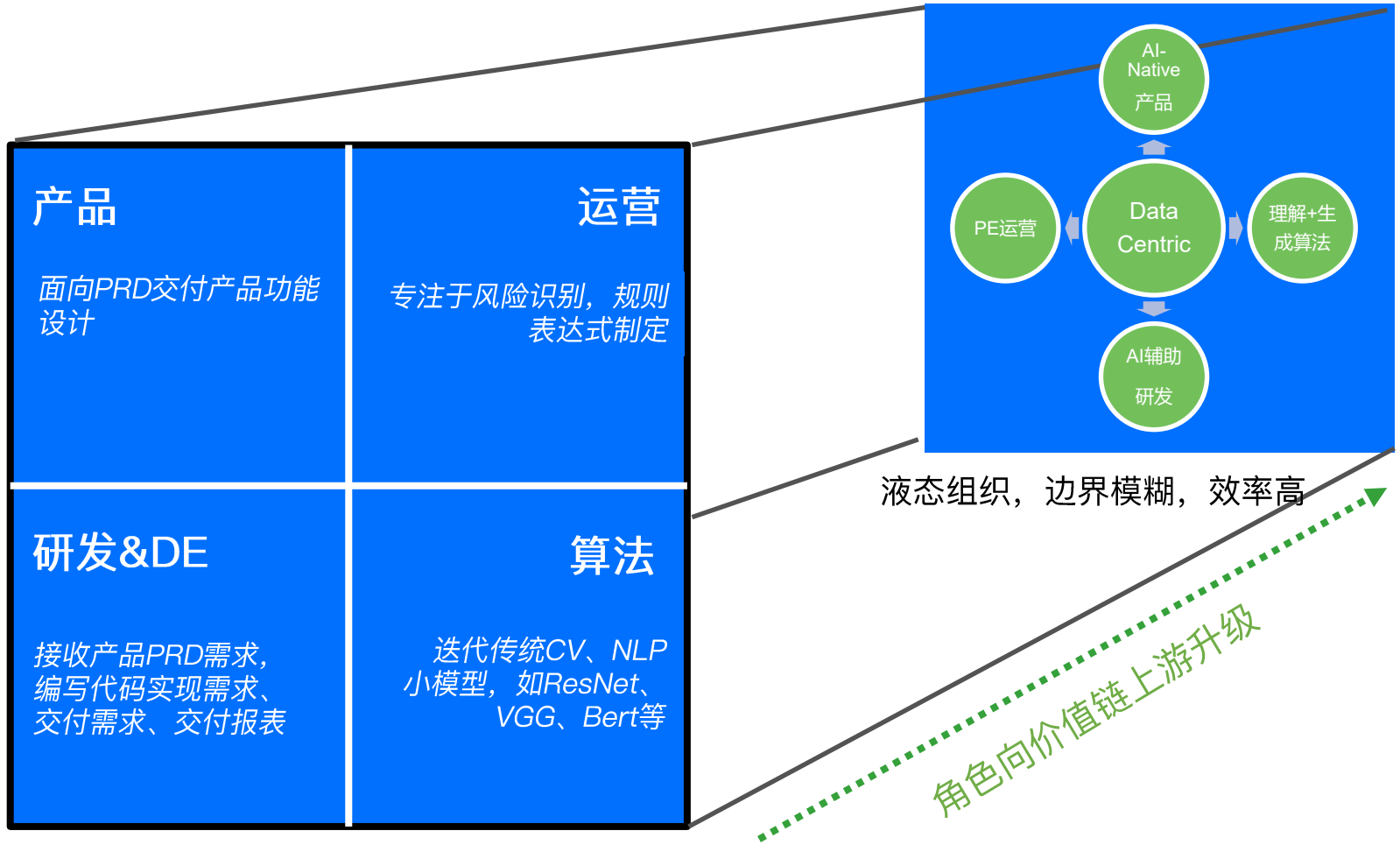
AI-Native组织：从“师级单位”到“AI合成旅”

从

“固态组织”（师）

到

“液态组织”（AI合成旅）



AI时代，给听众的3个行动建议

① 每个角色都应该持续向价值链上游升级

② 速度是唯一的护城河

③ 先完成、再完美，进化的速度比起跑的姿势更重要

极客邦科技 2026 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

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

InfoQ 极客传媒

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