

大小模型协同驱动安全升级

基于大小模型协同的数字内容风控实践

胡宜峰

网易 算法专家



目录

01 背景与挑战

02 大小模型协同解决方案设计与分析

03 基于大小模型协同的数字内容风控实践

04 未来展望

📍 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

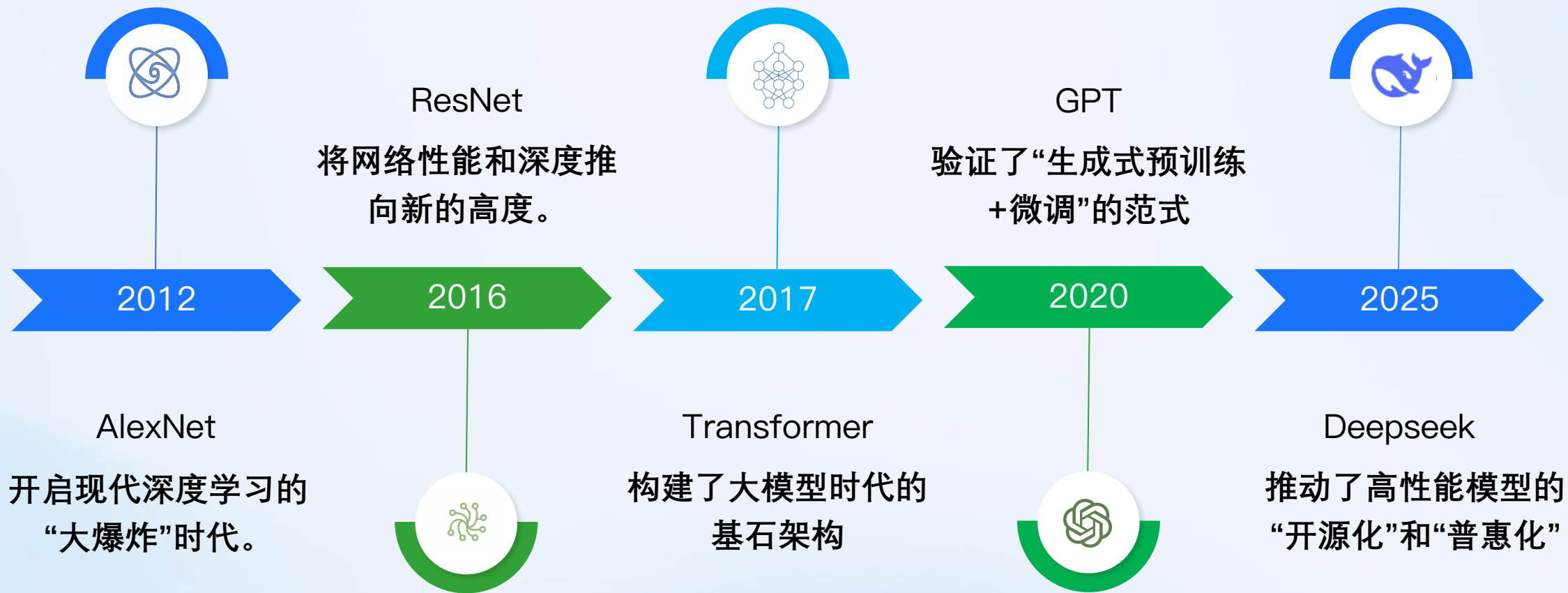


查看会议

01 背景与挑战

背景

发展历程



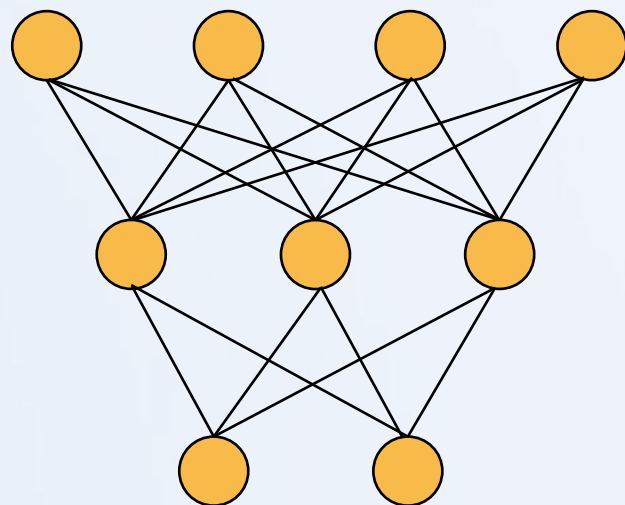
● 小模型



● 大模型



选择



小模型

VS



OpenAI

Gemini

LLaMA

Claude

deepseek

cohere

Baidu 百度

Qwen2.5

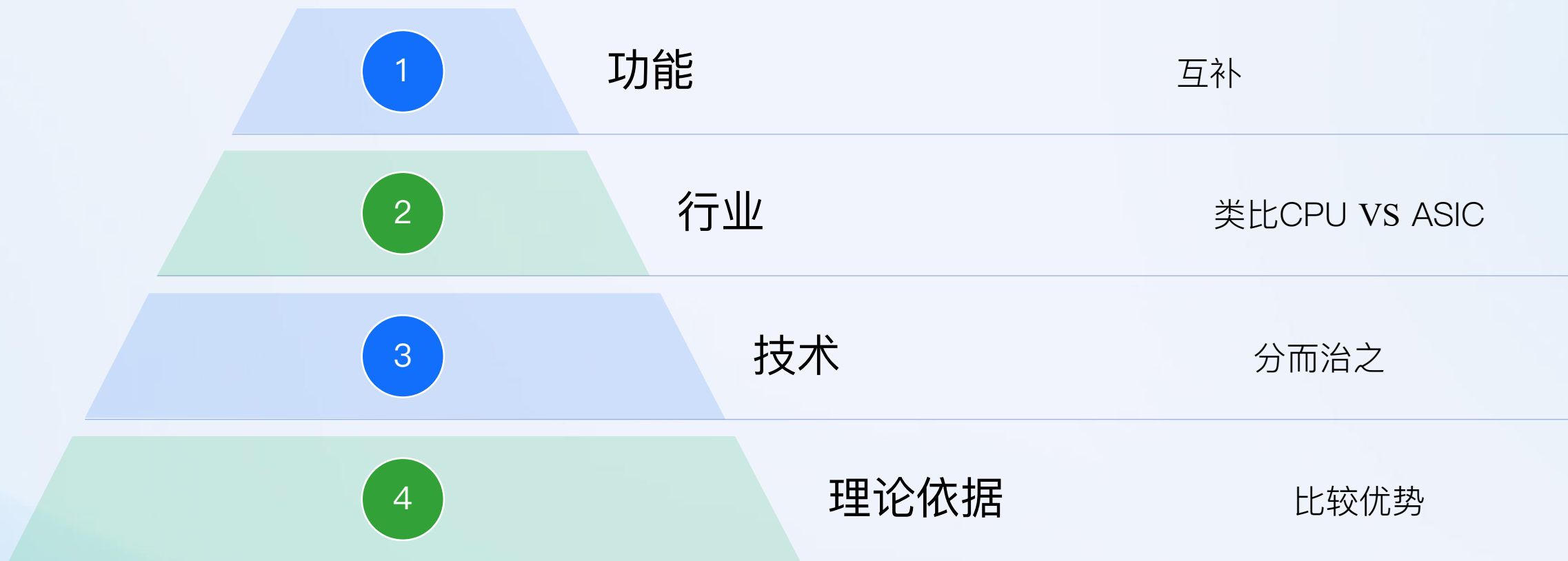
stability.ai

MISTRAL AI

IBM
Granite

大模型

选择

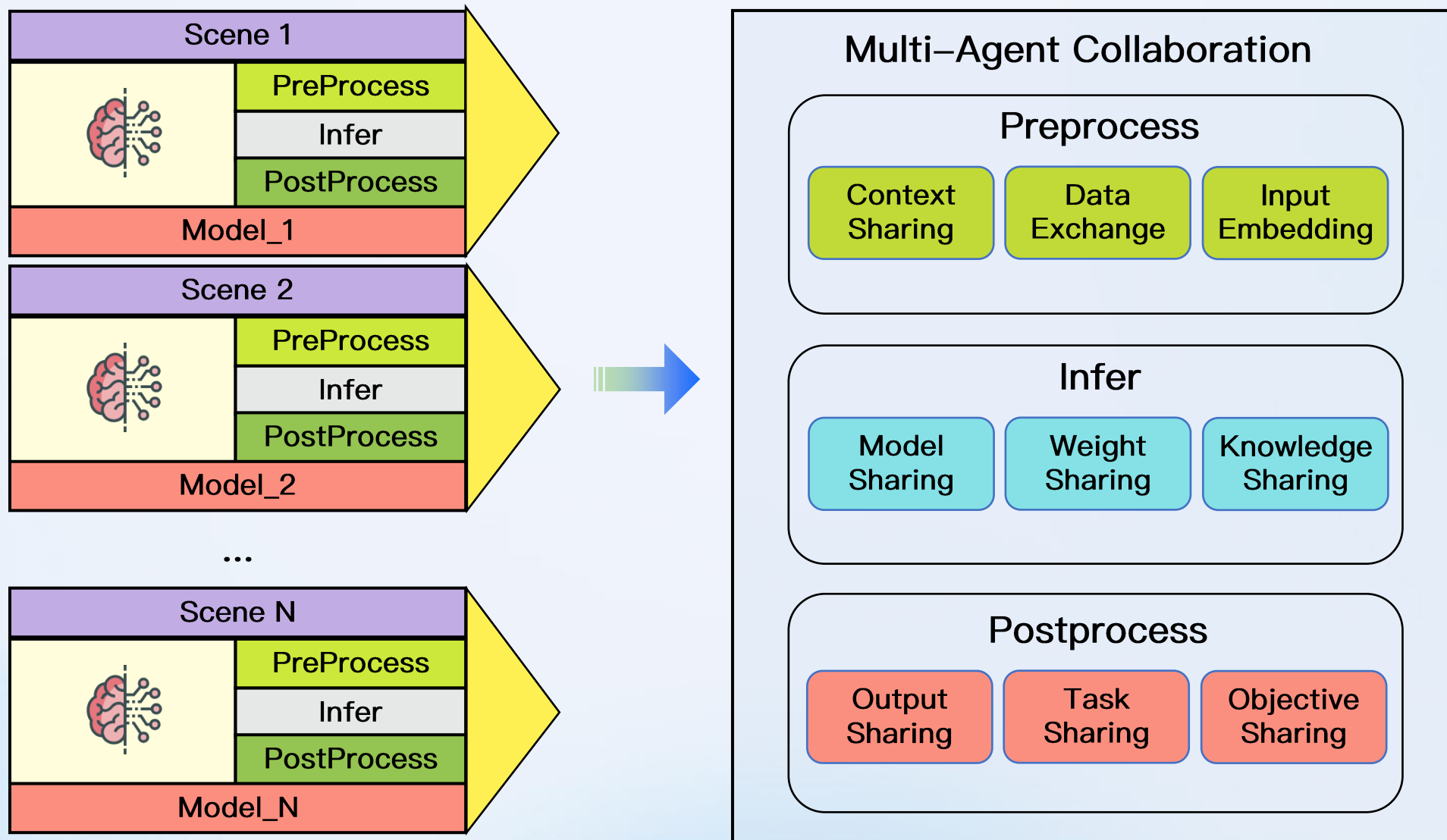


大模型 VS 小模型



协同发展

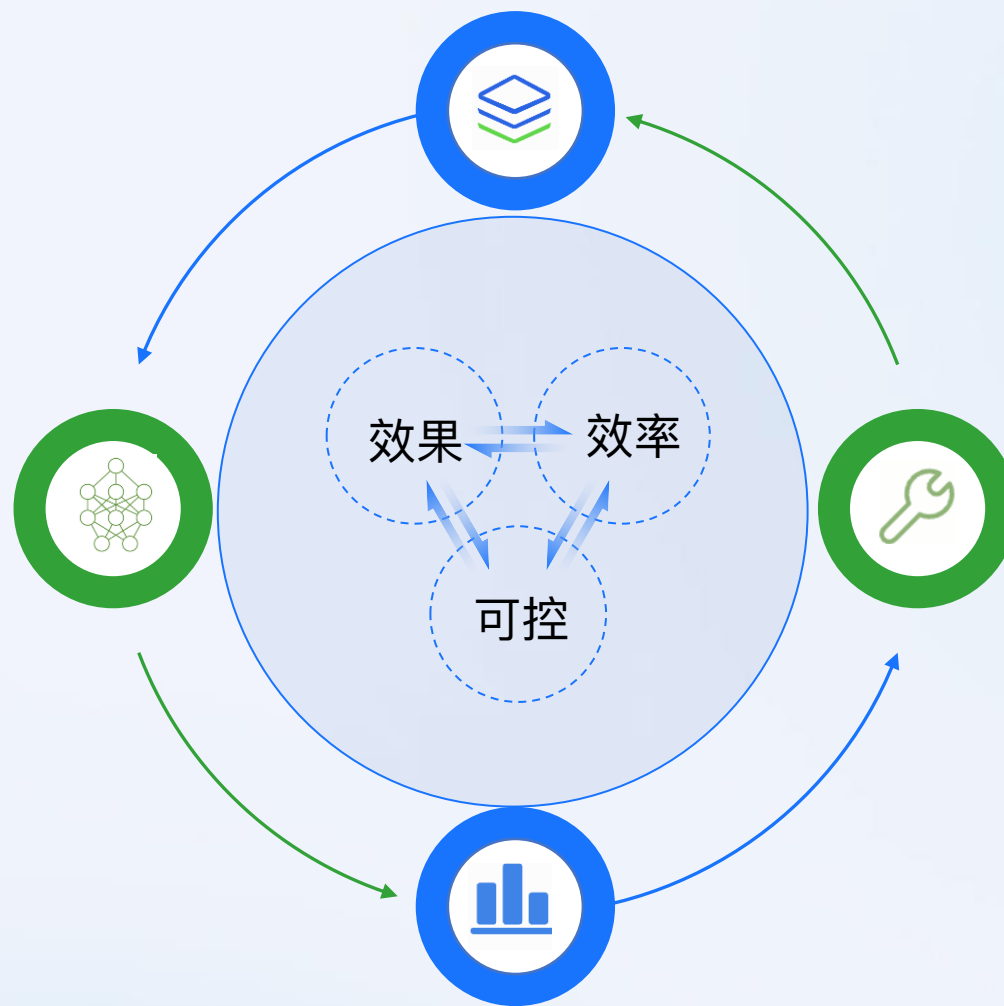
新范式-从孤立模型到协同系统



挑战

大小模型协同

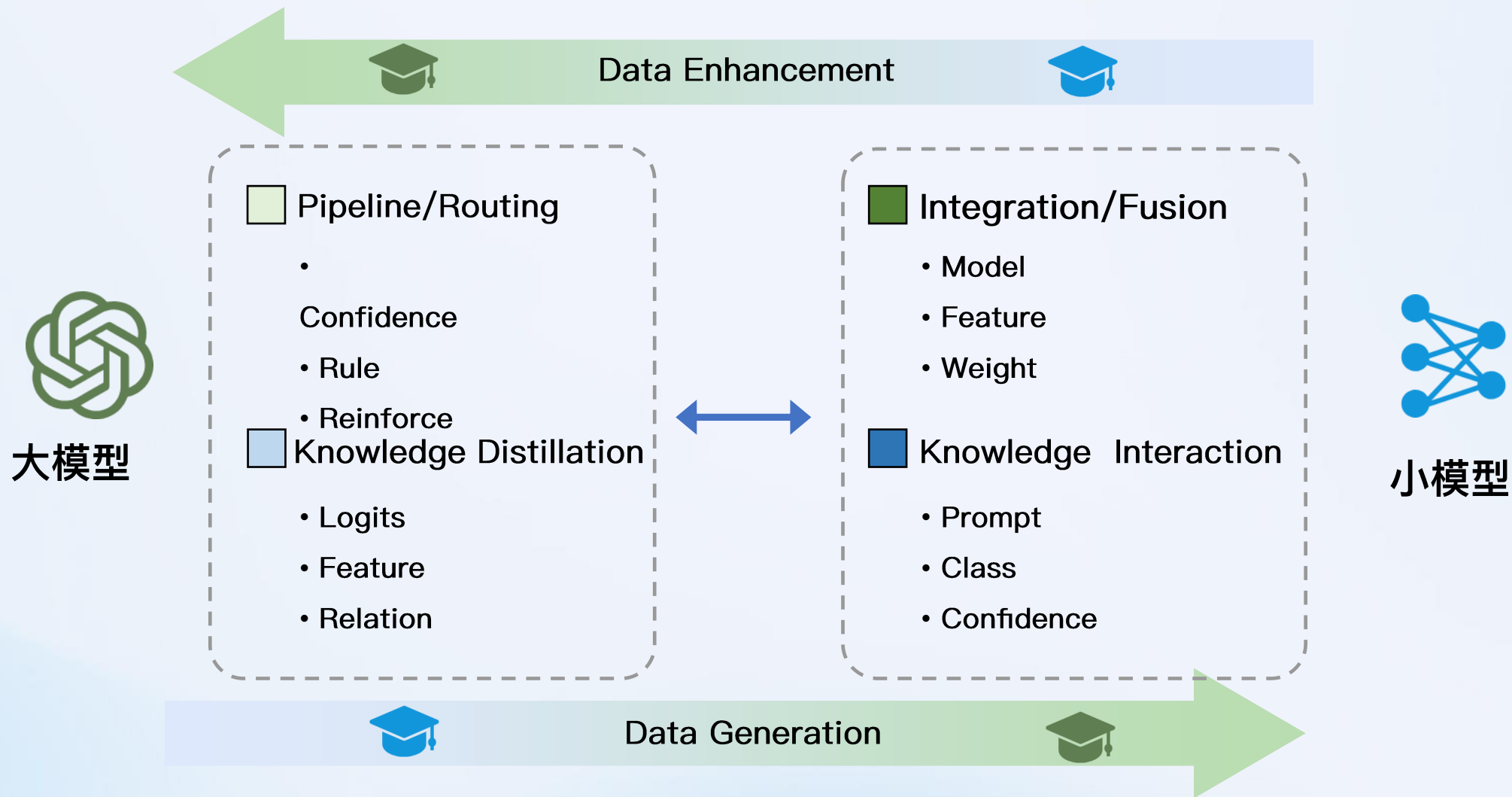
- 解决方案
任务拆解、模型选择、pipeline设计、模型路由、
信息交互、融合策略
- 模型
蒸馏、知识迁移
- 数据
数据量、准确性、差异性、价值
- 工程
时延、并发、迭代周期



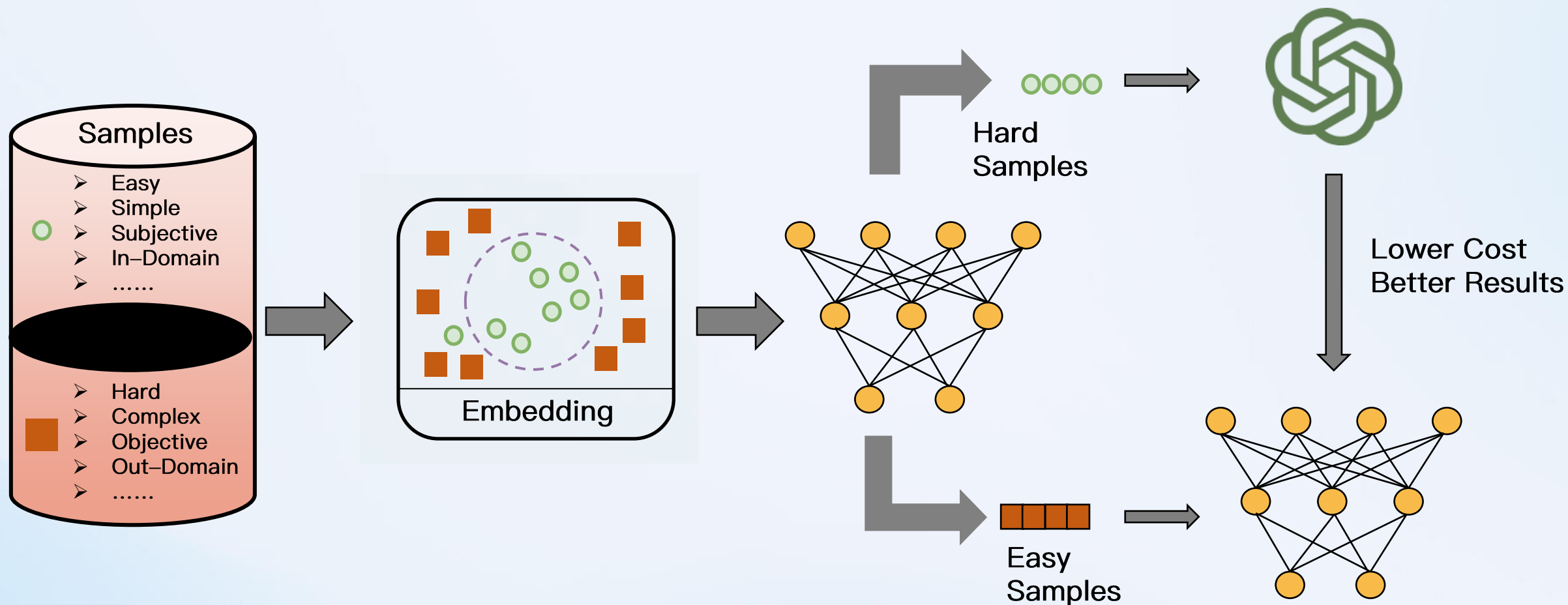
02

大小模型协同解决方案

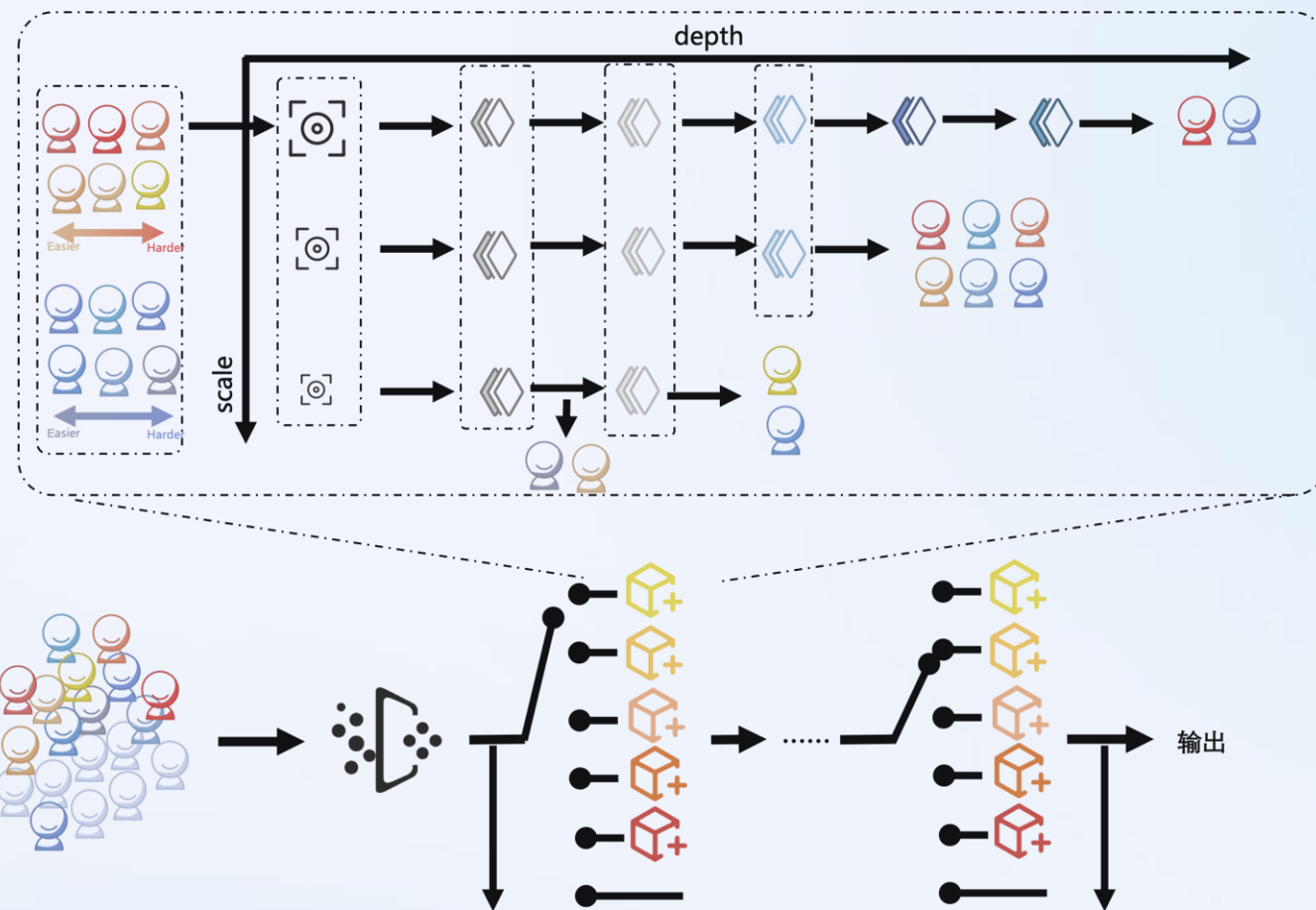
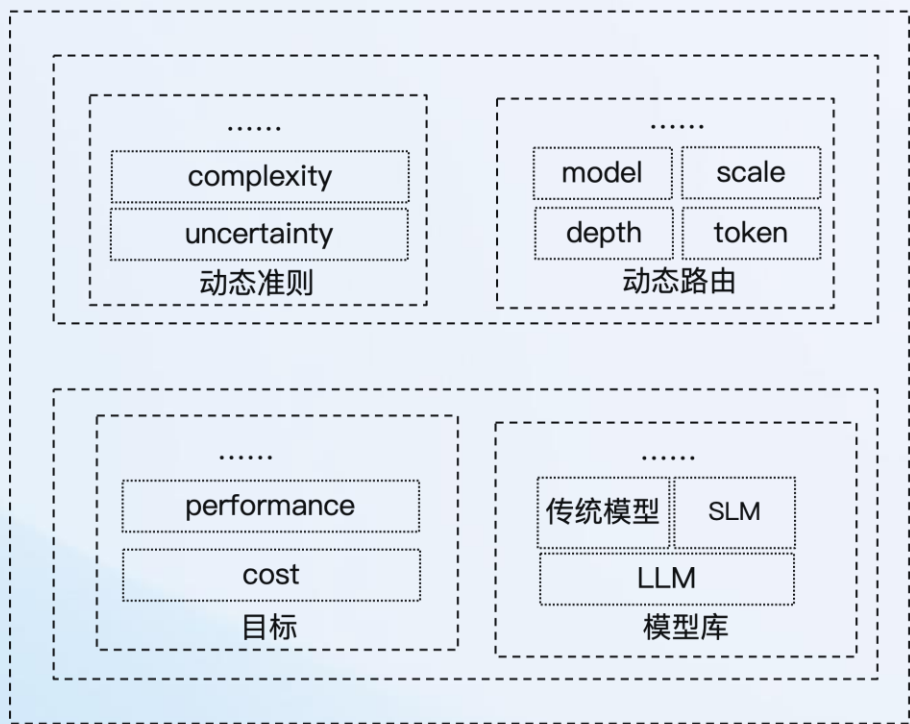
大小模型协同



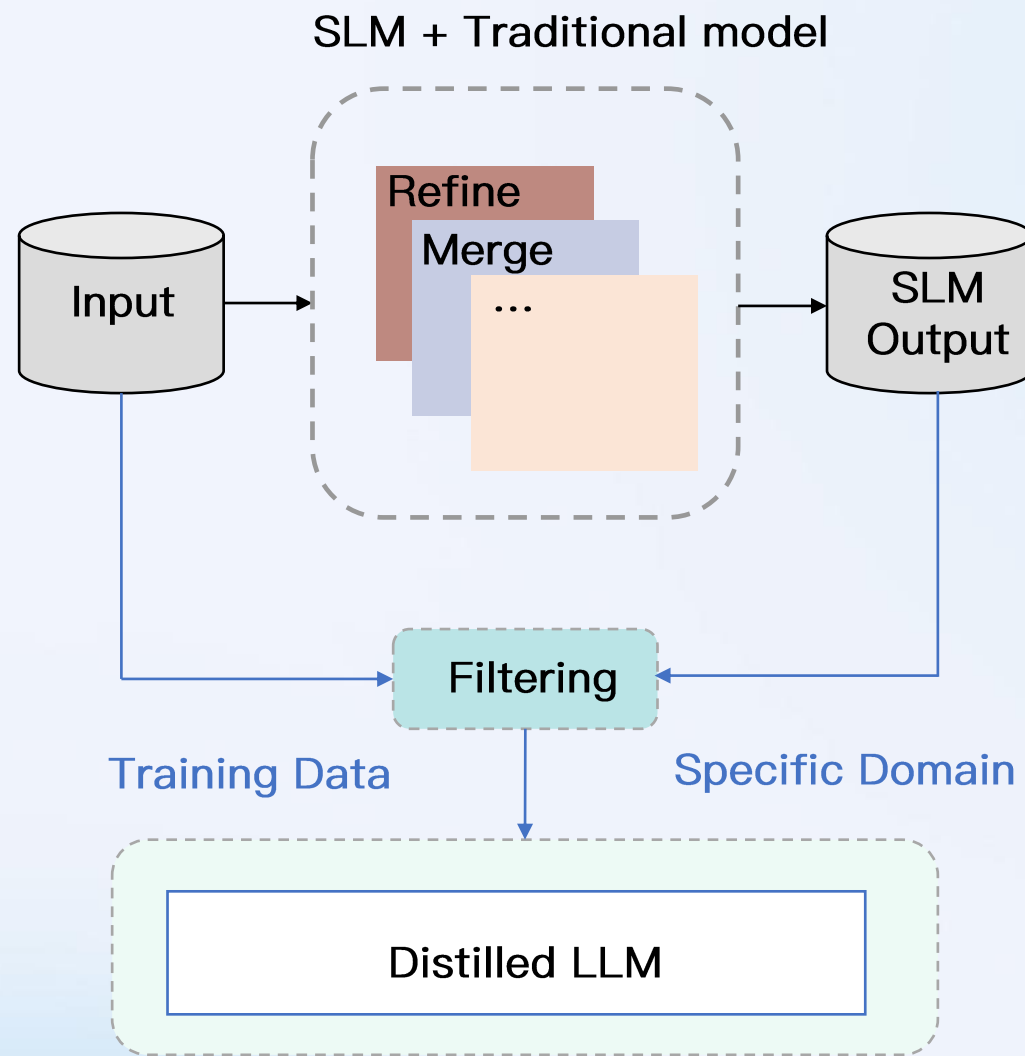
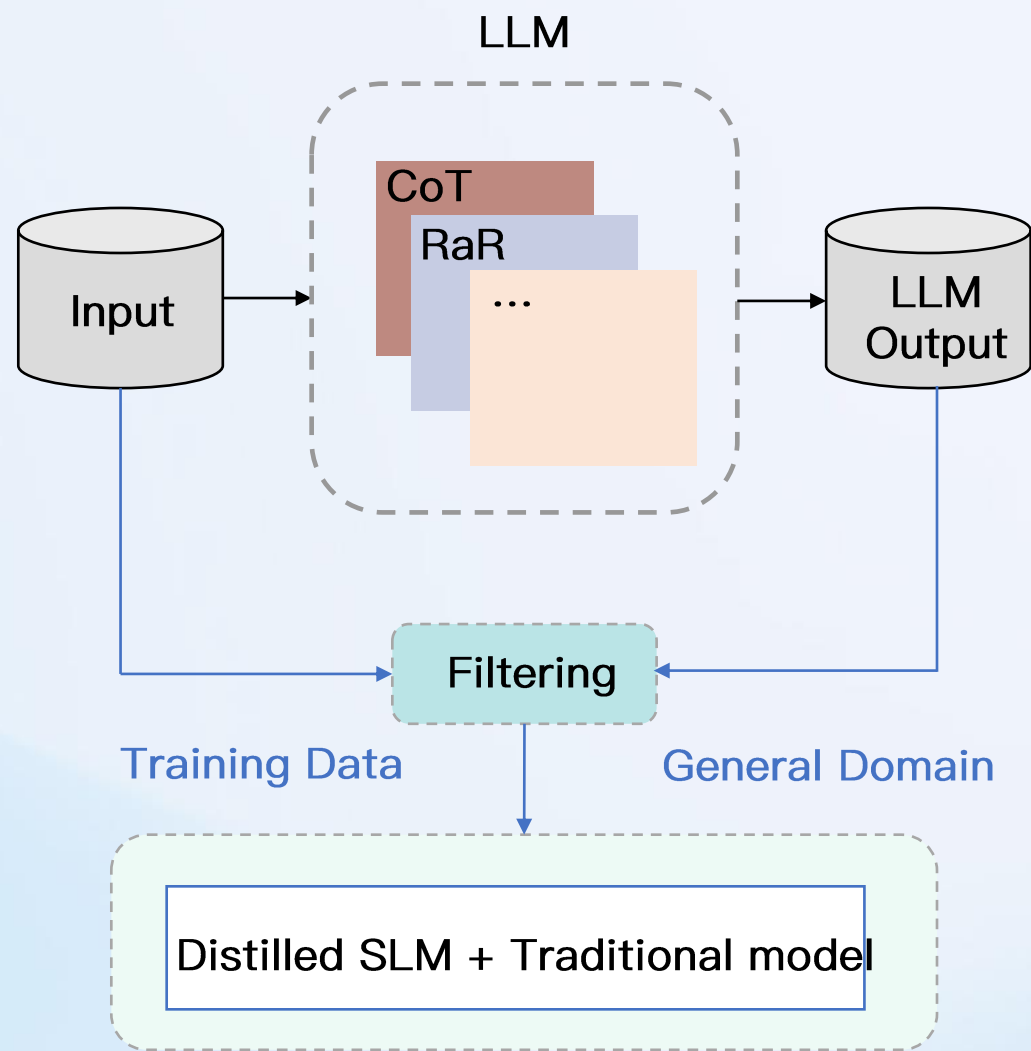
大小模型路由



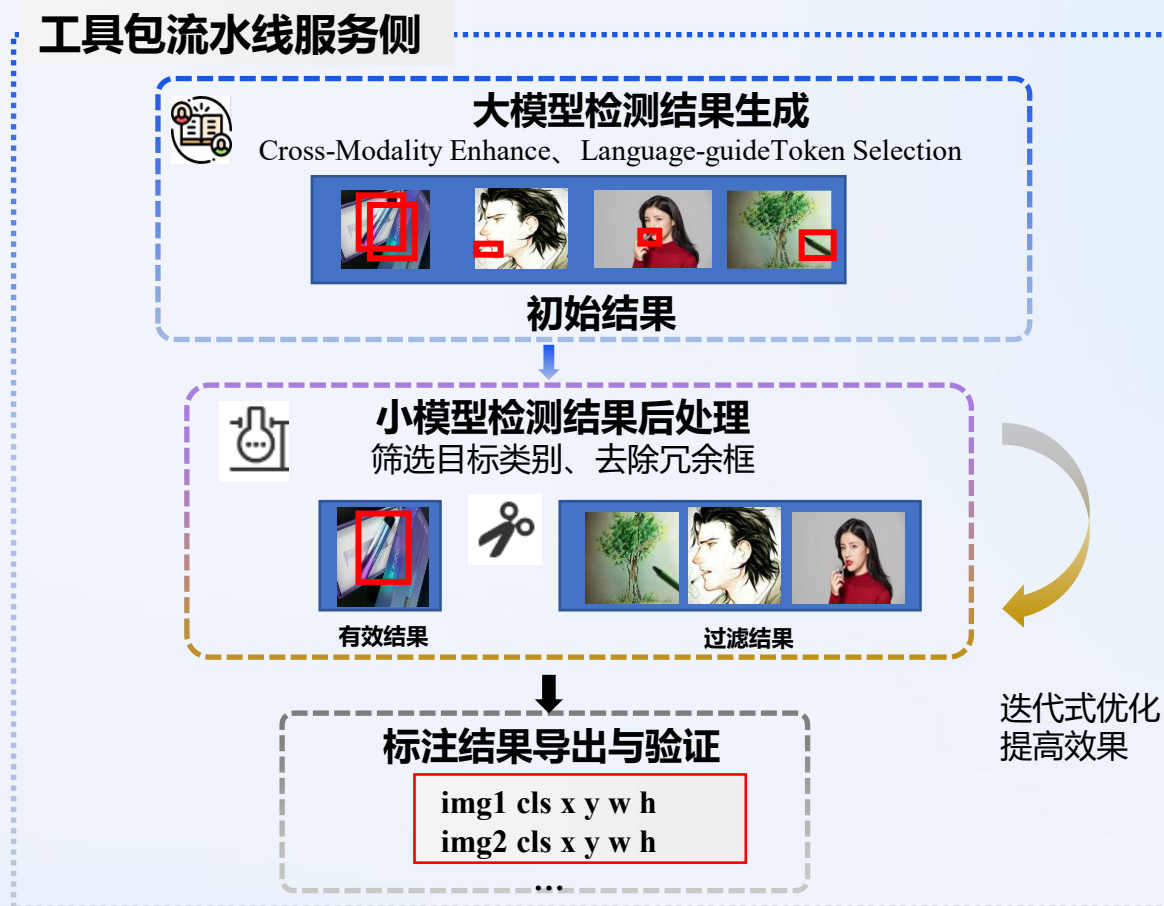
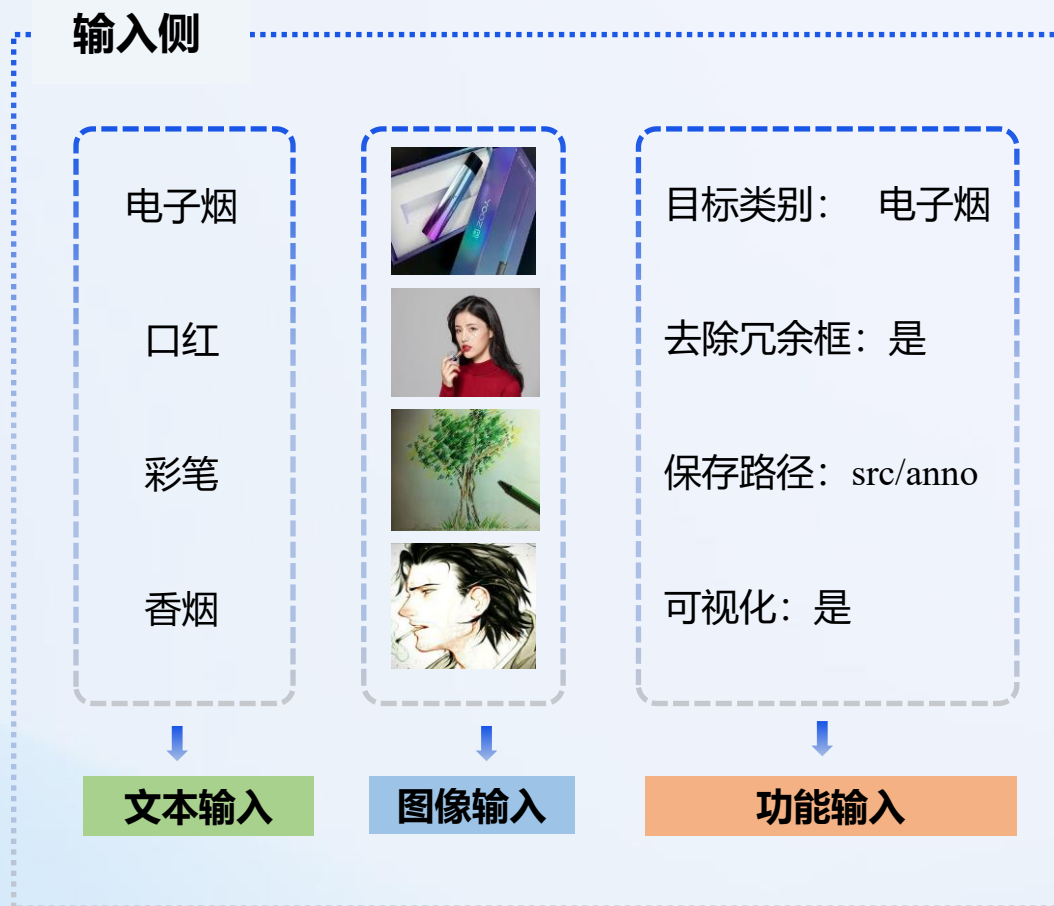
基于动态网络的大小模型路由



大小模型交互蒸馏

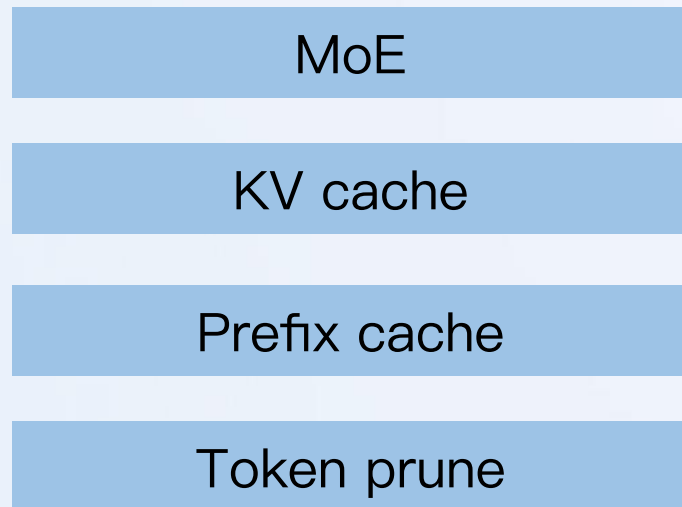


基于大小模型协同的数据挖掘



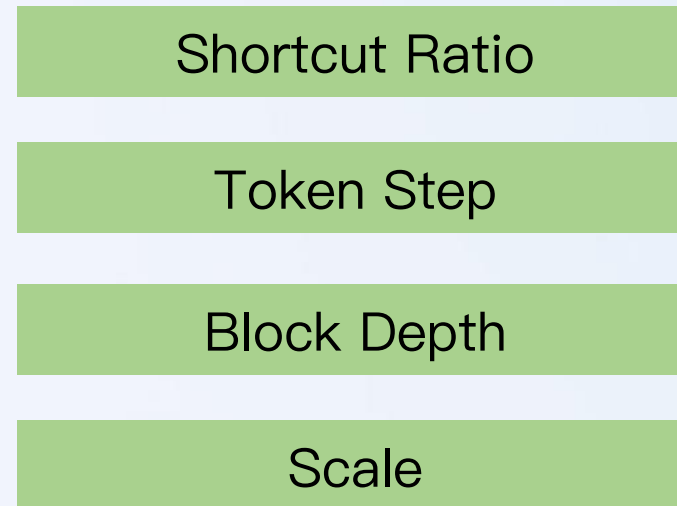
* 素材仅用于示意说明。

大小模型协同系统性能优化



大模型加速

Speed up



大小模型协同

03

数字内容风控实践

数字内容风控场景挑战

数据分布广泛

AIGC生成方式多样

强对抗

对抗样本攻击

识别难度大

生成效果逼真



语义层次深

复杂、隐晦表达

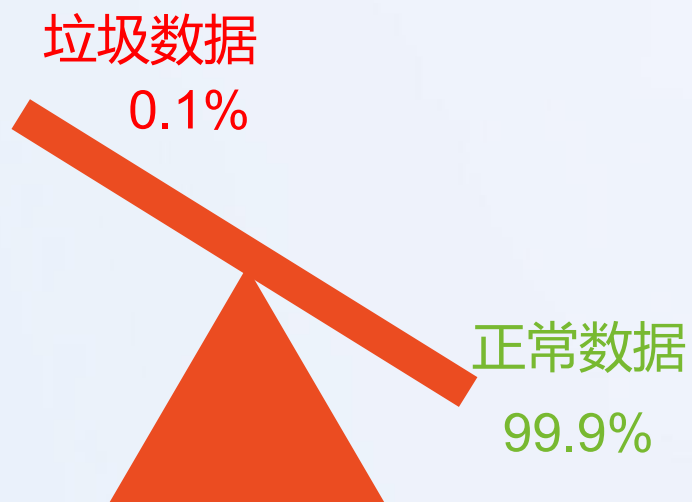
识别成本高

生成成本低

未知风险

AIGC技术发展快

内容风控场景的特点

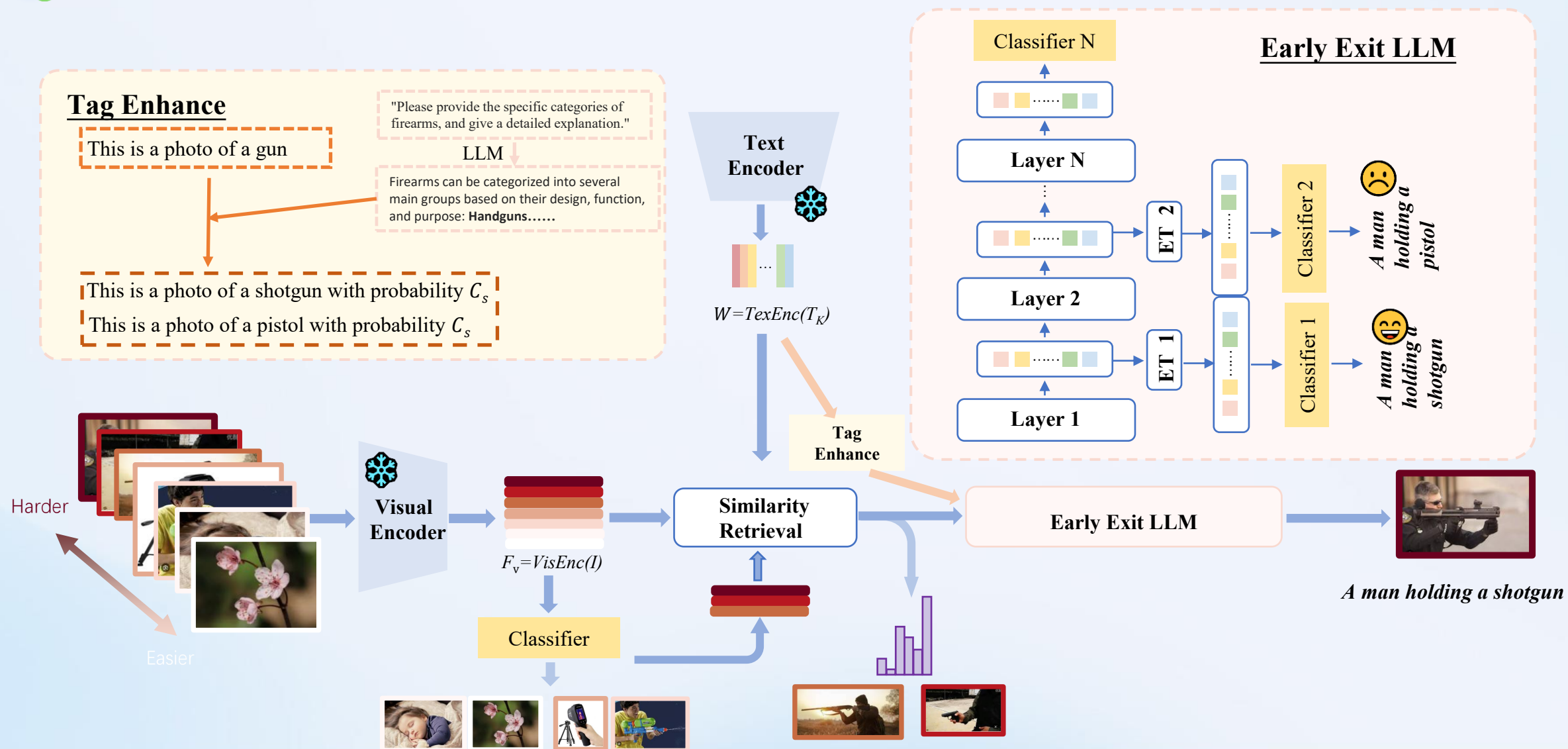


极度的数据不平衡

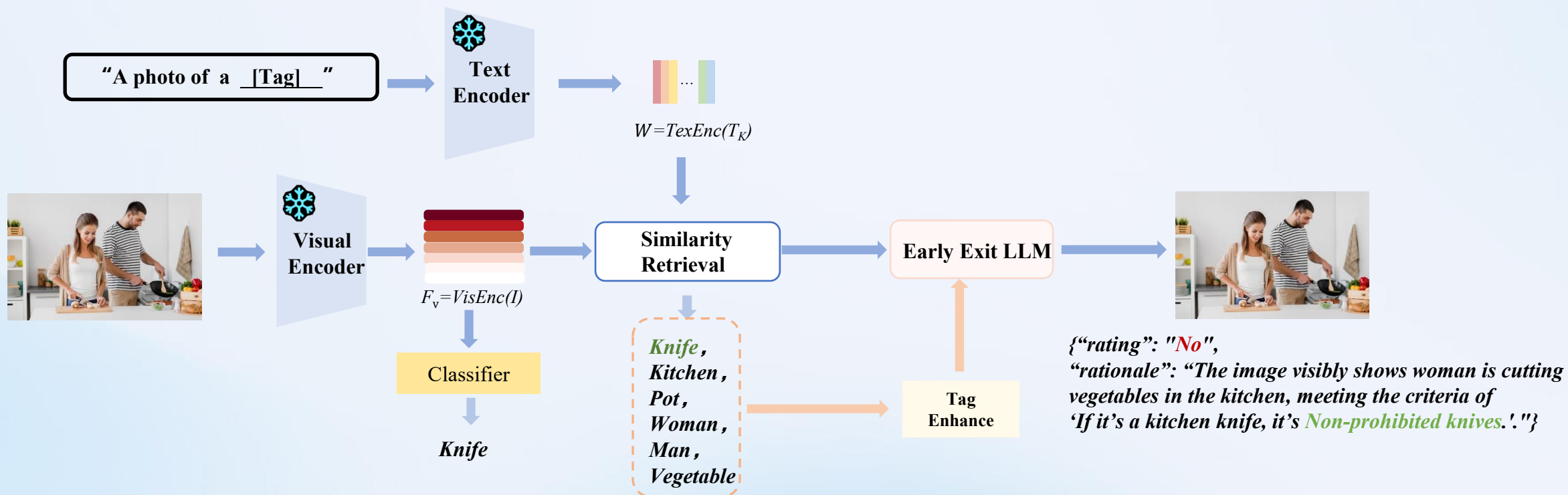
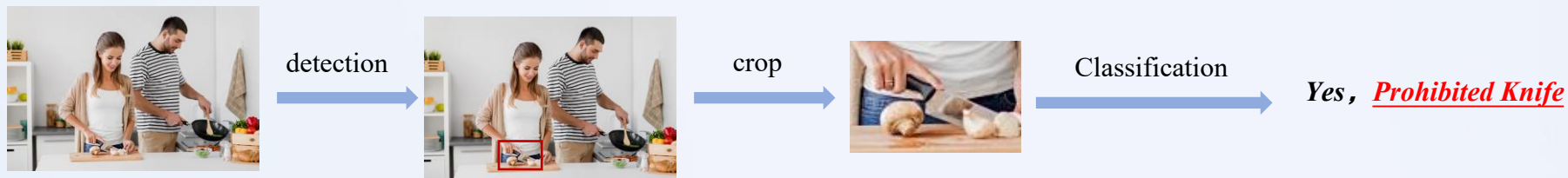
涉黄图片	暴恐图片	违禁图片	敏感图片	广告图片	恶心图片	涉价值观图片
男女下体	暴恐人物	违禁人物	敏感人物	二维码	内脏器官	抽烟
口交/性行为	暴恐旗帜	违禁图集	敏感旗帜	LOGO检测	疾病表征	喝酒
胸臀低俗	暴恐标识	违禁物品	敏感事件	商业推广	密集恐惧	纹身
吐舌摸胸	暴恐图集	违禁货币	敏感场景	广告刷量	腐烂食物	吃播
露沟/背/腿/足/裆	燃烧爆炸	公职服饰	中国地图	广告弹窗	排泄物	竖中指
情趣用品	打砸抢烧	宗教服饰	落马官员	医疗/迷信	恶心动物	劣迹艺人
性分泌物	分裂国家	特殊标识	英雄烈士	诱导欺诈	人类动物尸体	拜金炫富
儿童色情	恐怖主义	烛火祭奠	邪教迷信	极限用语	性器官特写	腐文化
卡通色情	极端宗教主义	骷髅灵异	热点敏感专项	法律风险	车祸	封建迷信
...	...	...	...	...	...	...

标签体系复杂

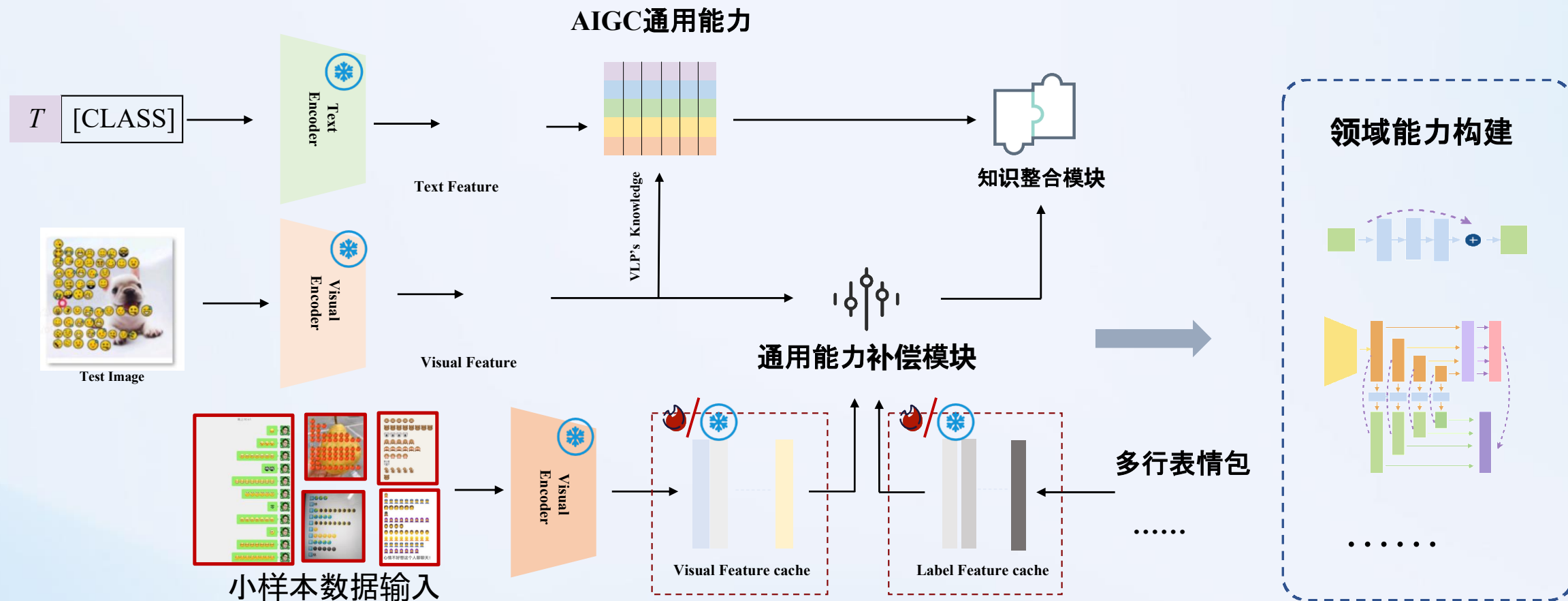
基于大小模型协同的数字内容风控方案



复杂风险治理



● 新型风险识别



● 风险处理实时性



标签体系
自动构建

['Active Wear Underwear', 'Bodycon Swimsuit', 'Fishnet Lingerie', 'Athletic Swimwear', 'Beach Cover Up', 'Beachwear Swimsuit', 'Bra and Panty Set', 'Maid Outfit', 'Tattoo Print Bodysuit', 'Anime Swimwear', 'Elegant Nightwear', 'Elegant Bridal Gown', 'Royal Lingerie Set', 'Bodybuilding Briefs', 'Fitness Bra', 'Athletic Lingerie', 'Seductive Hosiery', 'Bralette Top', 'Casual Intimates', 'Men's Swim Briefs', 'Fashion Intimates', 'Bodybuilding Attire', 'Glamour Swimwear', 'Cosplay Lingerie', 'Floral Pajamas', 'Fantasy Lingerie', 'Silk Pajamas', 'Glamorous Intimates', 'Luxury Intimates', 'Playful Lingerie', 'Backless Swimsuit', 'Cutout Swimsuit', 'Royal Bra Set', 'Party Outfit', 'Red Bra Set', 'Star Print Swimwear', 'Bunny Costume', 'Knitted Bralette', 'Bodycon Outfit', 'Swimwear Suit', 'Children's Swimwear', 'Anime Lingerie', 'Polka Dot Underwear', 'Red Swimwear', 'Bathwear Set', 'Silk Robe', 'Strapless Bra', 'Bodybuilding Gear', 'Casual Lingerie', 'Winter Swimwear', 'Fitness Wear', 'Fitness Lingerie']

细粒度标
签识别

N-shot
Training-free

实践效果

基于大小模型协同的内容风控实践

准确率

- 85%+ (5-shot training-free)
- 95%+(微调)

效果

研发周期

- 大部分新增能力研发周期由周级别缩短到日级别
- Case级别响应周期缩短到分钟级别

敏捷迭代

资源消耗

- 单个标签数据标注量下降到百量级
- 小模型短路比96%+。

资源消耗

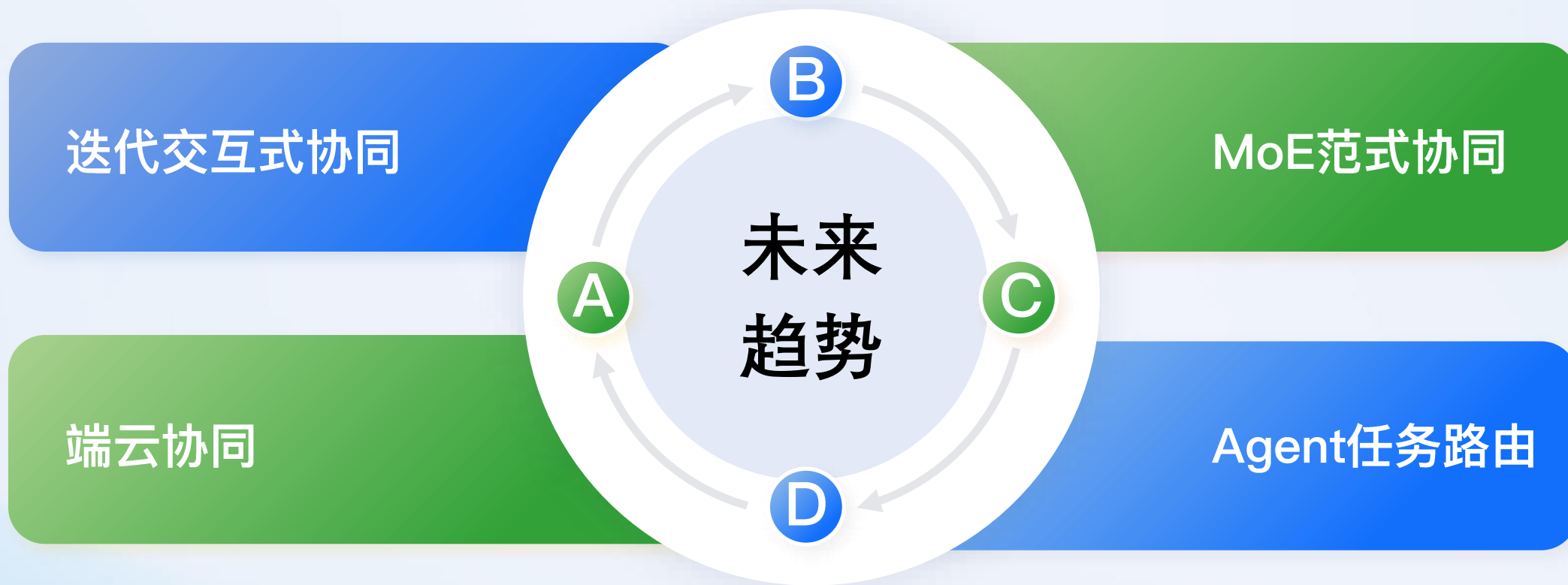
审核效率

- 人工审核效率提升70%

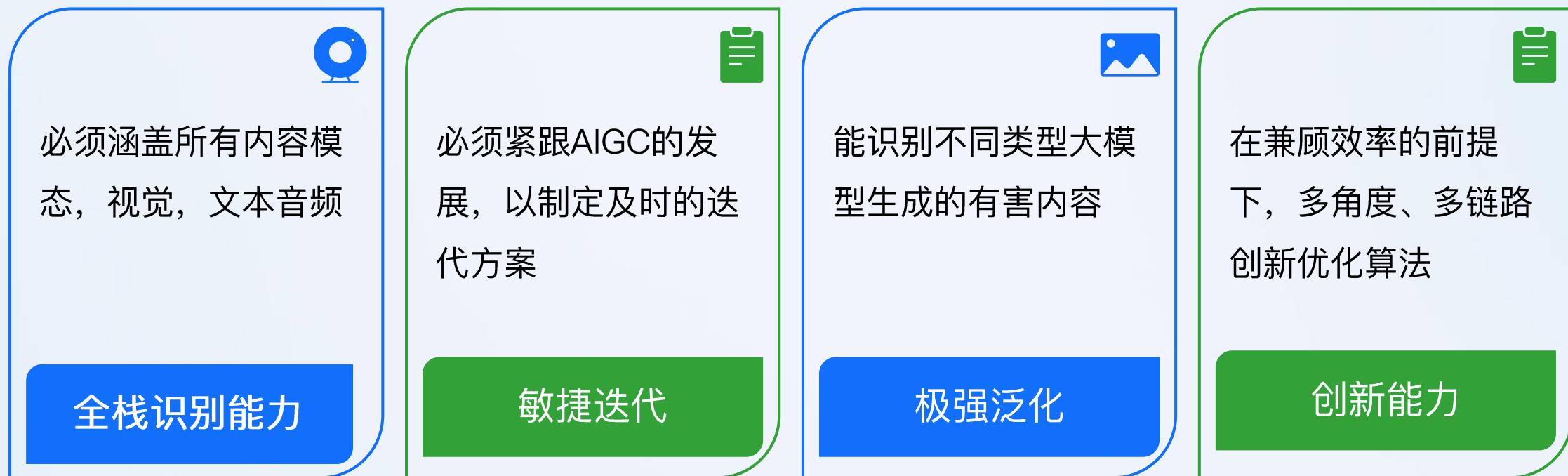
人工审核

04 未来展望

大小模型协同未来趋势



内容风控场景发展趋势



特征识别方案



结合通用知识

后知后觉



提前预判

内容感知



内容认知

📍 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

大模型正在重新定义软件

Large Language Model Is Redefining The Software

