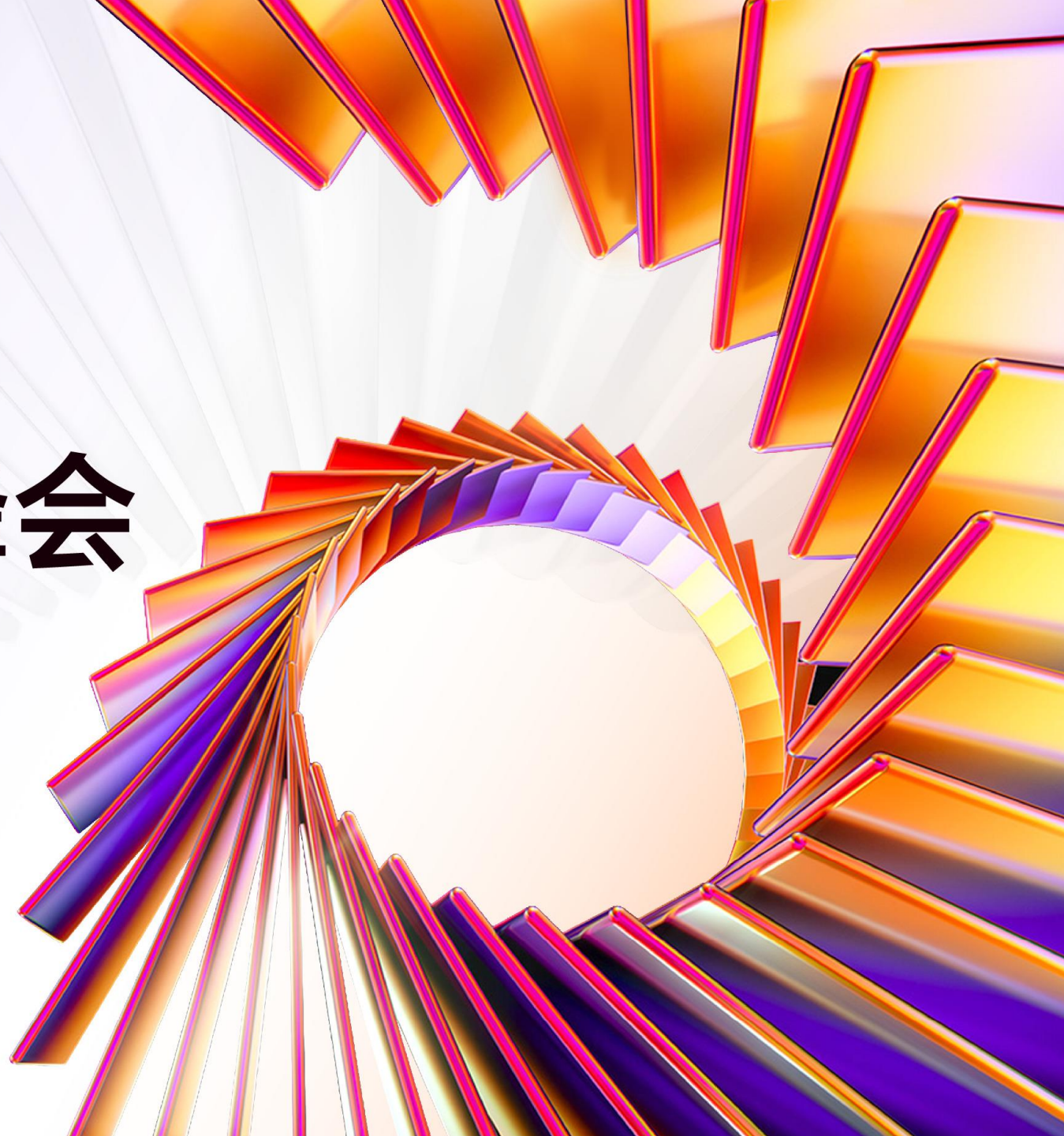
AiDD 6th 2025

2025 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发

05/23-24 | 上海站



AiDD 7th 2025

2025 AI+研发数字峰会

拥抱AI 重塑研发

AI+ Development Digital Summit

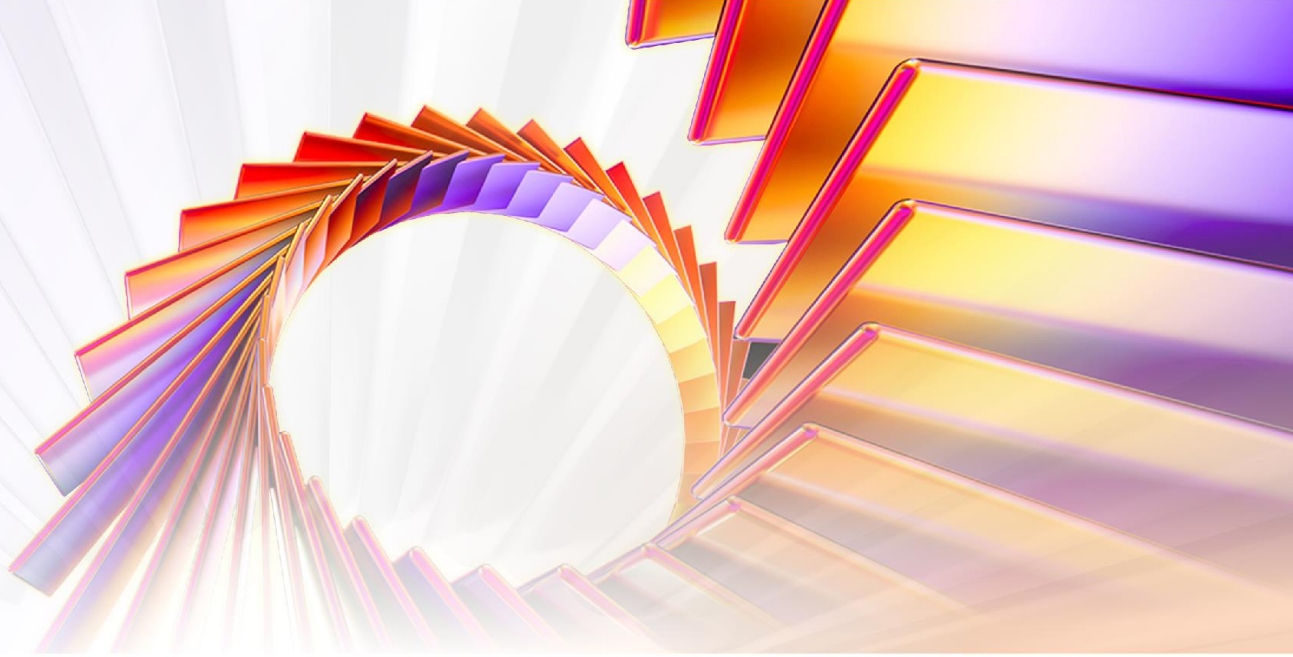
查看会议详情

下一站预告 08/08-09 | 北京站 11/14-15 | 深圳站

北京站论坛设置

- AI+ 金融业务创新
- 大模型安全与对齐
- 具身智能与机器人
- 大模型和 AI 应用评测
- 智能需求工程
- 大模型应用开发框架与实践
- 代码生成及其改进
- 智能存储与检索技术
- 智能体与研发效率工具
- 智能体经济（Agentic Economy）
- AI+ 新能源汽车
- 下一代知识工程
- AI 产品运营与出海策略
- 智能测试工具的开发与应用
- AI 前沿技术探索与实践

AiDD 6th 2025 | 05/23-24 | 上海站
2025 AI+ Development Digital Summit
AI+研发数字峰会
拥抱AI 重塑研发



抖音性能LLM分析体系建设 从智能诊断到决策推荐

姚凡、李文博 | 字节跳动

姚 凡

抖音测试开发专家

毕业于悉尼大学计算机专业，现就职于抖音专项测试团队，负责抖音体验平台的性能诊断、大模型应用等方向的开发工作。

李文博

抖音测试开发专家

毕业于悉尼大学计算机专业，毕业后即加入抖音，现就职于抖音基础体验团队，专注于客户端性能体验分析、测试效率优化等专项技术。

目 录

CONTENTS

PART 01

背景

性能优化的重要性

降低用户流失
加载速度直接影响用户留存率

提高品牌信任
性能问题会导致用户评分下降

驱动营收增长
加载速度与用户留存率直接相关

性能优化挑战：从表现层问题到优化落地

1. 性能问题归因复杂

- **分析数据繁多**：**性能**问题涉及客户端、网络、服务端等多个维度，数据源多，数据量大。
- **分析**门槛高：需熟练使用不同端/**数据源的多个分析工具**，如 Android Profiler、Instruments；**需熟悉**编程语言、技术实现原理等。

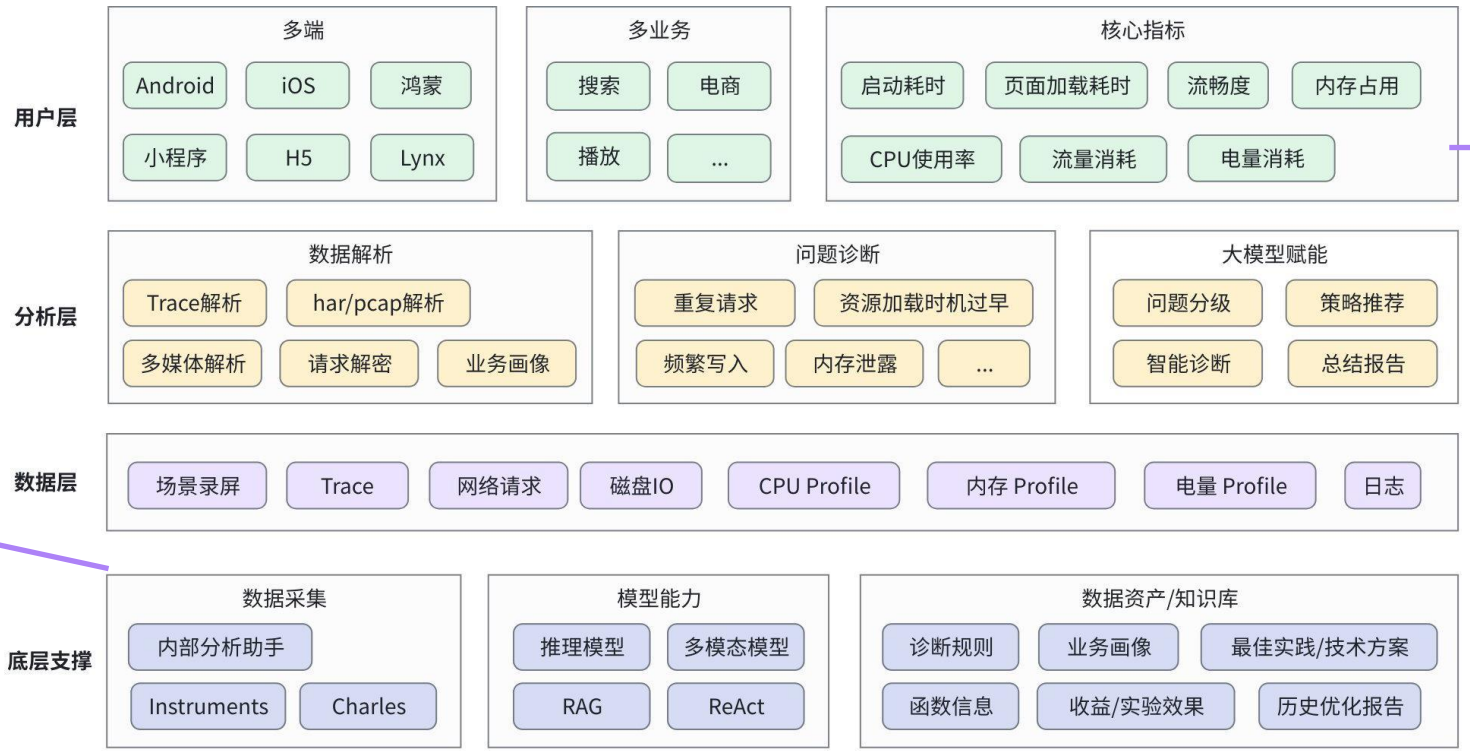
2. 优化方案选型复杂

- **决策困**难：优化路径多，效果不明确，导致试错成本高，难**以快速确定适合**业务的最优方案。
- 专家依赖：优化策略依赖少数资深专家，存在经验主观性、**知**识孤岛与人才瓶颈问题。

PART 02

整体解决方案

性能分析整体架构



☹ 分析数据繁多？

多数据同时采集与联合分析

☹ 分析门槛高？

用户层问题的自动分析能力

☹ 专家依赖？

经验知识沉淀

AI赋能

AI

- **数据整合与理解能力**
  - **多模**态模型：图片/视频理解
  - **代**码理解模型：代码/**函数理解**
- RAG：
  - **相似**场景的历史经验复用。
- SFT：
  - 专家知识应用于LLM中。

赋能

诊断分析

策略推荐

提升

覆盖率

准召率

采纳率

下面将分两个章节重点介绍AI在诊断分析与策略推荐中的难点与方案。

PART 03

核心模块 - 智能诊断

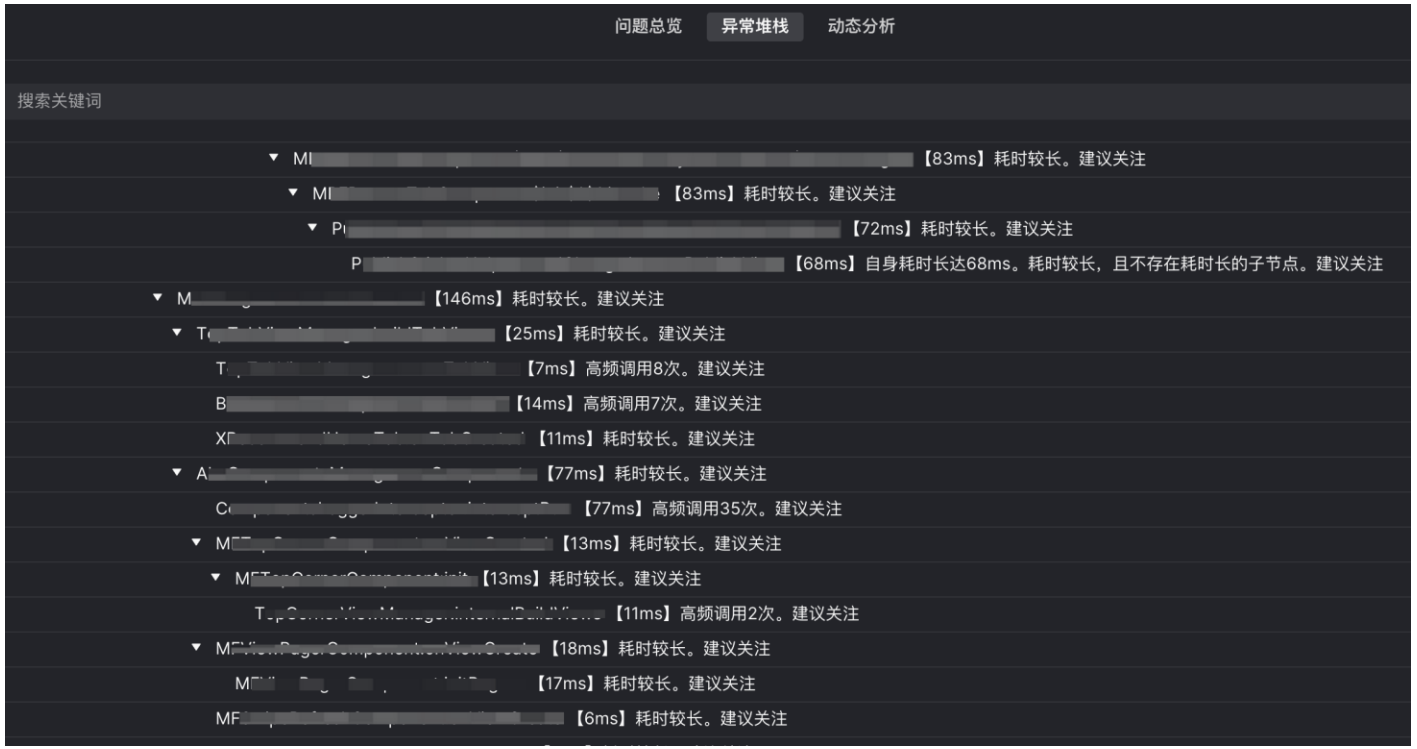
Trace分析平台现状

- 可视化分析
- 异常堆栈
- 规则分析

🙂 录屏+Trace同时采集，协同分析

Trace分析平台现状

- 可视化分析
- **异常堆栈**
- 规则分析



🙂 **自上而下**对慢函数/**高**频函数**的异常堆**栈定位

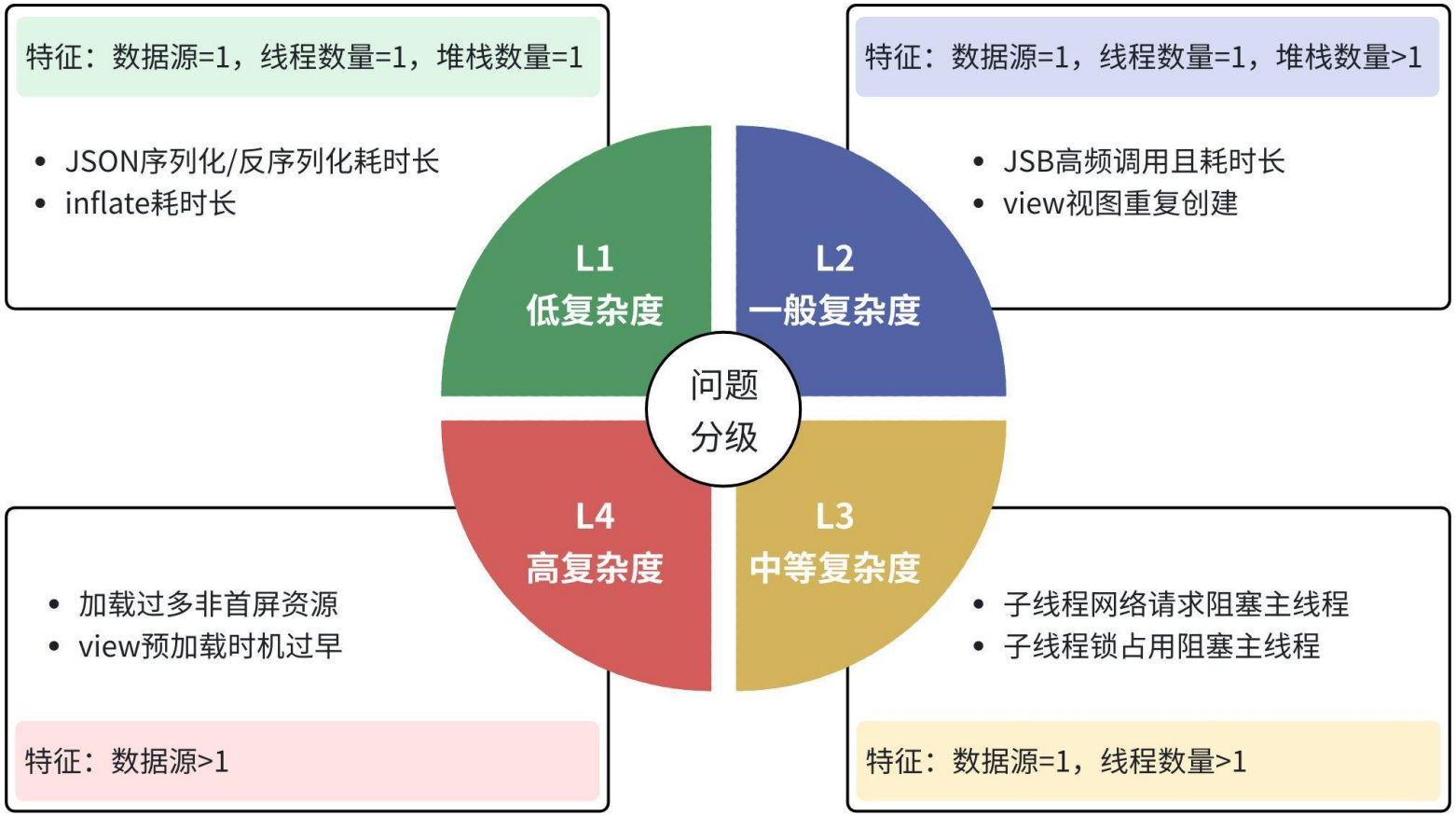
😕 **未与**场景结合，**无法确定**问题影响的具体阶段

Trace分析平台现状

- 可视化分析
- 异常堆栈
- 规则分析

🙂 支持组件耗时长、无效UI更新等L1/L2的部分规则

😕 但不支持L2+的规则



AI应用的诉求

很多规则无法靠硬编码实现，需要借助AI的语义理解、推理能力、图像理解等能力。

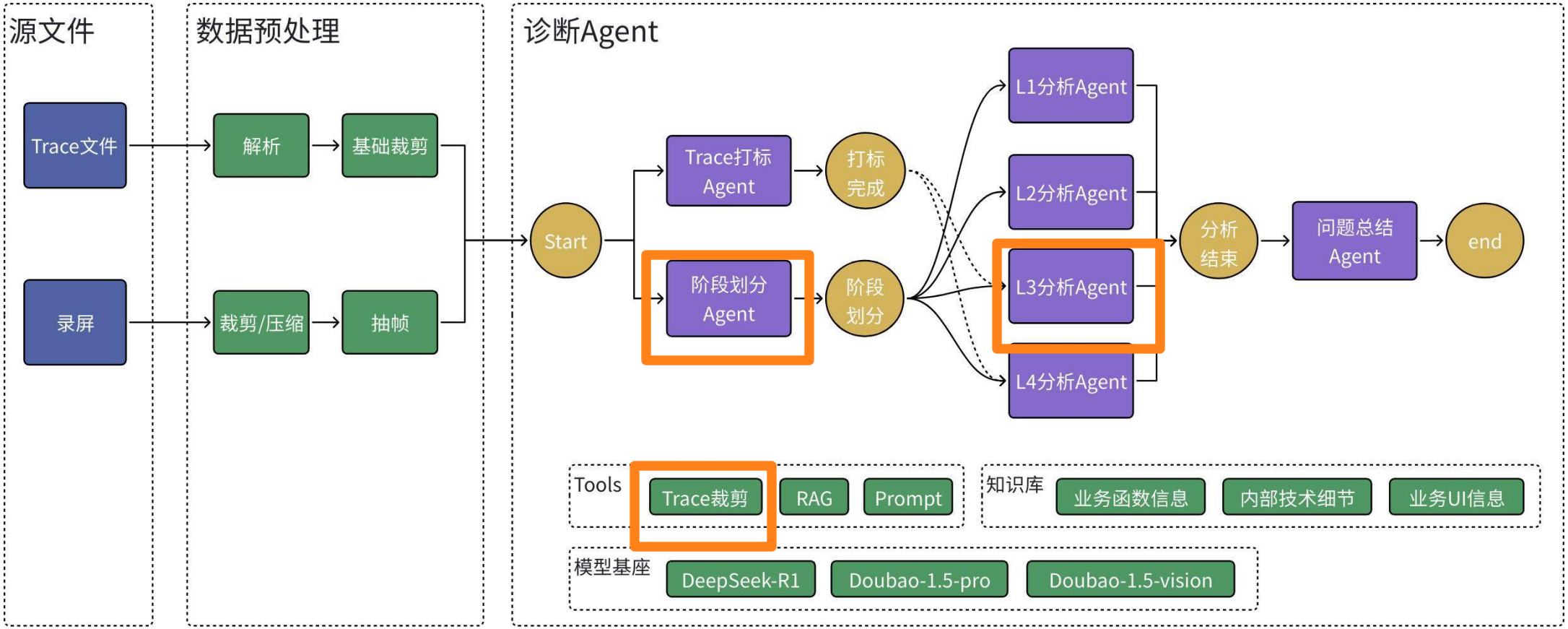
如何判断多个堆栈在做同一件事情，需要聚合分析？

如何判断资源的创建/渲染时机是否合理？

如何判断主线程与子线程之间的调度关系？

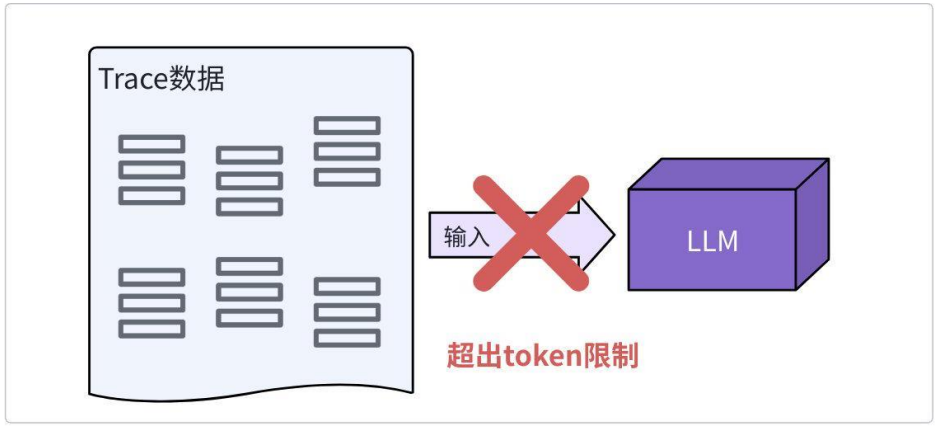
如何判断诊断出的问题具体影响的是哪个阶段？

AI智能诊断整体流程

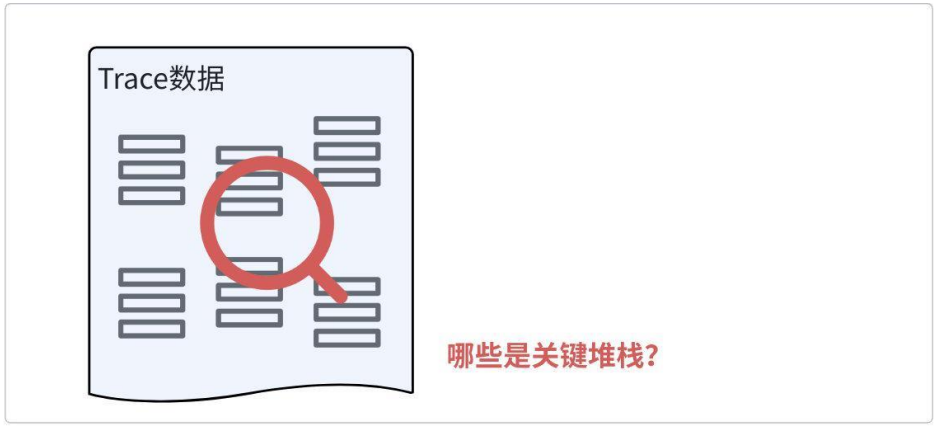

源文件
数据预处理
诊断Agent
Trace文件
录屏
解析
基础裁剪
裁剪/压缩
抽帧
Start
Trace打标 Agent
打标完成
阶段划分 Agent
阶段划分
L1分析Agent
L2分析Agent
L3分析Agent
L4分析Agent
分析结束
问题总结 Agent
end
Tools
Trace裁剪
RAG
Prompt
知识库
业务函数信息
内部技术细节
业务UI信息
模型基座
DeepSeek-R1
Doubao-1.5-pro
Doubao-1.5-vision

Trace裁剪 - 为什么要裁剪？

Trace数据量过大， 一个完整Trace文件在几百M左右。



单个堆栈的数据量就已超出LLM**上下文**token**限制**



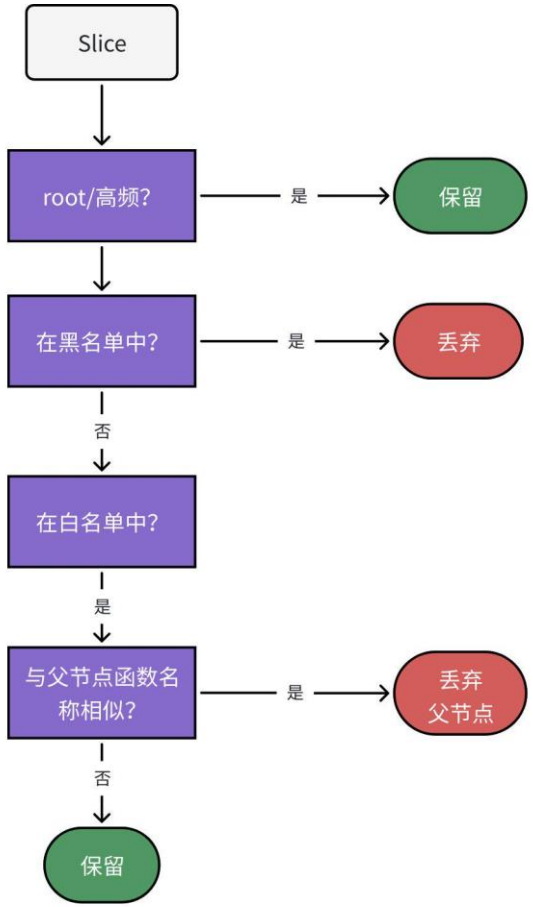
对大量数据进行分析可能造成准确率降低、处理耗时过久

Trace裁剪 - 裁剪方式

- 线程裁剪：**主**线程/**子**线程
- **堆**栈裁剪：丢弃 root 节点耗时小于阈值的堆栈
- 单堆栈slice**裁剪**
  - **高**频聚合
  - **最小耗**时阈值
  - **黑白名**单
  - **相似函数/相似子**树

可根据不同的场景，选择不同的裁剪方式

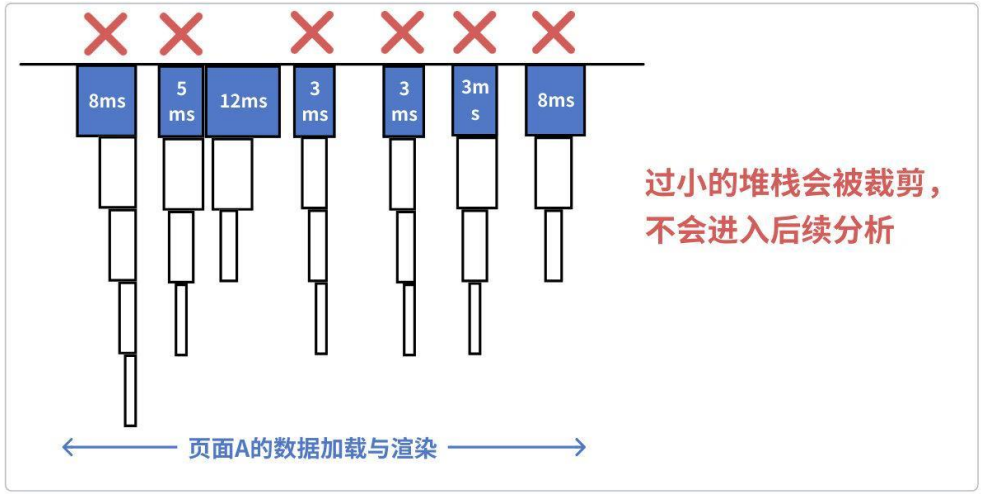
。



单堆栈slice**裁剪**

阶段划分Agent - 为什么要划分阶段？

合理的阶段划分可以**提高**问题发现的**准召率与采纳率**。

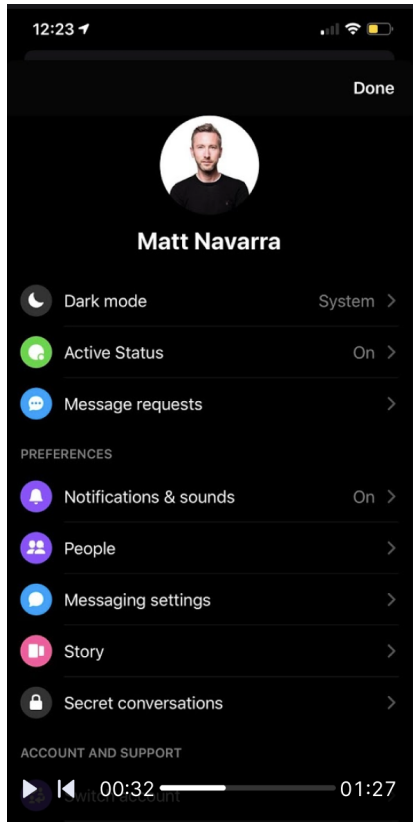


存在多个小堆栈在处理同一任务且整体耗时较长的情况，需聚合在一起分析

单纯的问题罗列，**不**绑定阶段场景的话，**用**户难以理解与消费

阶段划分Agent - 录屏场景识别

录屏理解 + Trace裁剪 => 阶段划分



录屏

Doubao-1.5-vision

prompt

```
{
  "appRunDescription": "用户在用户主页，点击Story后，页面跳转到Story详情页。",
  "pagesDetail": [
        {
          "name": "用户主页",
          "detail": "页面顶部有用户头像和名称‘Matt Navarra’，右上方有Done按钮。下方有Dark mode、Activite Status、xxx"
        },
        {
          "name": "Story详情页",
          "detail": "页面顶部显示xxx"
        }
  ]
}
```

录屏信息

阶段划分Agent - Trace裁剪

录屏理解 + **Trace裁剪** => 阶段划分

原Trace → 裁剪+ 格式转换 →

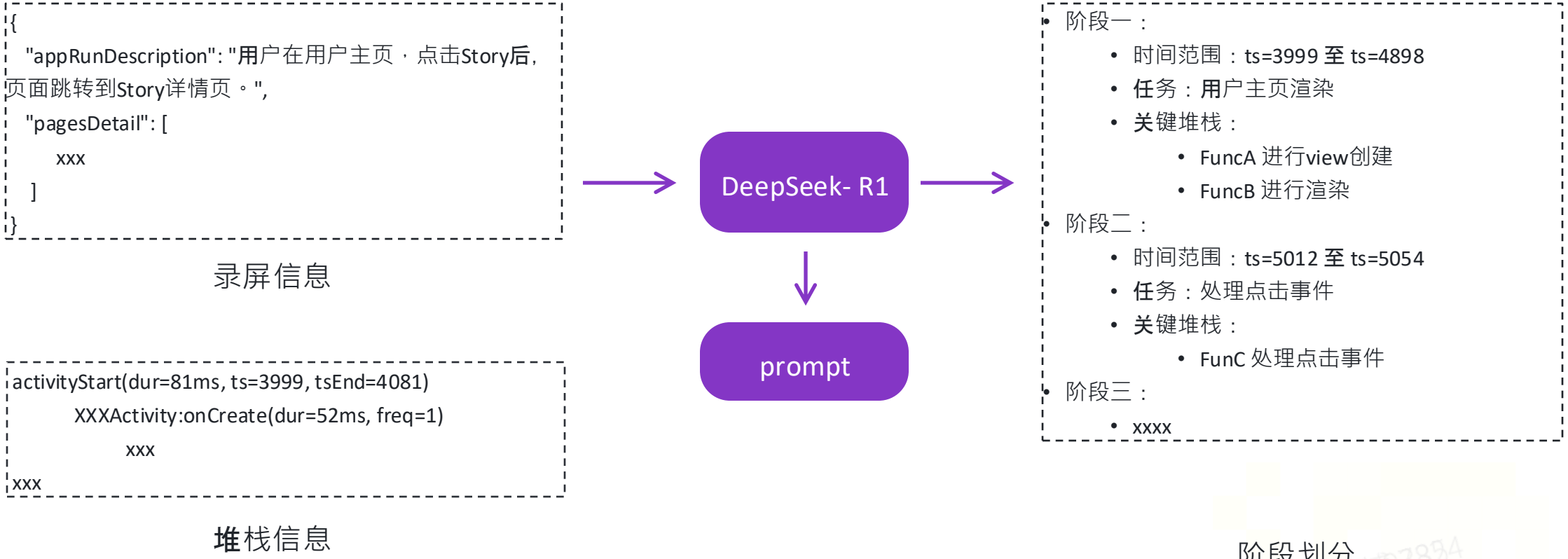
```
activityStart(dur=81ms, freq=1, ts=3999, tsEnd=4081)
        XXXActivity:onCreate(dur=52ms, freq=1)
                xxx
activityResume(dur=19ms, freq=1, ts=4135, tsEnd=4154)
        binder transaction(dur=4ms, freq=7)
        XXXActivity:onResume(dur=6ms, freq=1)
Choreographer#doFrame(dur=39ms, freq=1, ts=3956, tsEnd=3996)
        traversal(dur=39ms, freq=1)
                measure(dur=27ms, freq=1)
                        xxx
                layout(dur=7ms, freq=1)
                draw(dur=3ms, freq=1)
xxx
xxx
```

堆栈信息

- **裁剪方式**：**保留主**线程，root耗时 > 5ms**的堆**栈，并进行重度精简
- **格式**转换，提取函数名称、耗时、频次、开始/结束时间信息

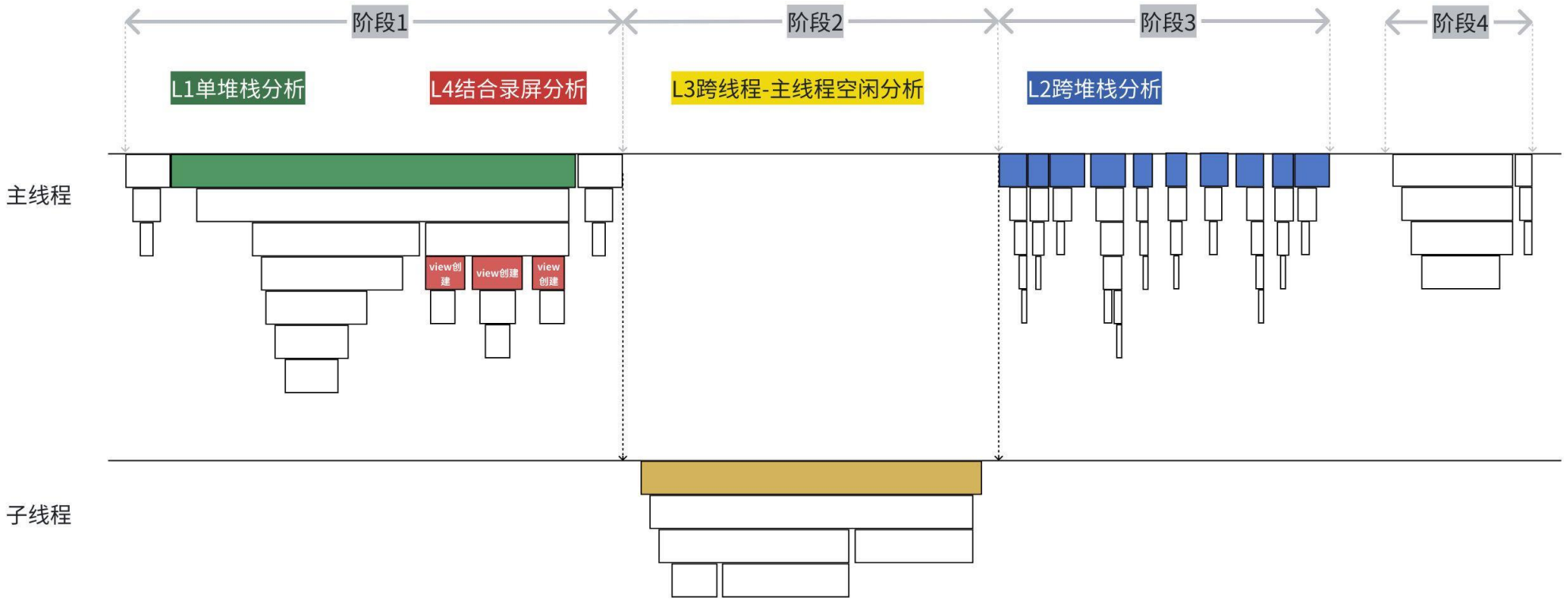
阶段划分Agent - 阶段拆分

录屏理解 + Trace裁剪 => 阶段划分

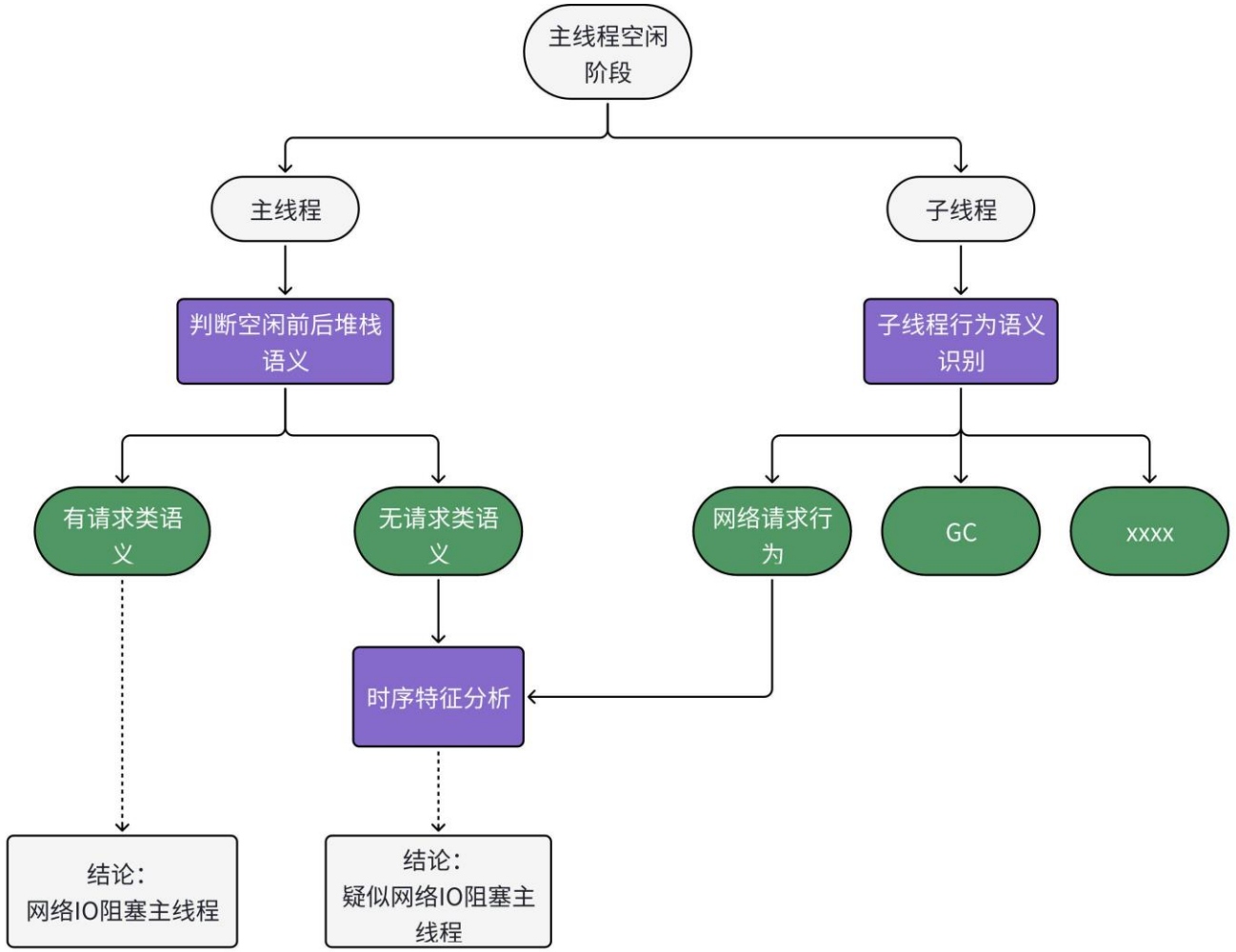


阶段划分Agent - 分阶段分析

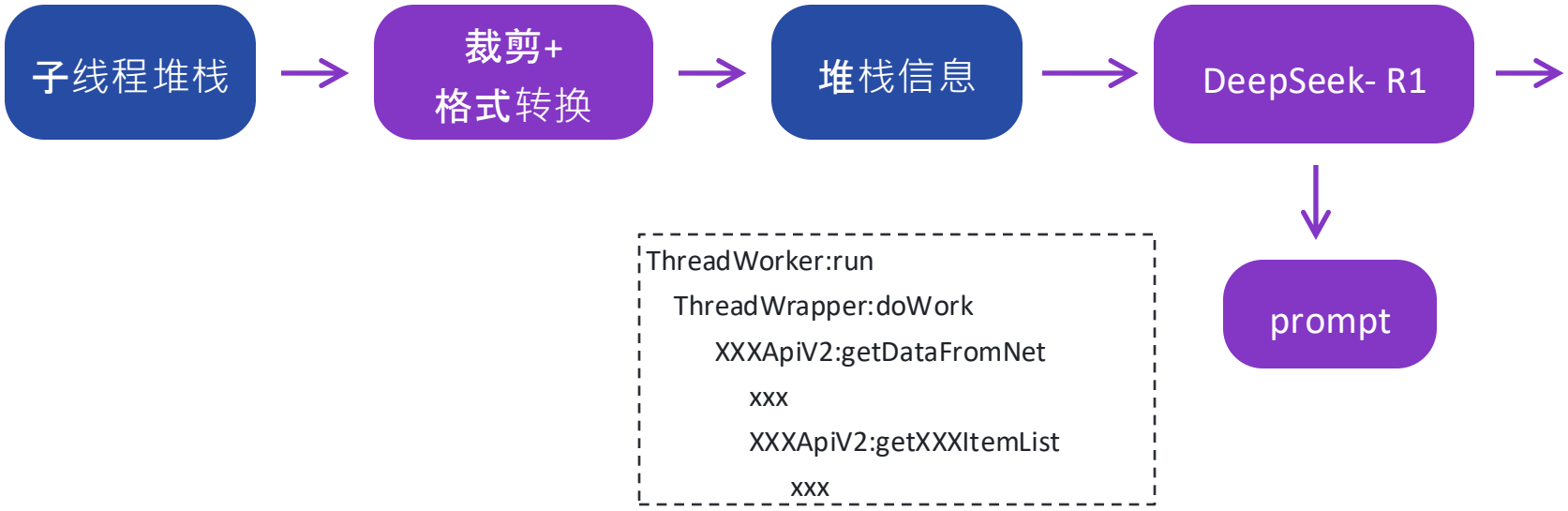
阶段划分完成后，阈值初筛出需要分析的堆栈，再根据堆栈与场景特征分别进行分析。



L3分析Agent - 主线程空闲分析流程



L3分析Agent - 子线程行为语义识别



语义识别

PART 04

核心模块 - 策略推荐

体验优化的核心挑战

问题难定位

- 用户在不同场景下遇到的各类体验缺陷分布于多渠道数据中，缺乏归纳和映射，导致定位和复现困难。
- 传统问题排查多依赖研发个人经验，从表现层问题到技术根因的排查链路长

策略难复用

- 性能优化方案分散于各个团队，优化经验非标准化，难以有效传递，重复造轮子，试错成本高昂。
- 具体业务的优化策略通用性不足，难以抽象出普适优化思路，跨场景迁移应用难。

优化难落地

- 性能优化涉及技术领域广泛，解决思路多样、优化手段接入路径复杂且不明确。
- 缺乏清晰的决策支持与收益评估，导致优化决策难以获得业务方认可与投入。

AI赋能解决思路

智能Agent：多源聚合精准推荐

- 结合大模型Agent的任务规划能力，聚合日志、监控、反馈等多渠道表现层数据，联通知识库和归因技术自动定位体验问题。使用专家系统多角度提升问题定位效率和准确性，降低对人工经验的依赖。

体验知识库：打破信息孤岛

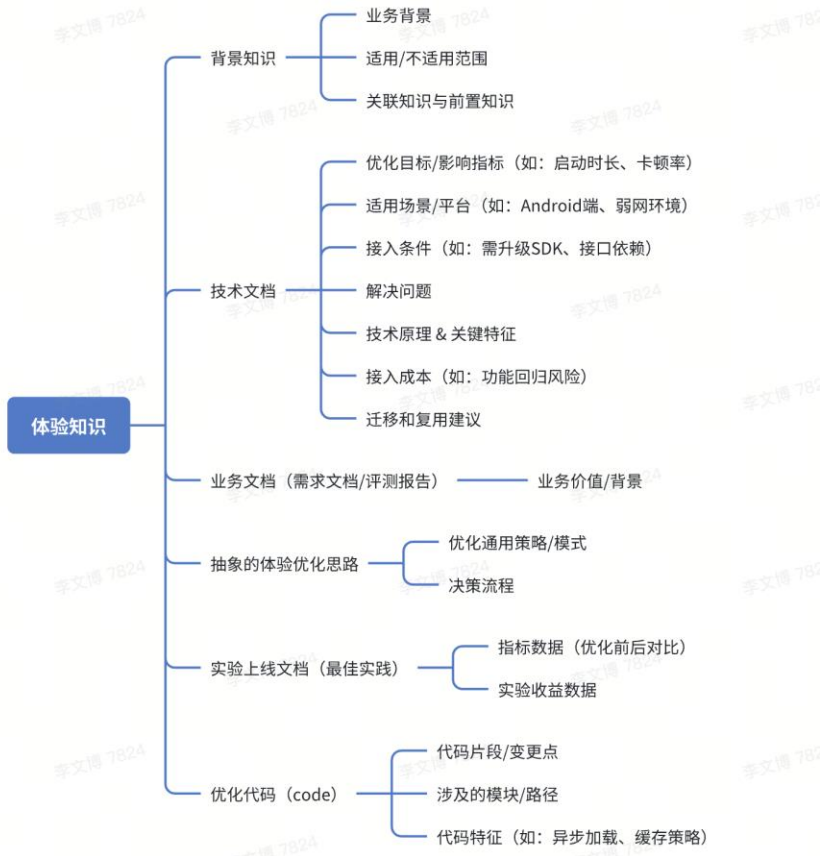
- 使用LLM能力自动抽取和结构化沉淀优化方案、最佳实践、实验收益等多源信息，构建统一知识库，实现跨团队、跨业务实现知识共享和高效检索，支撑科学决策与快速复用。

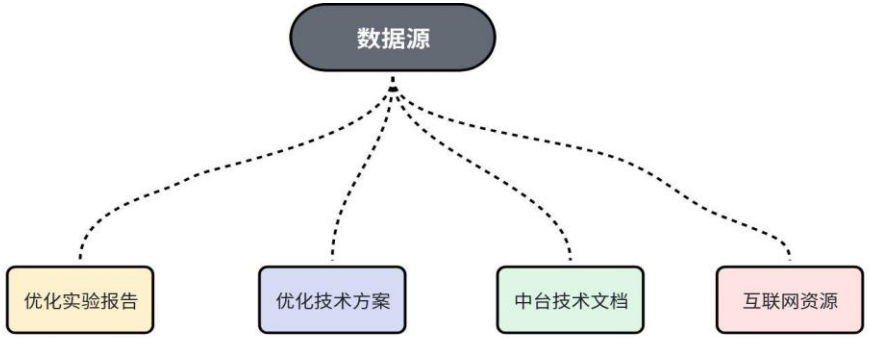
抽象推理：策略泛化创新

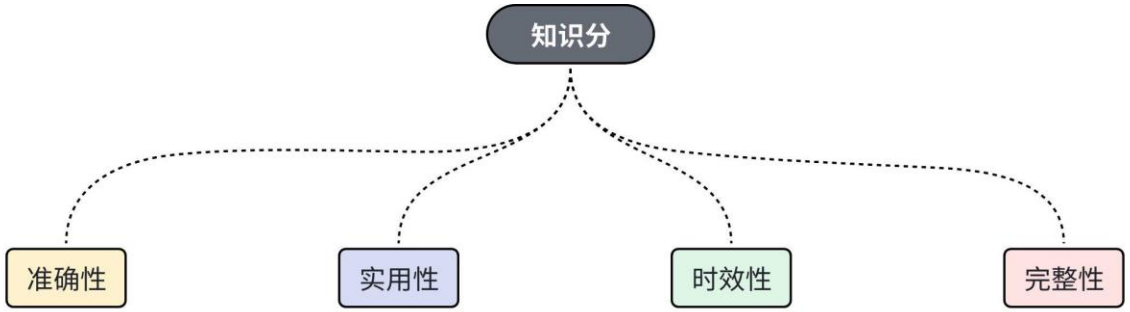
- LLM具备优化策略的抽象归纳和迁移能力。从历史案例中总结通用方法，不仅提升优化策略的适用性，还能发现并推荐新的创新优化思路。

体验知识库构建


文档输入
判断数据源
判断信息类型
摘要提取
标签推理
交叉验证
去重入库


体验知识
背景知识
业务背景
适用/不适用范围
关联知识与前置知识
技术文档
优化目标/影响指标（如：启动时长、卡顿率）
适用场景/平台（如：Android端、弱网环境）
接入条件（如：需升级SDK、接口依赖）
解决问题
技术原理 & 关键特征
接入成本（如：功能回归风险）
迁移和复用建议
业务文档（需求文档/评测报告）
业务价值/背景
抽象的体验优化思路
优化通用策略/模式
决策流程
实验上线文档（最佳实践）
指标数据（优化前后对比）
实验收益数据
优化代码（code）
代码片段/变更点
涉及的模块/路径
代码特征（如：异步加载、缓存策略）


数据源
优化实验报告
优化技术方案
中台技术文档
互联网资源


知识分
准确性
实用性
时效性
完整性

体验知识库应用

1. 事实性提问

2. 技术垂类策略检索 & 优化手册生成

3. 产品表现层问题解决方案推荐

体验优化推荐Agent设计

专家系统问题分析

- **技术实现：**基于多领域专家系统（网络、客户端、服务端、多媒体）构建智能诊断框架，结合RAG技术从知识库精准召回优化方案。
- **价值亮点：**自动化归因与方案推荐，显著降低性能优化试错成本，提升用户体验与业务指标。

多源知识图谱融合

- **技术实现**：根据技术方案、实验报告、最佳实践等多源知识构建知识库。通过知识图谱与RAG技术，实现跨文档、跨业务的深度语义检索与召回。
- **价值亮点**：为个性化优化方案提供精确支撑，实现全景知识的深度聚合与智能推理。

业务画像驱动策略推荐

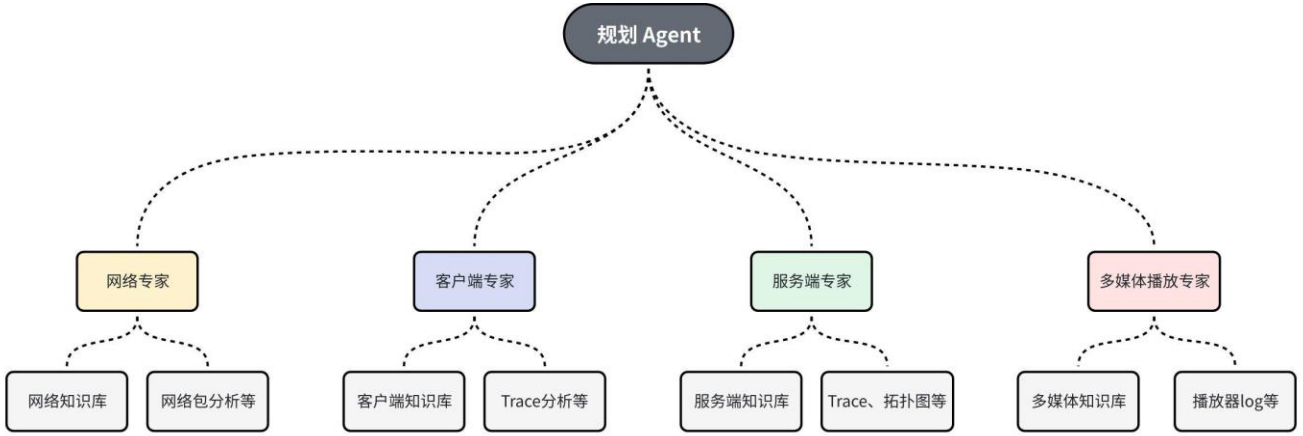
- **技术实现**：结合业务画像主动理解业务场景和用户需求。智能发现推理链路中的信息缺口，自动补全关键信息，保障方案完整性和准确性。
- **价值亮点**：显著提升推荐系统的可解释性、信任度及准召率。

智能推理和方案生成

- **技术实现：**结合CoT推理与专家知识融合，自动生成结构化、可落地的优化方案。具备策略泛化、举一反三与创新能力，支持跨场景的最佳实践迁移。
- **价值亮点**：方案推理过程透明可溯源，提升方案创新性、适应性与落地效率。

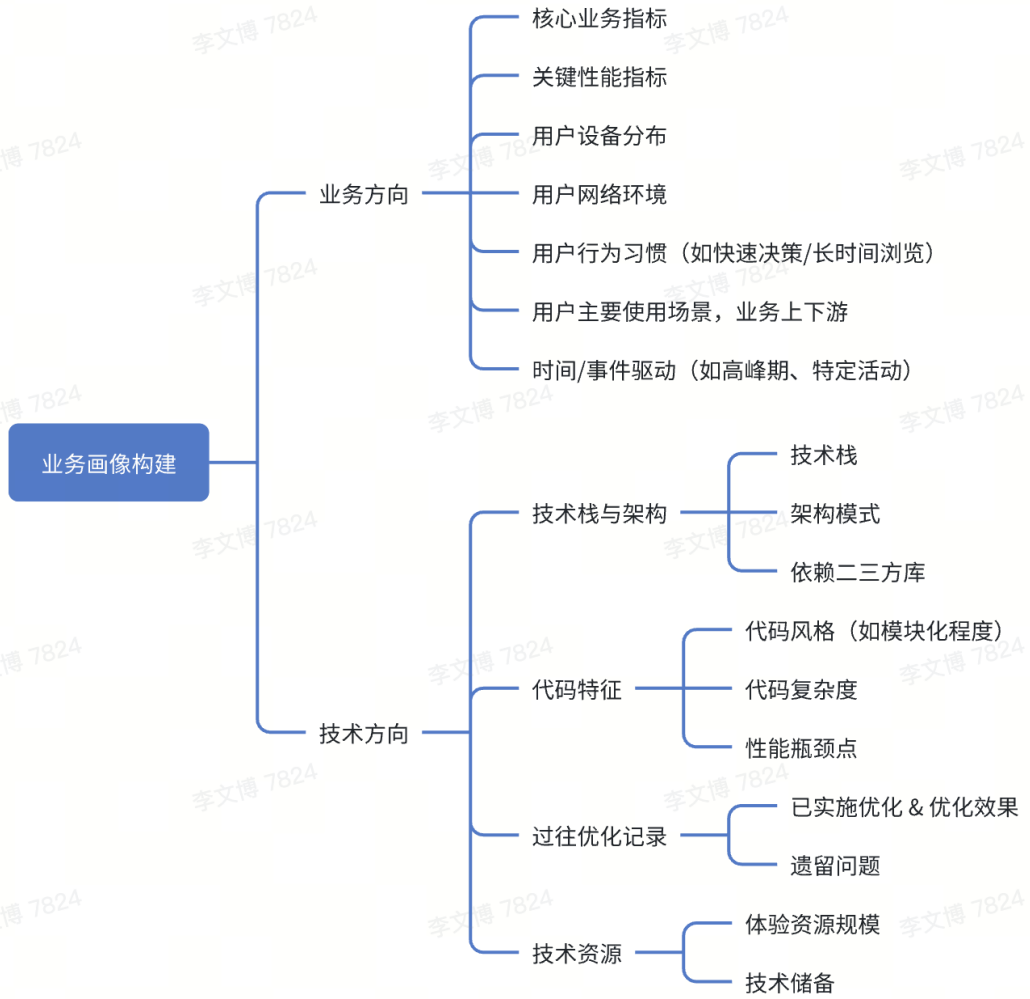
多领域专家系统协同推荐

- **基于大模型的**规划、**拆解能力构建虚**拟专家系统
  - **主**Agent**通**过语义分类和关键词**匹配，智能分**发问题至对应专家Agent。
  - **各**领域专家Agent**利用大模型上下文理解，生成**针对性子问题，**去各技**术领域知识库中检索。



业务画像驱动策略推荐

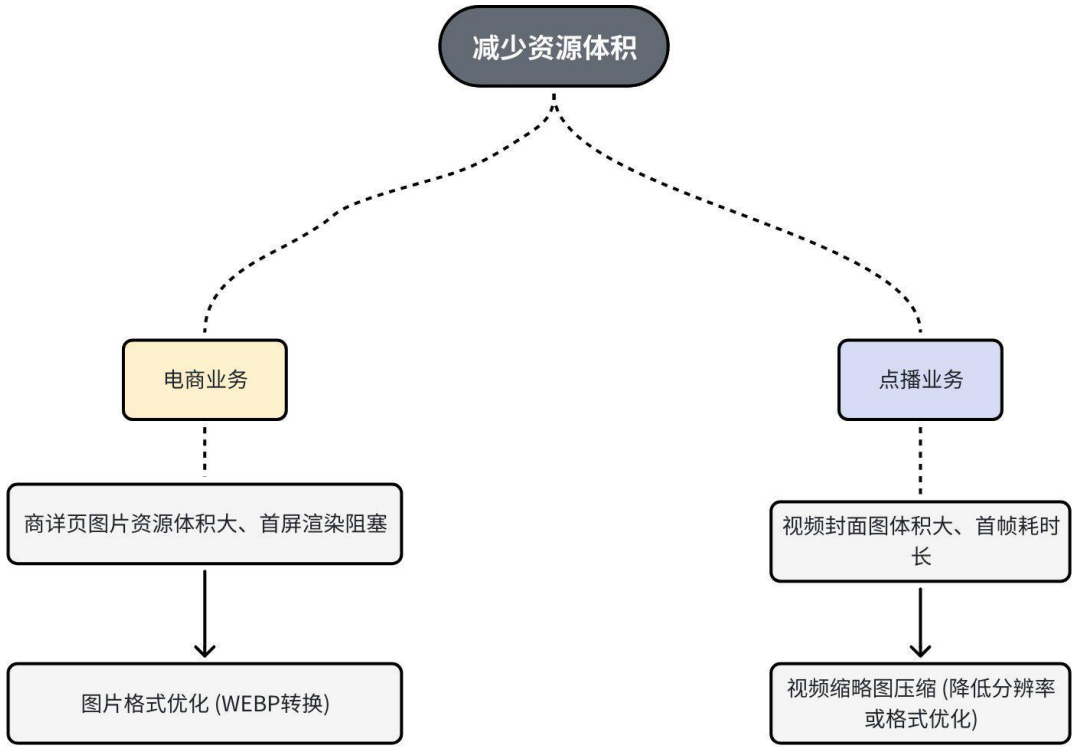
- **利用大模型的**语义理解能力，结合业务**画像**（**如**业务场景、核心指标、用户群**体特征、技**术上下文），主动解析业务**需求和痛点**。
- **通**过上下文推理和知识图谱关联，智能发现推理链路中的信息缺口（如未提及**的用**户设备分布对性能影响）。
- **自**动补全关键信息，基于历史数据或跨领域知识推测缺失维度（如“**弱网用**户占**比高可能**导致支付环节卡顿”）。



跨业务知识图谱融合

- **利用大模型构建跨**业务、多源知识图谱，整**合技**术方案、实验报告、最佳实践等多源知识，形成立体化知识库。

- **借助**图数据库检索方法，加速知识图谱中的**关**联推理，提升知识间关联性，确保高效精**准推荐**。



基于CoT的推理和方案生成

- 设计动态优先级评估机制，**通**过加权评分模型综合考量用户痛点紧急度、业务指标影响、**方案可行性和收益**预估，**精准排序各**专家建议，**确保核心**问题优先解决。

- 针对不同业务场景（**如**电商、**短**视频）总结定**制不同思**维范式，**确保推理**贴合业务需求。

应用场景

针对某页面加载慢问题，主专家，利用评分**模型**综合评估业务影响、方案可行性，排序图片格式优化和CDN调度优化建议，优先推**荐**“图片格式优化”，**因其效果**显著且接入成**本低**。

PART 05

总结与展望

未来规划

智能归因

全链路多模态融合：融合日志、监控、**用**户反馈、业务数据、**代**码运行态等全链路**数据**，实现体验问题的智能溯源与因果洞**察**，实现体验问题**“全景式”智能**归因。

代码级主动归因：自动收集与分析代码特**征**，**推**动归因深度延伸至架构和底层逻辑，实现异常实时智能发现。

智能推荐

全生命周期闭环优化：推荐系统直达代码层，自动生成修复方案并驱动落地。打通归因、推荐、**修复、**验证全流程，实现体验问题自闭环，，推动体验优化从**“被**动响应”**到“主**动发现”。**支撑**业务高质量发展。

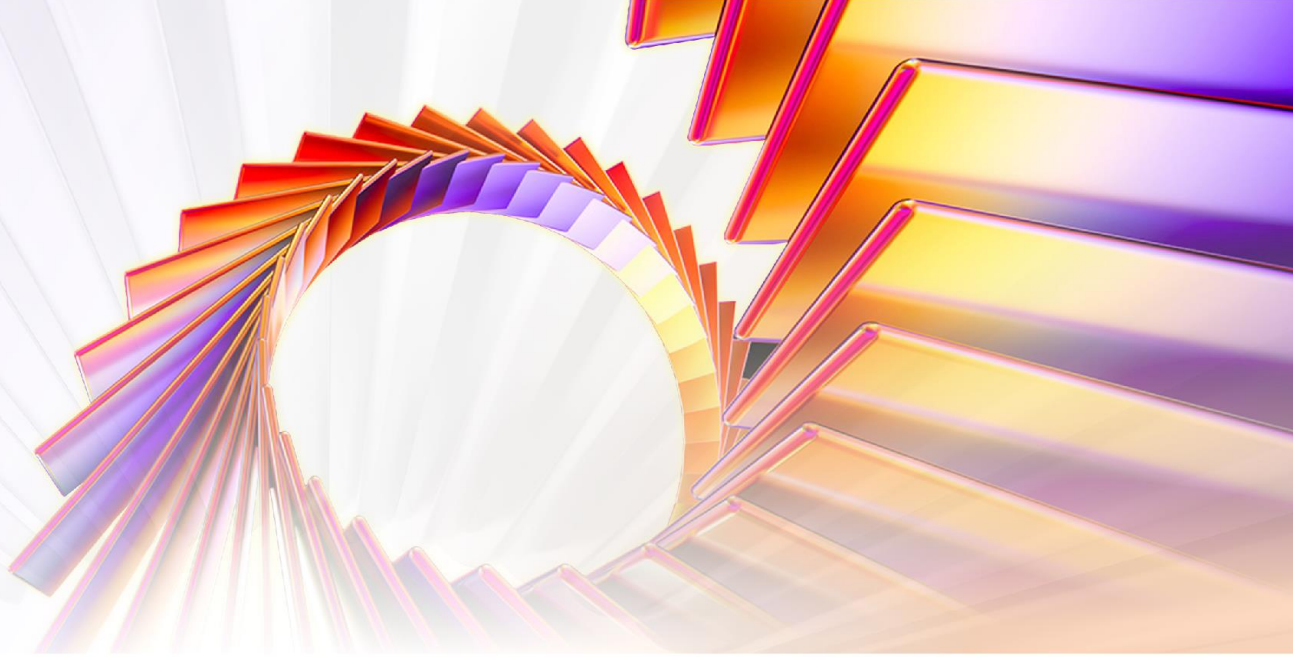
跨域知识迁移：拓展不同技术垂类的知识体系，**聚合客**户端、网络、图片、音视频等多技术领域**的**优化经验，智能迁移和复用最佳实践，赋能各类互联网产品的体验持续提升。

AiDD 6th 2025 | 05/23-24 | 上海站

2025 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发



大模型在得物部署优化实践

孟令公 | 得物

AiDD 6th 2025

参与调研您将优先获得

AiDD定制版
《AI+软件研发精选案例》

专属学习顾问
1对1需求对接

AiDD会后小调研

AiDD峰会致力于协助企业利用AI技术深化计算机对现实世界的理解，推动研发进入智能化和数字化的新时代。作为峰会的重要共建者，您的真知灼见对我们至关重要。衷心感谢您的参与支持！

扫码参与调研

2025 AI+研发数字峰会

拥抱AI 重塑研发

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

K+峰会

K+峰会 敦煌站
K+ 思考周®研习社
时间：2025.08.29-30

K+峰会 上海站
K+ 金融专场
时间：2025.09.26-27

K+峰会 香港站
K+ 思考周®研习社
时间：2025.11.17-18

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2025.05.23-24

AiDD峰会 北京站
AI+研发数字峰会
时间：2025.08.08-09

AiDD峰会 深圳站
AI+研发数字峰会
时间：2025.11.14-15

AiDD峰会详情

AiDD 6th 2025

2025 AI+研发数字峰会

AI+ Development Digital Summit

感谢聆听！

扫码领取会议PPT资料