AiDD AI+研发数字峰会
AI+ Development Digital summit
第5届

蚂蚁数科AI Agent质量保障体系建设探索

AI质量团队 | 蚂蚁数字科技

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

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：2025.10.17-18

K+峰会 · 香港站
K+ 思考周®研习社
时间：2025.11.25-26

K+峰会详情

AiDD峰会

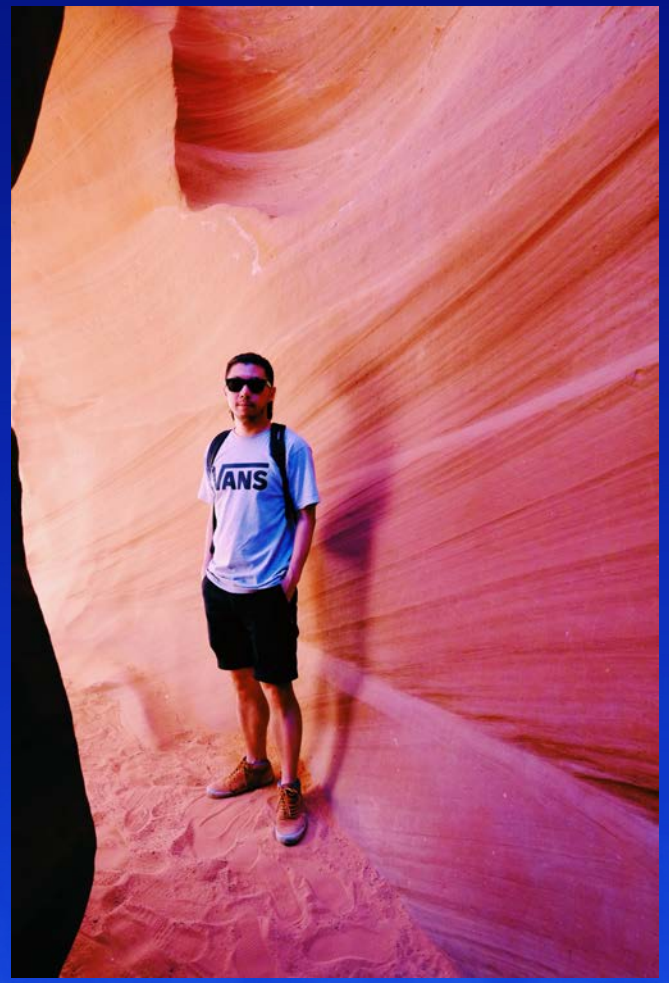
AiDD峰会 · 上海站
AI+研发数字峰会
时间：2025.05.17-18

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：2025.11.28-29

AiDD峰会详情

演讲嘉宾



李赫

蚂蚁数字科技线测试开发专家

10余年软件开发及测试经验，在测试工具开发、质量效能平台建设等方向有丰富的落地建设经验，先后就职于网易，淘宝，腾讯音乐，现任蚂蚁数科AI业务质量&质量效能平台负责人 。

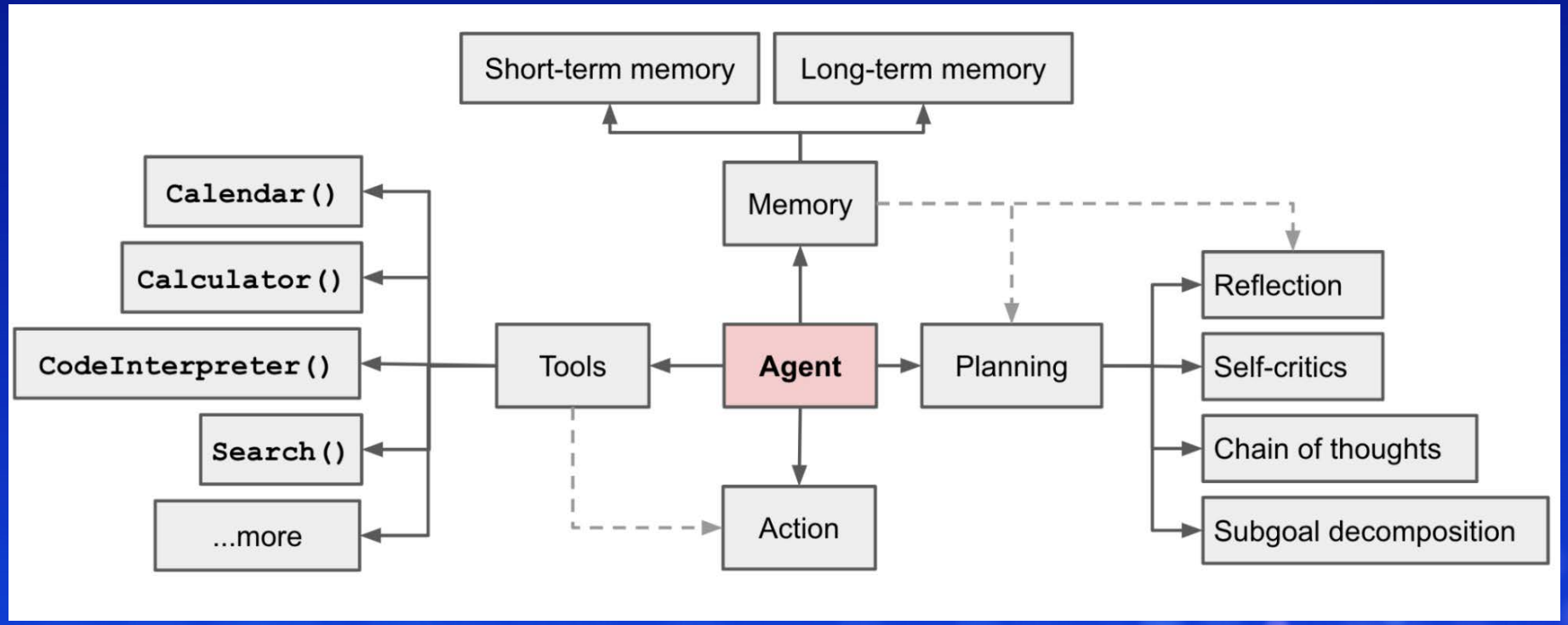
目录
CONTENTS

PART 01

蚂蚁数科Agent重要性和挑战

什么是Agent

Agent = LLM + Planning + Feedback + Tool use



Lilian Weng关于Agent的定义

蚂蚁数科Agent的应用场景

智能助手

数字人一体机

售后运维

Agent质量保障重要性和挑战

错答问题

性能+幻觉问题

安全问题

Agent质量保障重要性和挑战

业内挑战

- 缺乏业内成熟理论支撑
- 缺乏统一的评估标准
- 缺乏丰富的场景化测试样本

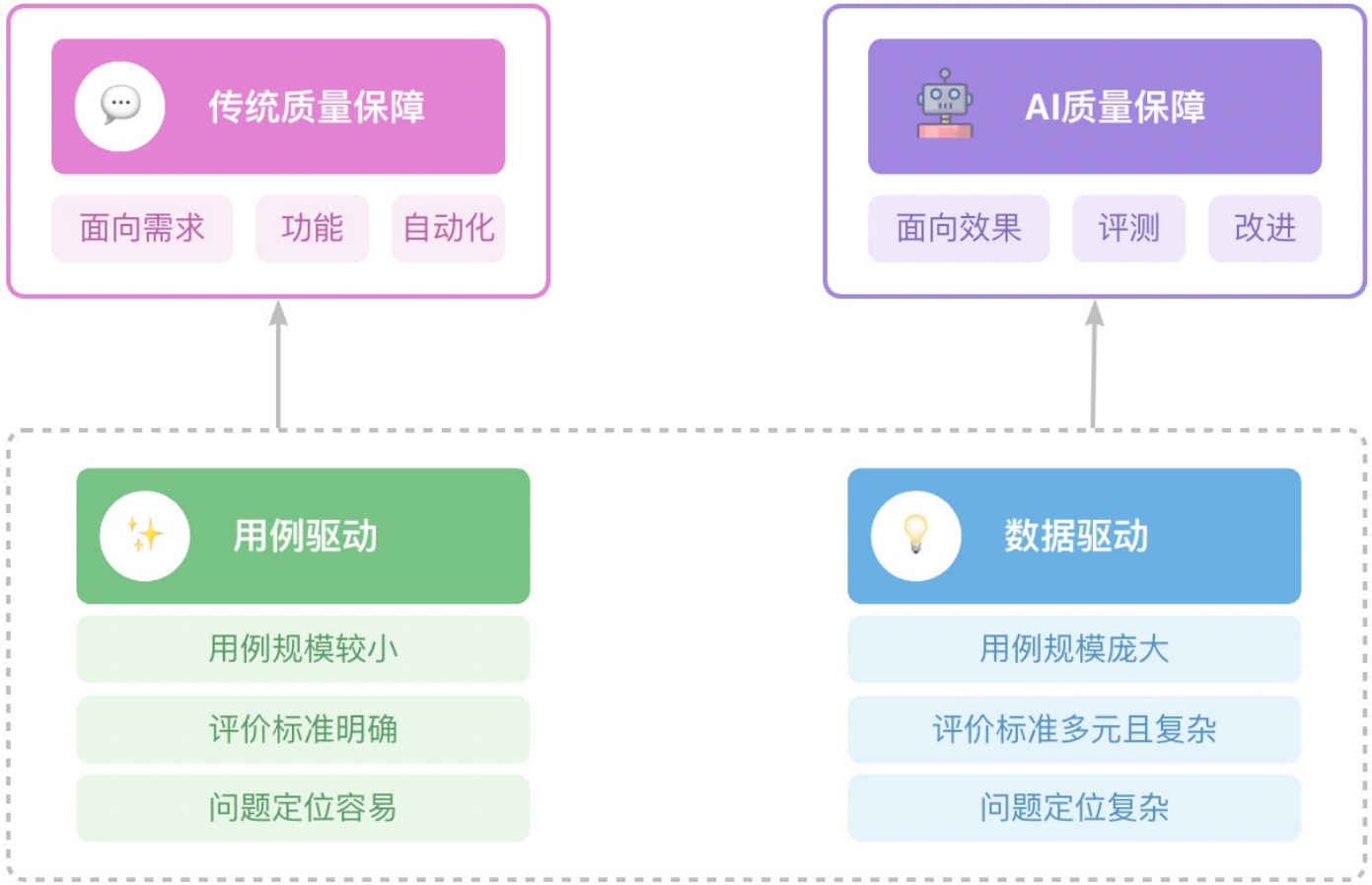
蚂蚁数科挑战

- 商业化产品快速交付
- 场景化的领域知识构建
- 标准化接入和集成

PART 02

Agent质量保障的整体策略

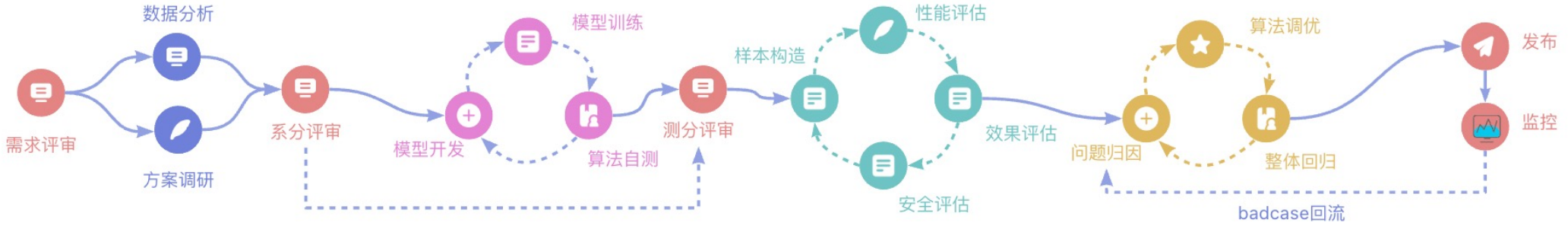
传统质量保障 vs AI质量保障



面向AI质量保障转型

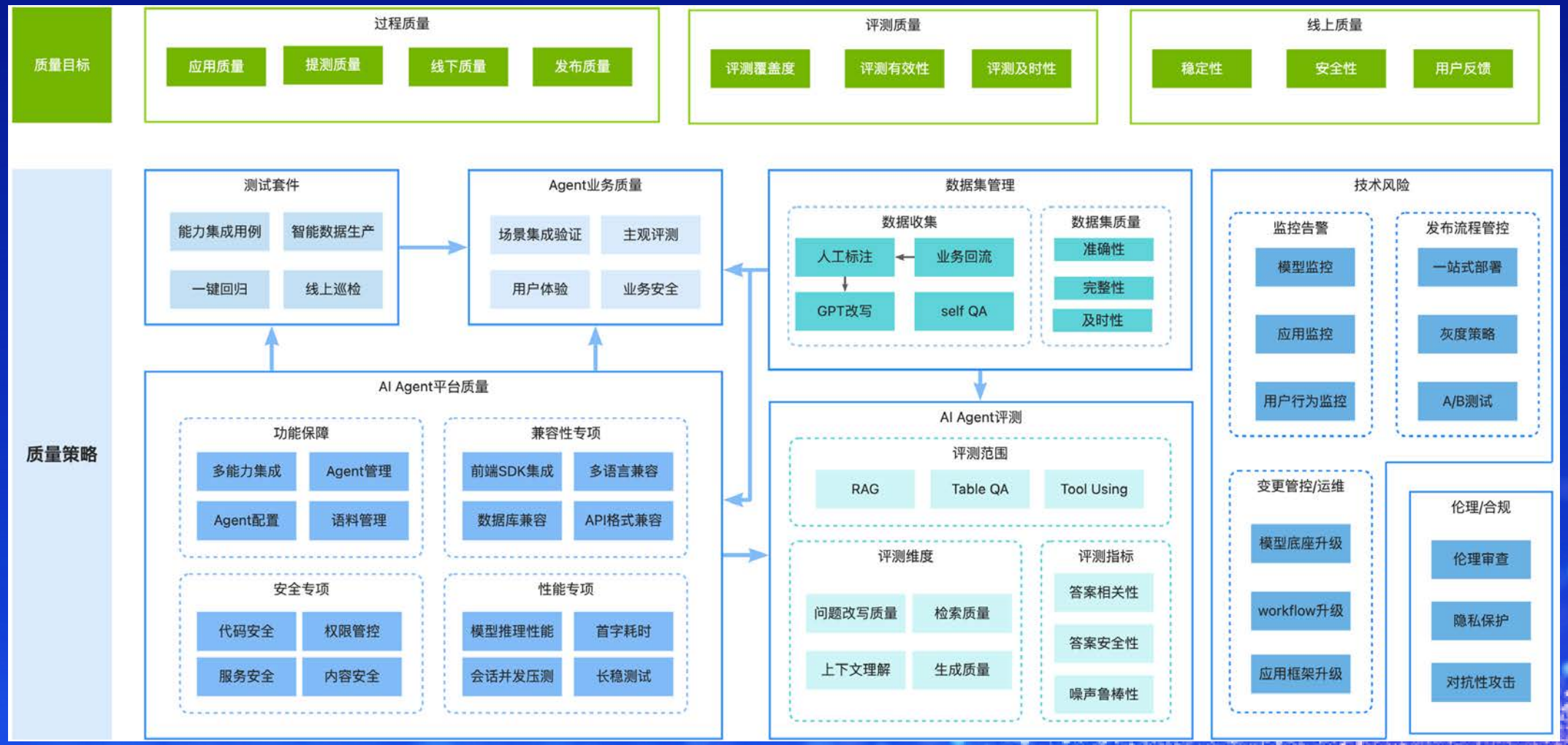
- 算法工程能力
- 数据分析能力
- 大模型应用能力
- 模型训练和部署能力
- 大模型安全知识
- 业务领域专业能力

Agent质量保障整体流程

阶段	需求阶段	模型开发	测试阶段	模型调优	上线&监控
流程					
任务	数据分析 方案调研	训练集构造 模型训练	功能验收 整体评估	调优&回归	业务验收 整体交付
参与人	PM、PD、运营、算法、测试	算法、测试	算法、测试	算法、测试	PD、业务、算法、测试
产出	✅需求文档 ✅调研文档 ✅系分文档	✅模型文档 ✅自测报告 ✅测分&测试用例	✅测试样本 ✅评测方案 ✅评测工具	✅效果评测报告 ✅性能评测报告 ✅安全评测报告	✅验收报告 ✅业务反馈

Agent质量保障整体方案

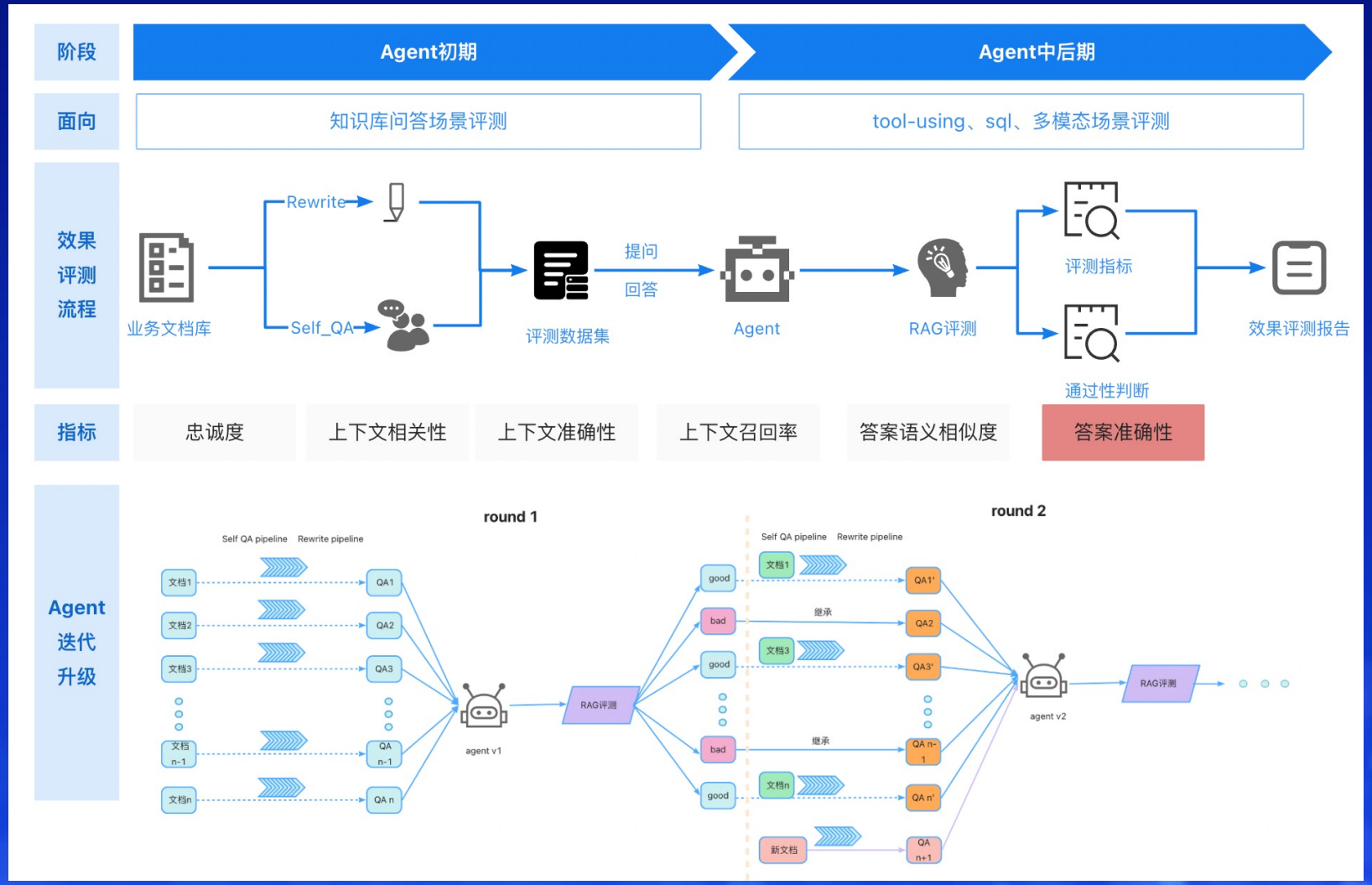
构建面向商业化Agent的质量保障体系与全面评测能力



PART 03

Agent质量保障体系建设实践

Agent效果评测流程



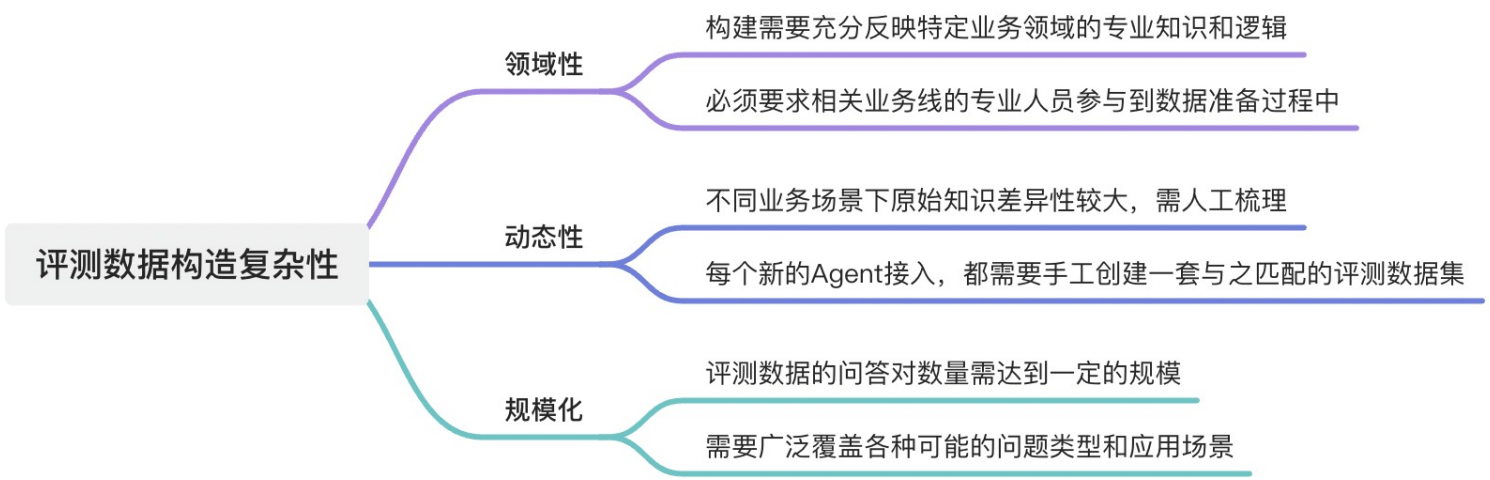
全流程自动化

- 评测数据集生产
- RAG评测

快速交付

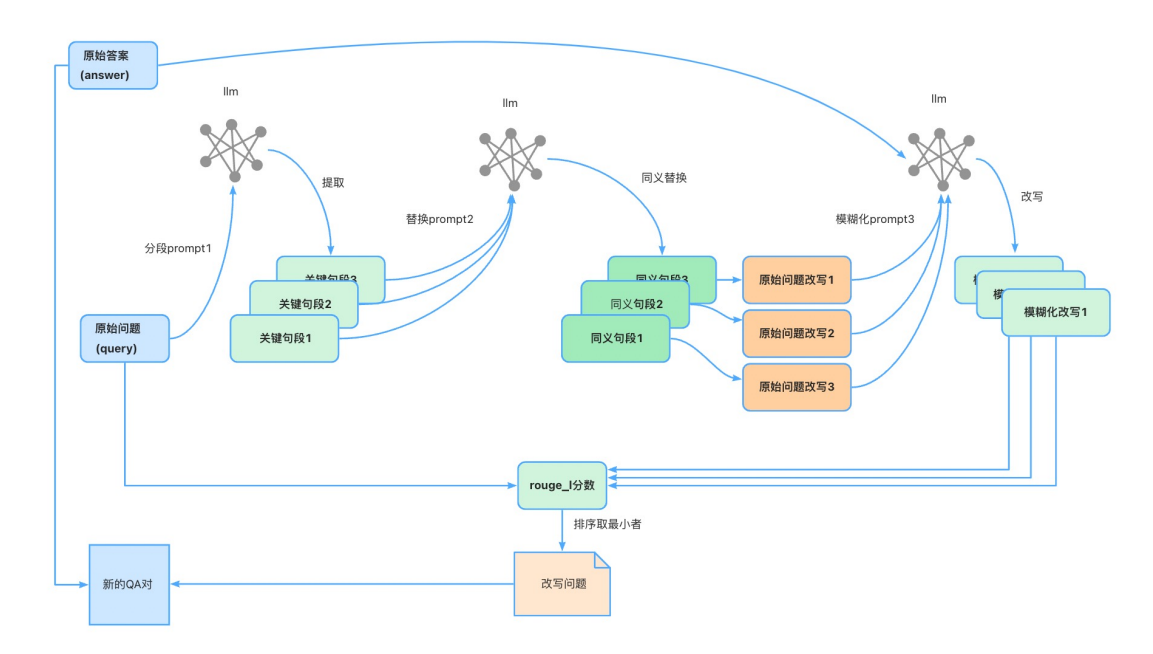
- Self_QA: 10s/文档
- Rewrite:20s/问题
- Ragas打分：20s/样本

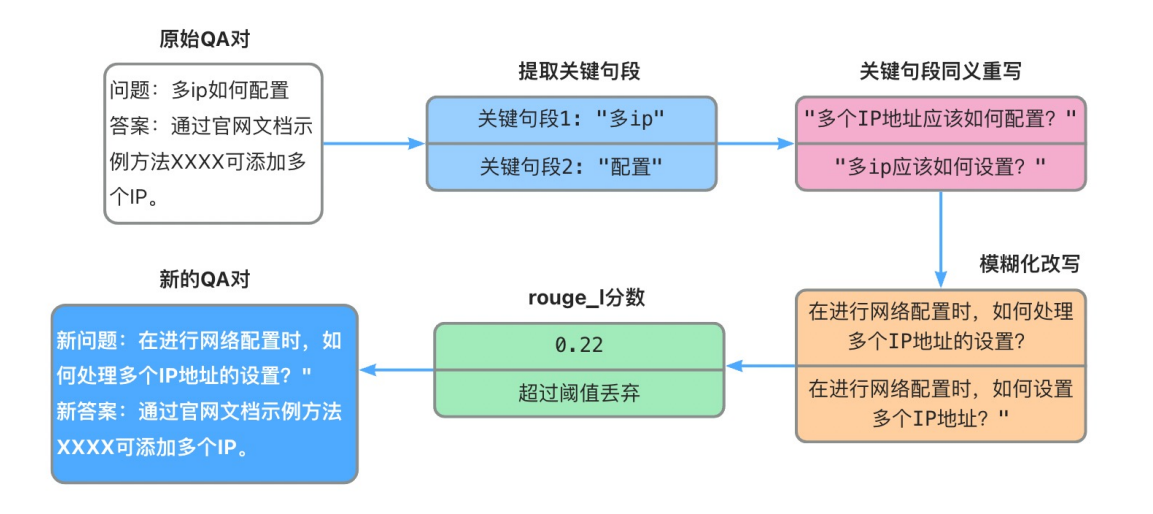
Agent评测数据生成自动化的挑战



在RAG评测中人工准备数据集不仅效率低下、耗时费力，而且难以满足高精度与全面性要求。因此，**开发自动化评测数据集生成策略**变得至关重要迫切。所以**Self-QA**和**Rewrite**两大关键自动化方法来应对这一难题。

评测数据构造方式： Rewrite





原始问题改写

- 关键句段提取
- 关键句段同义重写
- 模糊化改写
- rouge_l相似度过滤

评测数据构造方式：Rewrite（问题改写）

原始问题：**多IP如何配置**

普通prompt问题改写的结果

最终改写结果	与原问题相似分数
如何进行多IP的设置?	0.33
如何设置多个IP地址的配置呢?	0.46
如何对多个IP进行设置?	0.33

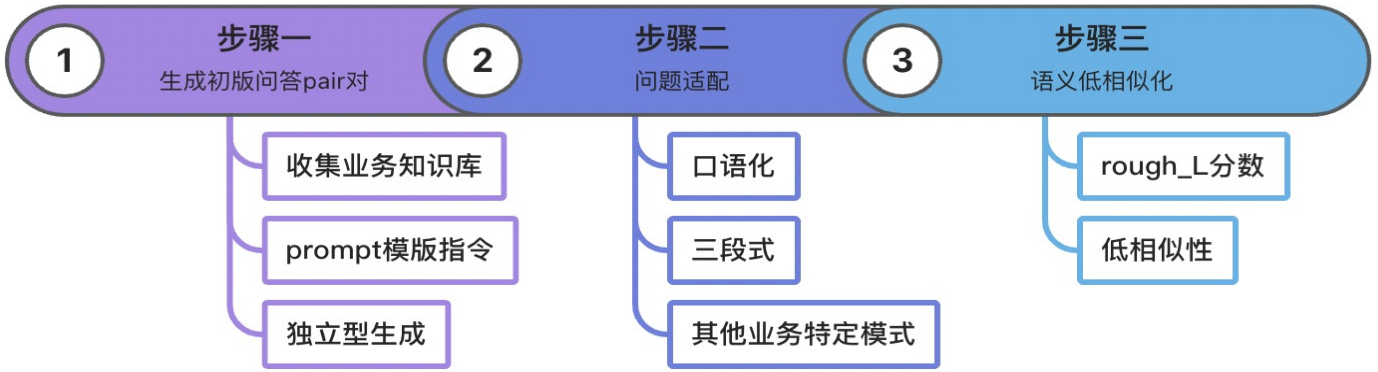
VS

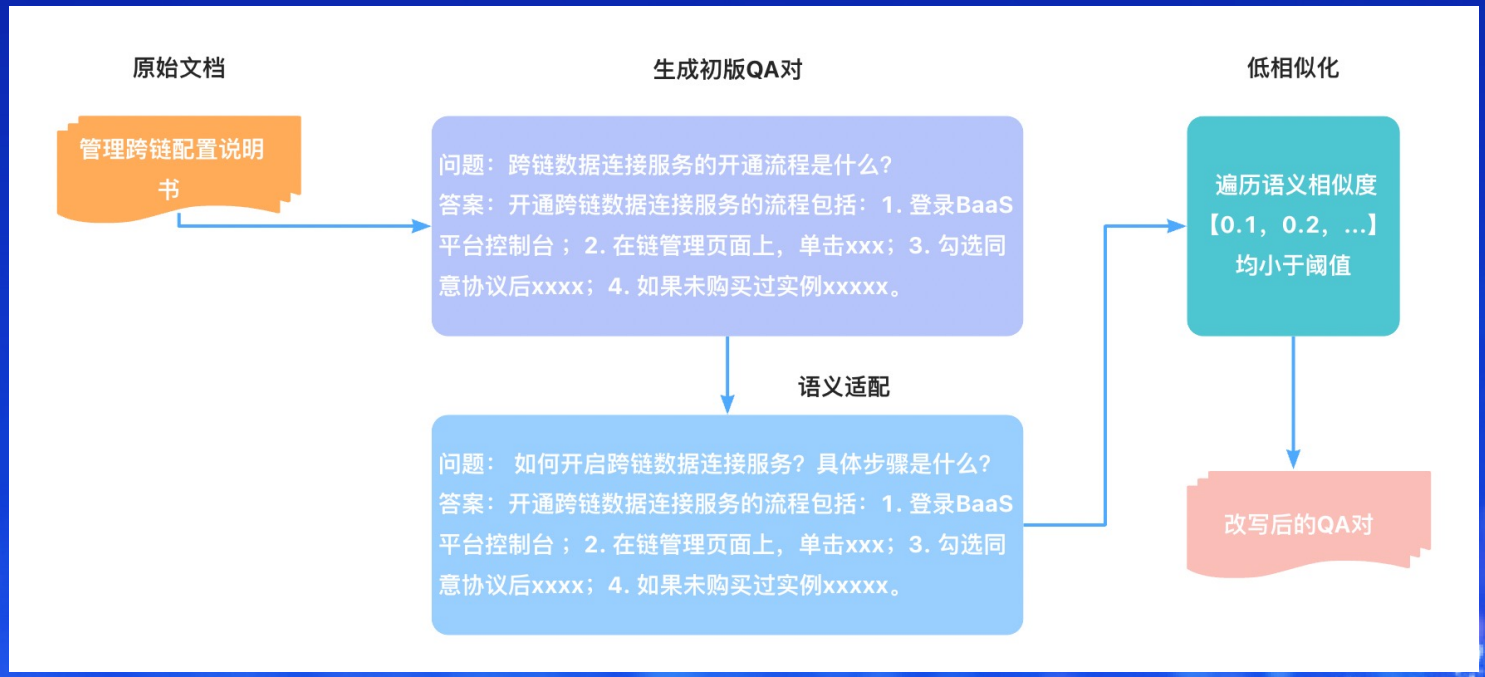
Rewrite pipeline问题改写的结果

关键字段	最终改写结果	与原问题相似分数
多ip	在进行网络配置时，如何处理多个IP地址的设置?	0.22
配置	在进行网络配置时，如何设置多个IP地址?	0.25

相比普通prompt版本，Rewrite pipeline方法改写问题的相似分数更低，句式也进行了改变，但仍然是相同含义的问题。

评测数据构造方式：Self-QA（自动生成问答对）





- **Self-QA生成**
  - **初版问答对生成**
  - **口语化改写**
  - **三段式生成**
  - **rouge_l相似度过滤**

评测数据构造方式：Self-QA（自动生成问答对）

普通prompt生成的QA对

```
{
  "question": "在使用区块链连接不同网络的时候，我们如何确保这些连接是安全的，并且可以正确地传递信息？",
  "answer": "确保跨链合约的安全性和信息传递的有效性，可以采取以下几个措施：首先，在部署跨链合约前，进行智能合约的全面审核，包括第三方代码审计，以识别潜在的漏洞和漏洞利用风险。其次，采用完善的权限管理机制，仅允许联盟管理员及经过授权的用户进行合约的创建和管理。再者，增加链上日志记录和监控功能，实时追踪合约的执行情况，以便及时发现和处理异常活动。此外，在开通跨链服务之前，充分评估外部数据源的信誉，确保其数据的真实性和可靠性，从而有效减少合规性风险和数据篡改的可能性。最后，定期更新合约并进行安全性评估，保持与最新的安全标准同步。",
 },
```

VS

Self-QA生成的QA对

```
{
  "question": "如何开启跨链数据连接服务？具体步骤是什么？",
  "answer": "开通跨链数据连接服务的流程包括：1. 登录BaaS平台控制台，进入联盟后，单击目标联盟链上的主链管理；2. 在链管理页面上，单击跨链管理标签并从下拉菜单中选择跨链服务，然后单击申请开通；3. 阅读《蚂蚁链跨链数据连接服务》，勾选同意协议后单击确定即可使用跨链数据连接服务。如果已购买过实例，开通后可直接使用；如果未购买过实例，开通后需购买一个跨链通道数后才能使用。"
 }
```

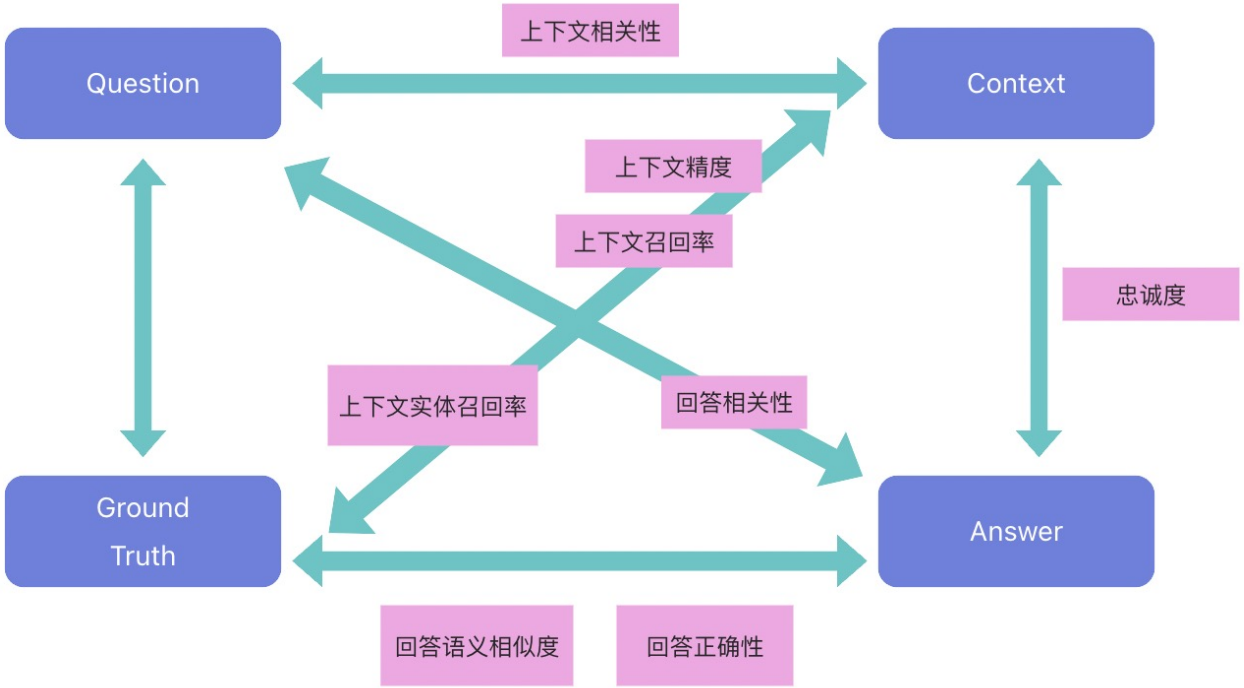
- 普通prompt生成方案针对概念性的提问，导致文字性的内容较多
- Self-QA生成的侧重于某个技术难点询问其操作步骤，更加聚焦，也更加的口语化

大模型评测&Agent评测差异

评测方面	大模型评测	Agent评测
关注点	模型的基础架构、训练数据和算法	模型在特定任务或应用中的效果
评测指标	精度、召回率、F1分数等综合性能	生成内容的相关性、用户满意度等
测试数据	标准数据集和基准测试集	针对具体应用场景自定义的数据集
应用场景	广泛的自然语言处理任务	特定应用，如信息检索、问答系统
评测目的	了解模型的整体能力	评估在实际应用中的性能和效果
反馈周期	更新基于新算法或新的训练数据	更新基于用户反馈和应用需求

Agent效果评测方案

Ragas指标体系



Ragas评测指标

指标	描述	范围
准确性	衡量答案与上下文事实一致性，值越高越好。	(0, 1)
答案相关性	评估答案与提示的相关性，分数高表示更好。	(0, 1)
上下文recall	衡量上下文与真实答案的一致程度，数值越高表示性能越好。	(0, 1)
上下文精确度	评估上下文中相关项的排名，理想为前几名，分值越高越准确。	(0, 1)
上下文相关性	根据问题和上下文评估检索到的相关性，数值越高表示相关性越好。	(0, 1)
上下文实体recall	衡量检索到的上下文对真实答案实体的召回率。	(0, 1)

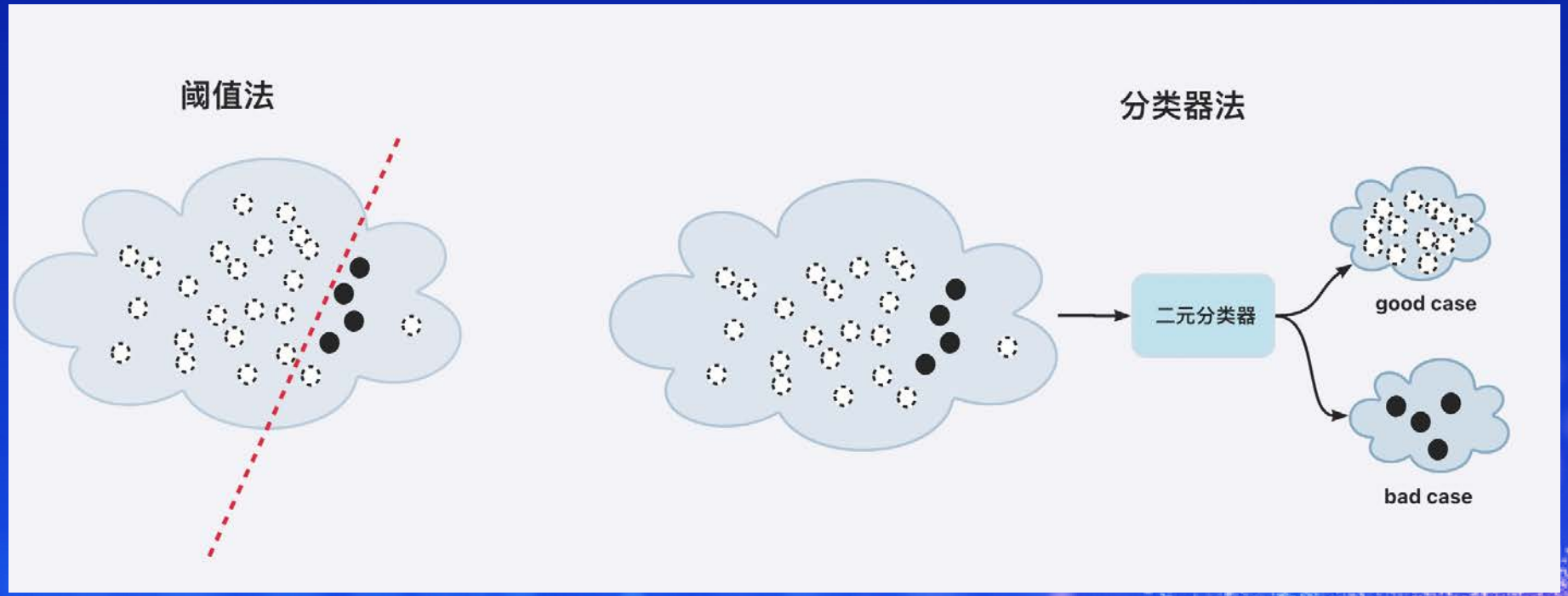
Agent效果评测方案

Ragas评测示例

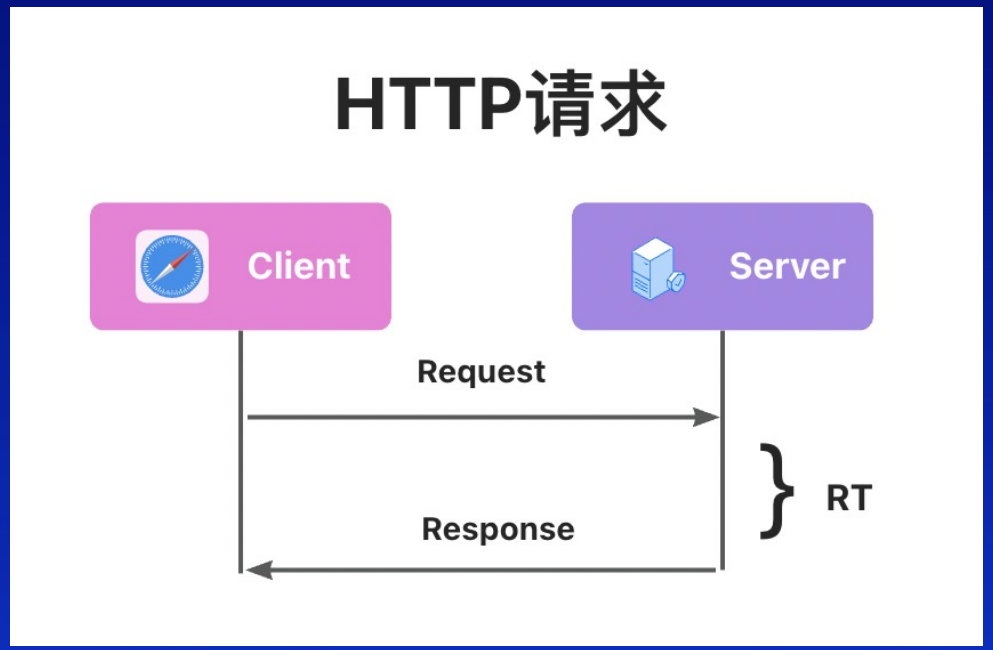
Question	Answer	Ground_truth	Precision	Recall	Answer correctness	f1_score	similarity_score	结果
有本书的作者是海子，书名好像是关于太阳的，书名是什么，出版社是哪个，出版年份是哪一年?	书名: 你来人间一趟，你要看看太阳 出版社: 湖南文艺出版社 出版年份: 2019-3	本书的书名是《你来人间一趟，你要看看太阳》，出版社是湖南文艺出版社，出版年份是2019年	1	0.6	0.89	0.88	0.94	正确
有本书讲述了一个绝症患者如何面对绝望并发起全面进攻的故事，书名是什么，作者是谁，哪个出版社出版的?	"书名：《活着》 作者：余华 出版社：作家出版社 （注：此回答基于常识，高质量知识未提供具体信息。《活着》虽然不是直接描述绝症患者的故事，但深刻描绘了主人公面对生活绝望的坚韧与抗争，与问题描述的情境相契合。）"	这本书名为《相信》，作者是蔡磊，由中信出版社出版。	0	0	0.21	0	0.85	错误

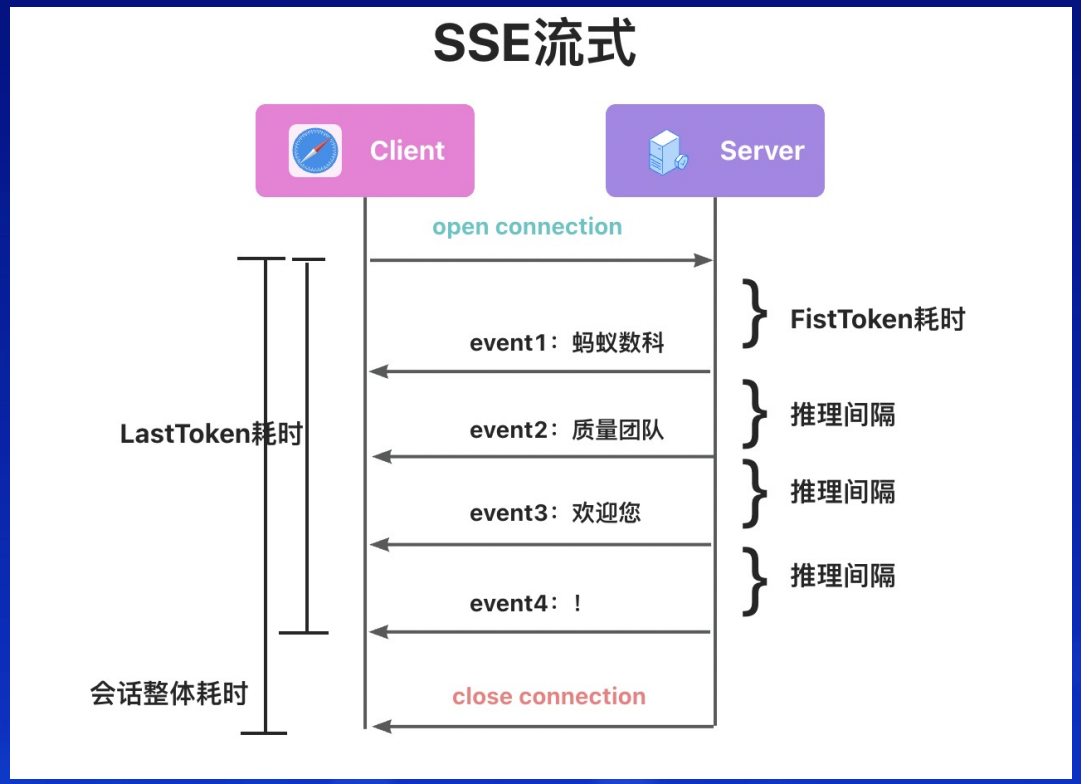
Agent效果评测通过判定

- 当前市面上并没有类似的方案应当如何判定评测通过，我们在实践中，主要是探索了两类方案：阈值法和分类器法：
- 阈值法：划定一批指标，这些指标在对应阈值a下的通过率（而非全部）是否高于某个值b，来作为判定是否通过的准则；
- 分类器法：给定多个相关评测指标的结果分数，需要分类器进行二元分类，设置整体的通过率应高于某一阈值；



Agent性能评测方案





- Concurrency：并发数
- QPS/TPS：每秒查询/事物数
- RT：响应时间

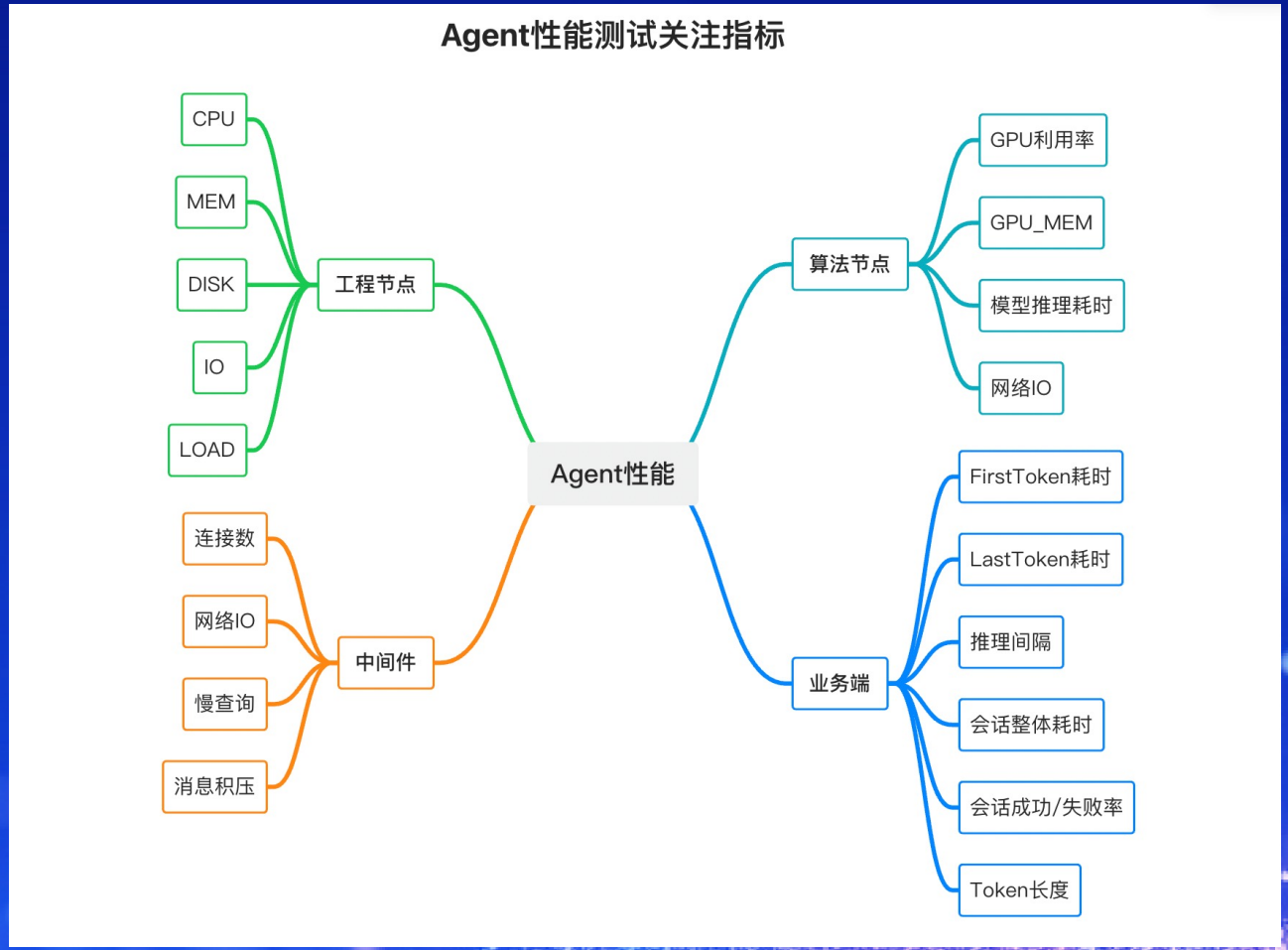
- 同时在线会话数
- FirstToken/LastToken耗时
- 推理间隔
- 会话整体耗时

Agent性能评测方案

打造必要的性能测试能力

- 稳定的发/控压能力
- 灵活的场景建模能力
- 准确的性能分析能力
- 实时的数据可视化能力
- 完备的报告生成能力

确定必要的性能测试指标

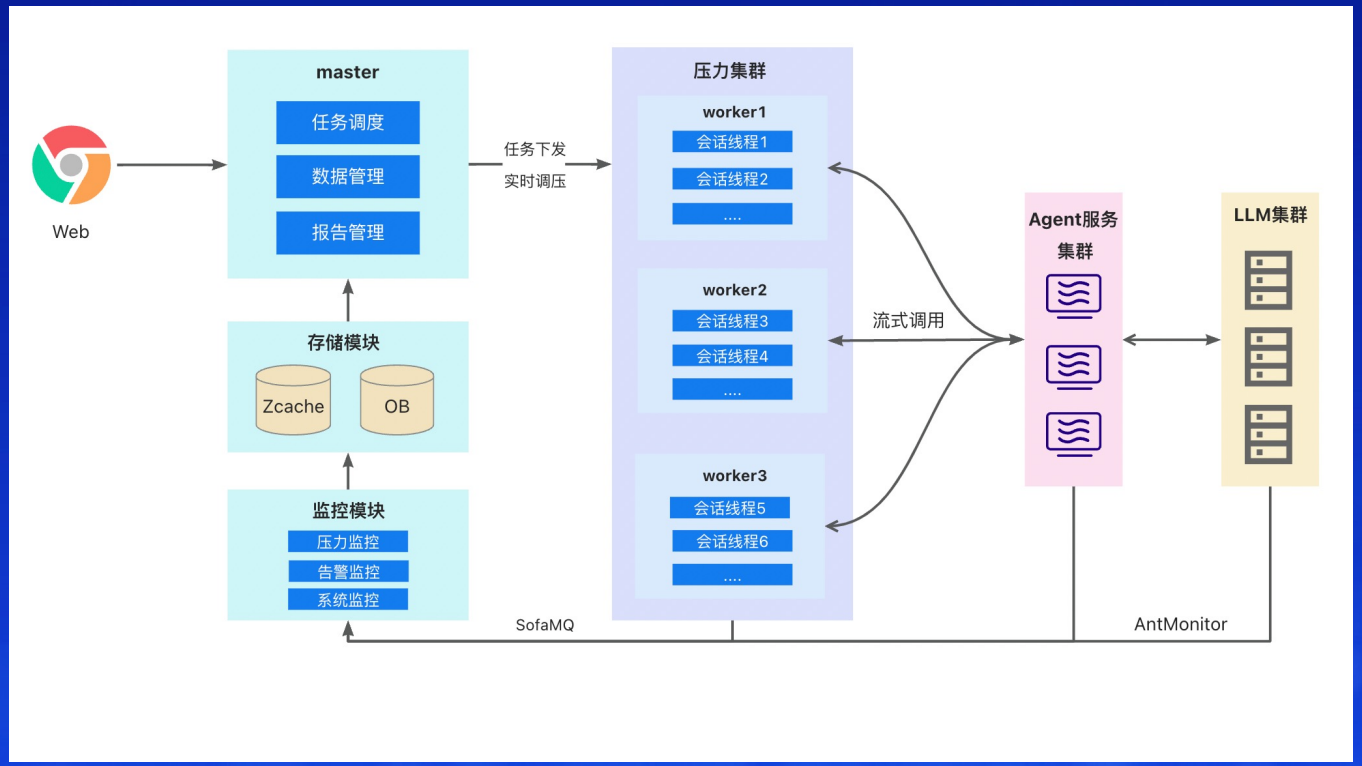


Agent性能评测方案

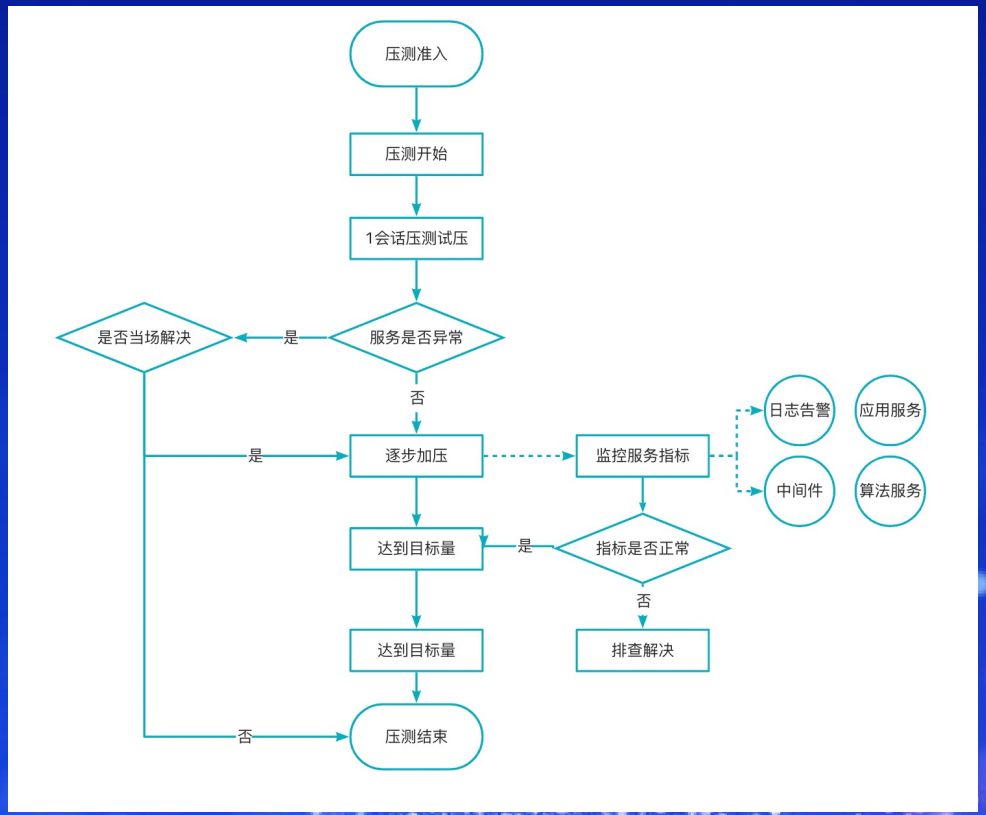
性能测试模块整体架构

Agent性能评测方案

Agent性能测试技术架构



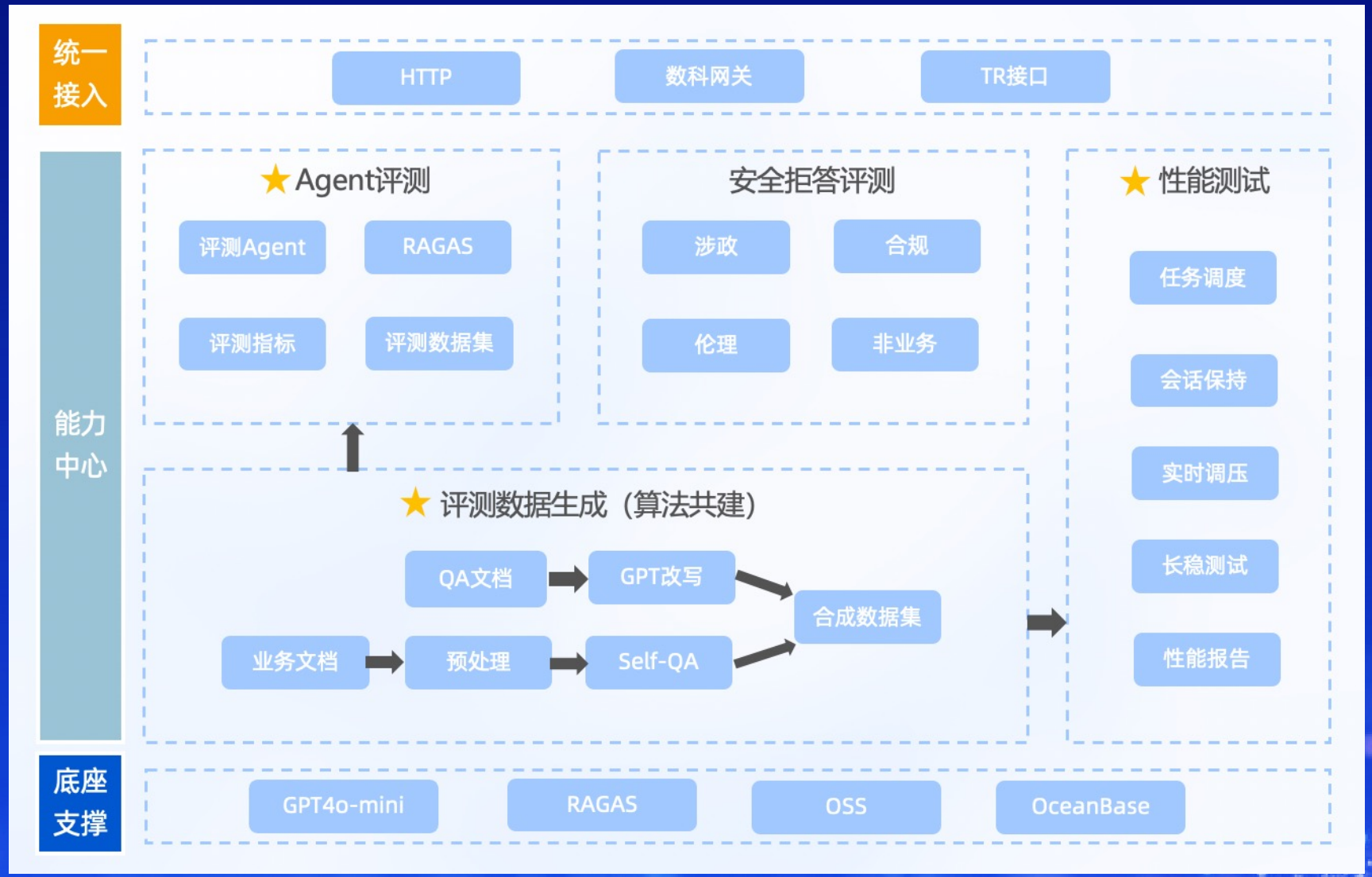
Agent性能测试执行流程



PART 04

Agent一站式质量平台建设实践

Agent一站式质量平台整体架构



- 一键自动化评测
- 分布式高效评测
- 全面的评测能力
- 灵活的能力扩展

PART 05

总结&展望

总结&展望

总结

- ✓ 场景化评测方法的探索
- ✓ 场景化评测数据的生成
- ✓ 场景化Agent评测实践
- ✓ Agent一站式评测平台建设

未来展望

- ➢ 评测自动化流水线建设
- ➢ Tool Use评测能力建设
- ➢ 评测数据质量分能力建设
- ➢ 多模态质量评测能力建设
- ➢ 质量评测闭环能力建设

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

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：2025.10.17-18

K+峰会 · 香港站
K+ 思考周®研习社
时间：2025.11.25-26

K+峰会详情

AiDD峰会

AiDD峰会 · 上海站
AI+研发数字峰会
时间：2025.05.17-18

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：2025.11.28-29

AiDD峰会详情

AiDD AI+研发数字峰会 第5届
AI+ Development Digital summit

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时代