QCon
全球软件开发大会

InfoQ 极客传媒

演讲主题

蚂蚁 DeepInsight 智能分析 Agent 在业务场景的落地实践

姓名：余志鹏（百恼）

Title：高级技术专家

1 业务问题定义及当前进展

2 ChatBI六大难题及解决方案

3 评测集构建

4 长程任务探索

个人简介

余志鹏，蚂蚁集团高级技术专家，在性能优化、架构设计、数据分析以及 AI + BI 等方向，有着深厚的专业积累和实践经验。

2018 年加入蚂蚁集团，深耕大数据领域 7 年，目前负责蚂蚁数据分析平台 DeepInsight ，带领产品进入到智能化时代。此前，在阿里巴巴工作 5 年，负责 Aliexpress 营销平台，积累了丰富的平台运营与管理经验。

01

业务背景介绍

大模型在分析领域解决两大难题：① 分析领域门槛； ② 分析产品门槛

分析领域门槛

- 描述性分析
  - 趋势分析
  - 分布分析
  - 离散程度分析
- 推断性分析
  - 假设检验
  - 回归分析
  - 相关分析
  - 时间序列分析
- 探索性分析
  - 假设检验
  - 回归分析
  - 相关分析
  - 时间序列分析

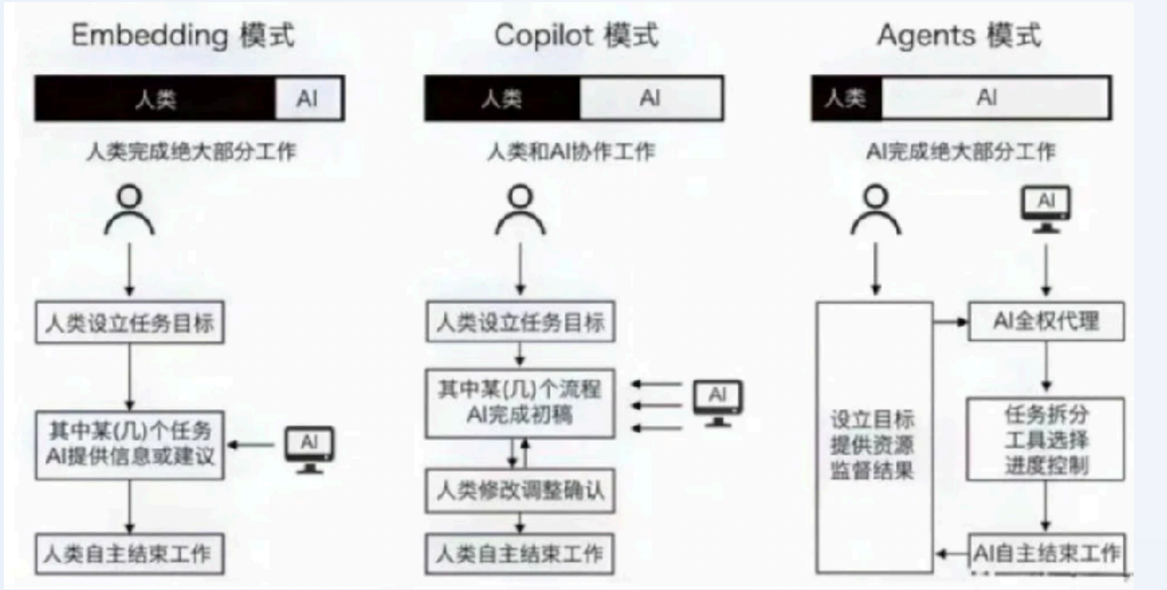
分析产品门槛

- 分析语言
  - SQL
  - Python
  - DAL
  - DAX
  - DAL
  - DAL
- 产品门槛
  - 数据准备
  - 报表制作
  - 报表阅读
  - 自助分析
  - …
  - …

传统BI时代，因为分析领域有很强的专业性，以及分析产品有较高的门槛，企业需要招专业的人才能解决

AI的演进路径：目前大部分场景还处在Copilot模式，小部分场景向Agents阶段探索

人类与AI协作的3种模式



书籍：《The AI Alignment Problem》

- **Embedding模式（过去）**：任务主线流程基本不变，其中某（几）个步骤由AI驱动完成，提高单个步骤的效率和质量。
- **Copilot模式（现在）**：在人的委托下，AI对整个流程的全部或部分进行智能化驱动，提高整个流程的完成效率和质量。
- **Agent模式（未来）**：完全把任务托管给AI并获得结果，不再关心中间过程，由AI自主完成任务。

模式演变的核心关键

小模型 → 大模型 → 大模型 Plus

DeepInsight从2023年开启智能化

开启智能化 | Copilot | ChatBI & DataSage

2017年 | 2019年 | 2020年 | 2021年 | 2022年 | 2023年 | 2024年 | 2025年

DeepInsight正式上线

用户随时随地看报表

新春红包，落地创新的数据驱动决策方案

运营完成一次决策耗时从天缩短到分钟级

DAL—创新的数据分析语言

基于DI新分析语言，红包场景中实现小时级报表研发

DI-NLQ一期上线自然

通过DI-NLQ，用户临时取数分析耗时缩短到分钟级

人群洞察分析

服务支付宝用户场景，指标人群分析、人群洞察自动化

EasyInsight 智能分析

数据科学和基于算法分析分析场景落地

性能大脑

大数据量下性能自动优化解决方案

知数据

蚂蚁统一看数门户

解决方案建站平台

一站式数据分析应用低代码搭建平台

GroupInsight 增长分析

数据驱动业务增长的数据分析产品

取数Copilot上线

基于LLM的数据分析智能助理首次应用落地

DI 智能化系列全面开放

报表/自助分析/知数据4个Copilot全面开放，6月首批通过信通院测评。

DeepInsigh商品化

正式对外输出，上架数科售卖商品

DI 智能化持续升级，打造行业竞争力（进行中）

深度智能化，底层模型和产品体验优化，一流行业竞争力。

当前DeepInsight重点建设智能问答ChatBI & 长程任务产品DataSage

智能取数ChatBI

智能分析DataSage

口径问答Agent

度量Agent

智能解读Agent

02

ChatBI六大难题及解决方案

可能有人会说2025年了，还在分享取数Agent是不是有Low，毫无新意

智能问答

长程任务

归因分析

行研分析

趋势分析

DataResearcher

取数Agent

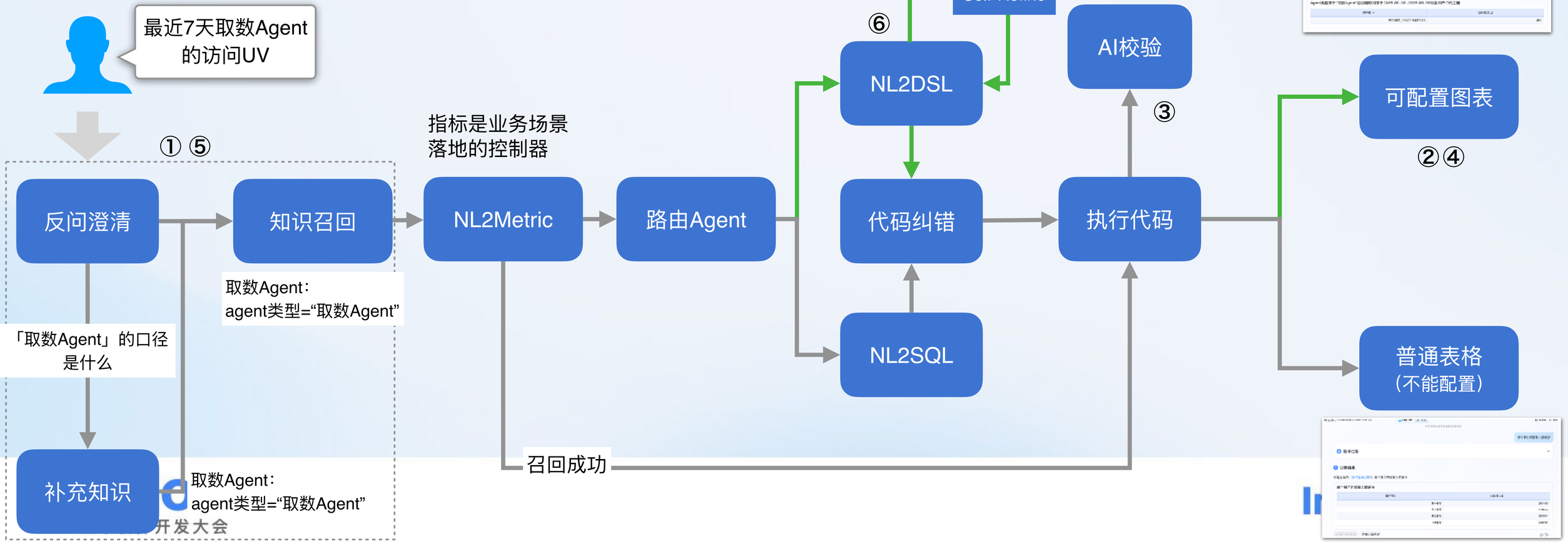
分析Agent

报告Agent

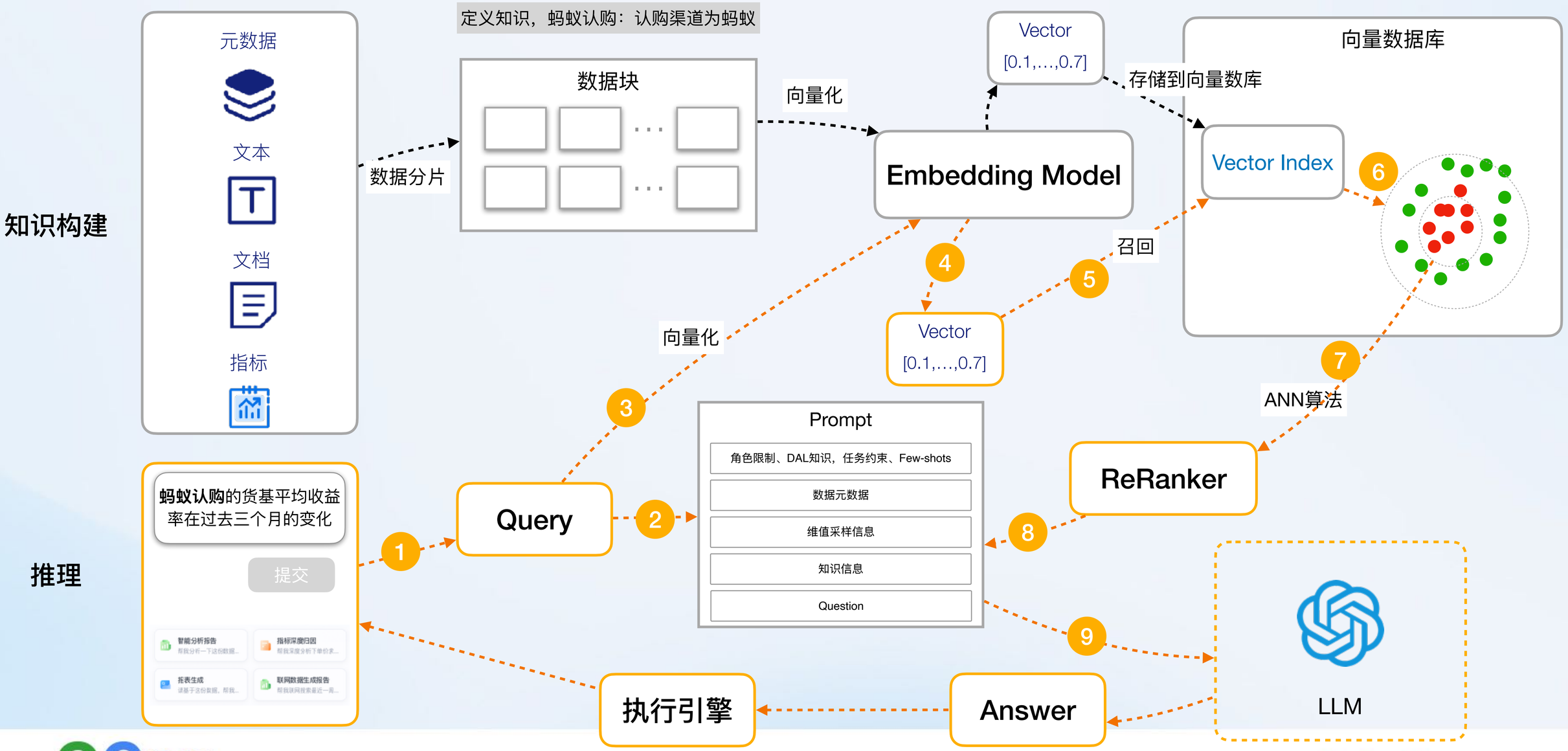
不管上面以何种方式呈现，最关键的仍然是取数，取数的准确率决定上面用户问答的质量
今天我们要讲的是如何通过技术+产品实现取数Agent的落地

ChatBI要落地，面临以下6大挑战

① 特领域下的知识，以及个人习惯问法，大模型如何识别
② NL2Code的正确率取决于使用的程序语言，Metric>Python>SQL>DSL，但是Python无法处理大数据量，SQL转化为可配置图表有较大的限制，Metric覆盖低，且泛化能力差
③ 不是所有需求通过自然语言描述的效率都更高，协同产品交互更有性价比，那么自然语言和产品交互两种模态如何协同
④ 模型生成的结果，用户如何验证是否正确。
⑤ 模型生成错误后，用户应该如何继续操作。
⑥ 模型响应时间和性能如何进行权衡，模型越大，正确率越高，但是响应时间也越长。

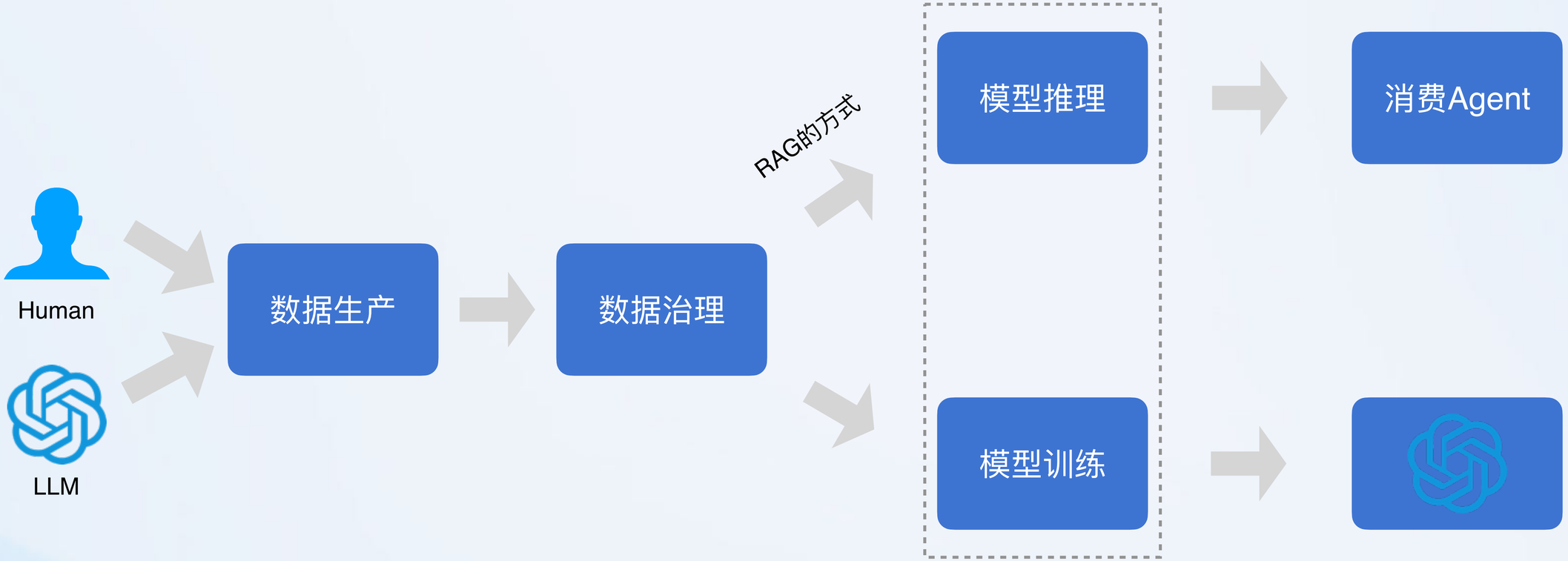


一、知识是“燃料”，构建知识体系，才能实现在具体的领域场景进行落地



数据治理会成为AGI落地的关键，基于大模型让数据质量有了标准

数据治理面临两个难题：① 标准化的数据如何产生价值难以衡量；② 怎么样的数据才是高质量没有一个判断标准



而在AGI时代，所有的数据都是给大模型消费，那么治理数据的的好坏就有了判断标准，
让数据治理成为了可能

通过反问澄清规范用户的输入，让模糊的问题精准化，沉淀成知识

岐义识别 → 精准澄清 → 知识沉淀 → 智能召回

岐义检测
识别未定义口径或不完整条件

强制澄清
最少、最关键的问题强制澄清

结构化转化
存入个人知识库

自动消歧
减少重复问答

岐义识别

用户问题 + 上下文

岐义类型识别

语义岐义 | 细化需求 | 泛化需求 | 指代现象

自动消歧

① 术语表默认无岐义

聚合方式 | 运算方式 | 常识逻辑

② 字段、维值消歧

字段 + 维值

精准匹配 / 编辑距离 → 自动消歧

岐义置信度评估

① 评分计算公式

Ambiguity Score ∈ [0,100]
=(100 - schema_match) × 0.4
+ term_ambiguity × 0.3
+ context_gap × 0.3

② 触发阈值

- ≥80：必须澄清
- 60–79：最小化澄清
- <60：无需澄清

知识沉淀和总结

① 知识沉淀

岐义反问 + 用户补充 → 结构化知识

② 知识检索

Rag → Embedding / 向量数据库 / 召回/精排

一次澄清，长期收益

智能系统将用户澄清答案转化为结构化知识，形成个人与公共知识库，实现持续复用

核心优势

- 最小化强制澄清，避免无意义反问
- 一次澄清，长期收益
- 个人知识沉淀

应用场景

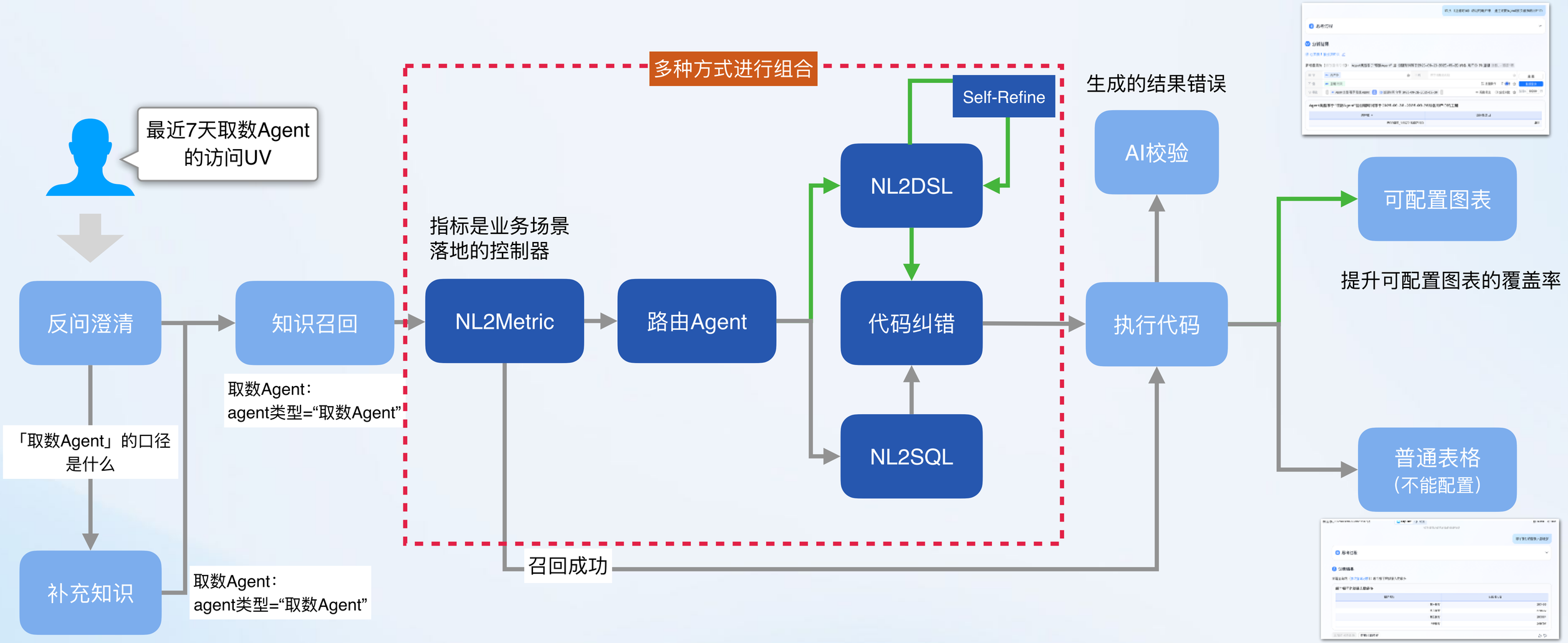
- 领域口径定义
- 时间范围定义
- 指标口径澄清
- 业务习惯定义

产品展示

二、如何提升NL2Code的正确率，不同方案的优劣对比

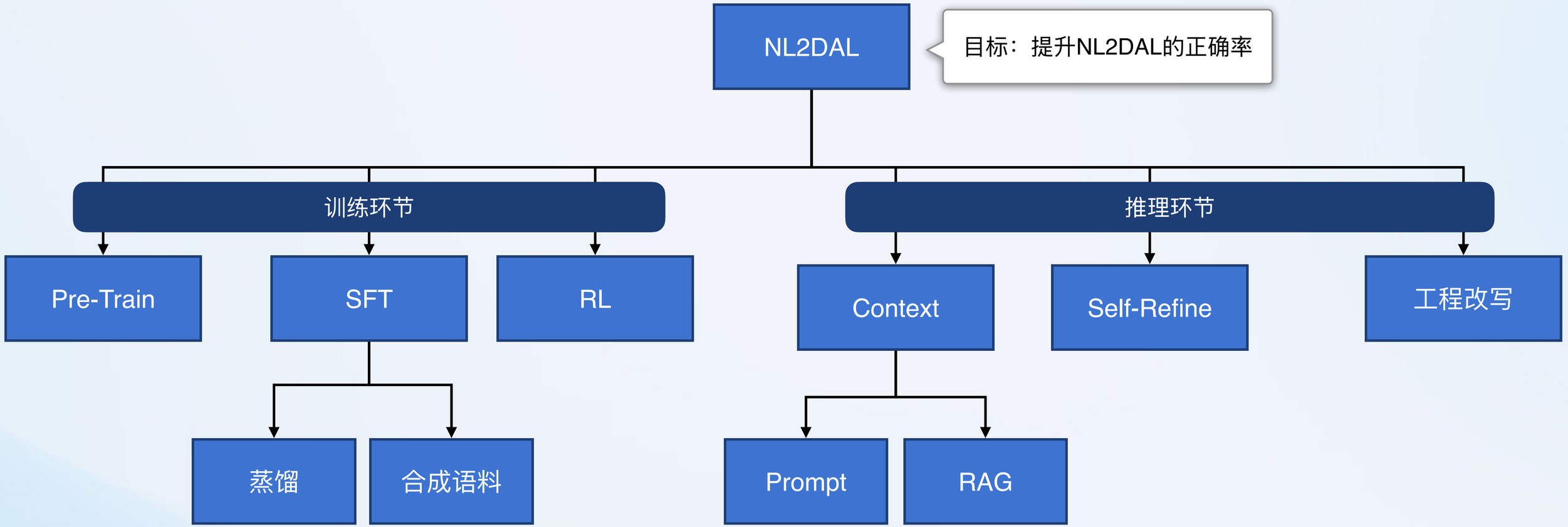
	NL2DSL	NL2SQL	NL2Python	NL2Metric
优点	可以转化为产品交互，进行相关配置	正确率次高 能实现大数据量查询	正确率最高	从生成转化为召回，正确率可控
缺点	正确率上限低	无法转化成图表配置	无法查询大数据量 无法转化成图表配置	正确率取决于指标的覆盖率，无法广泛使用。

多路由方案提升NL2Code的正确率



NL2DSL的正确率，如何提升?

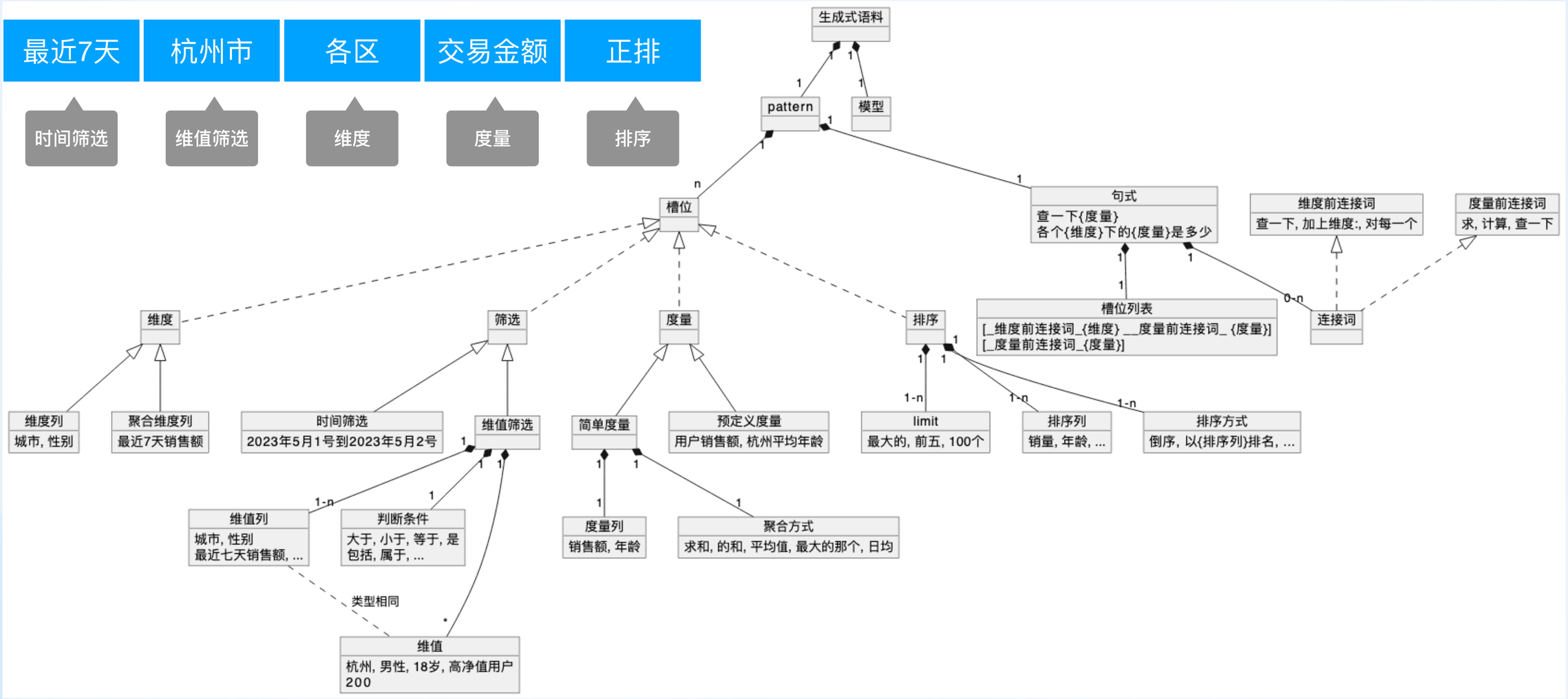
如何提升NL2DSL（在蚂蚁是NL2DAL）的正确率



SFT，一场漫长的炼丹之路，最终都回到一个命题：如何得到高质量的语料

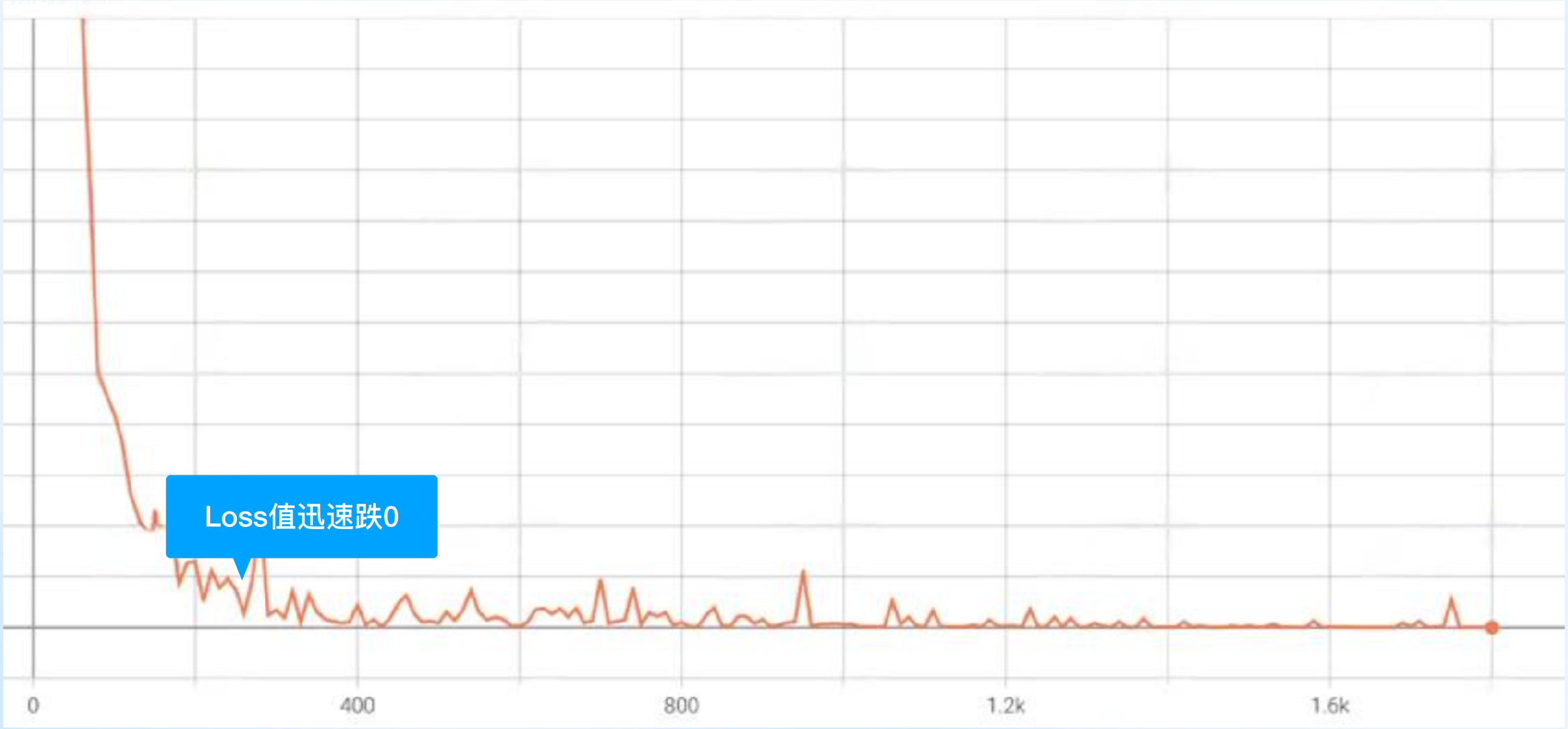
语料来源	合成语料	线上语料	手写语料	大模型生成
优点	成本低，效率高， 可以快速生成大量语料	来源真实， 直接反映现实的情况	有人味，人比较容易理解	成本低，效率高 可以大量生成语料
缺点	质量低，天花板低， 模型训练后泛化能力差	需要对客，场景有偏 需要大量的标注和筛选	标准难统一，成本高、效率低 对于人员要求高	需要大量标注 场景受限

我们的案例：通过原子结构化的方式组装NL2DAL的语料



把取数提问结构化，并拆解成原子结构，类似于搭积木的方式拼接成目标语料

合成语料，语料模式单一，导致模型训练过拟合



增加语料的种类和多样性，对SFT有非常显著的效果

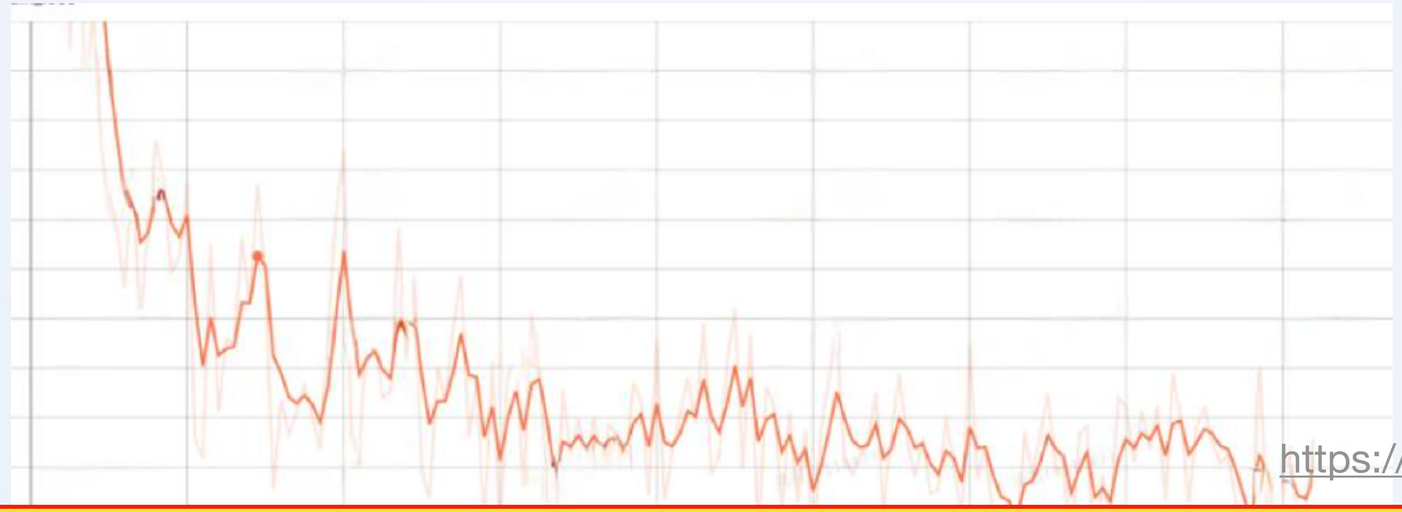
语料类型

合成语料

CoT语料

函数文档

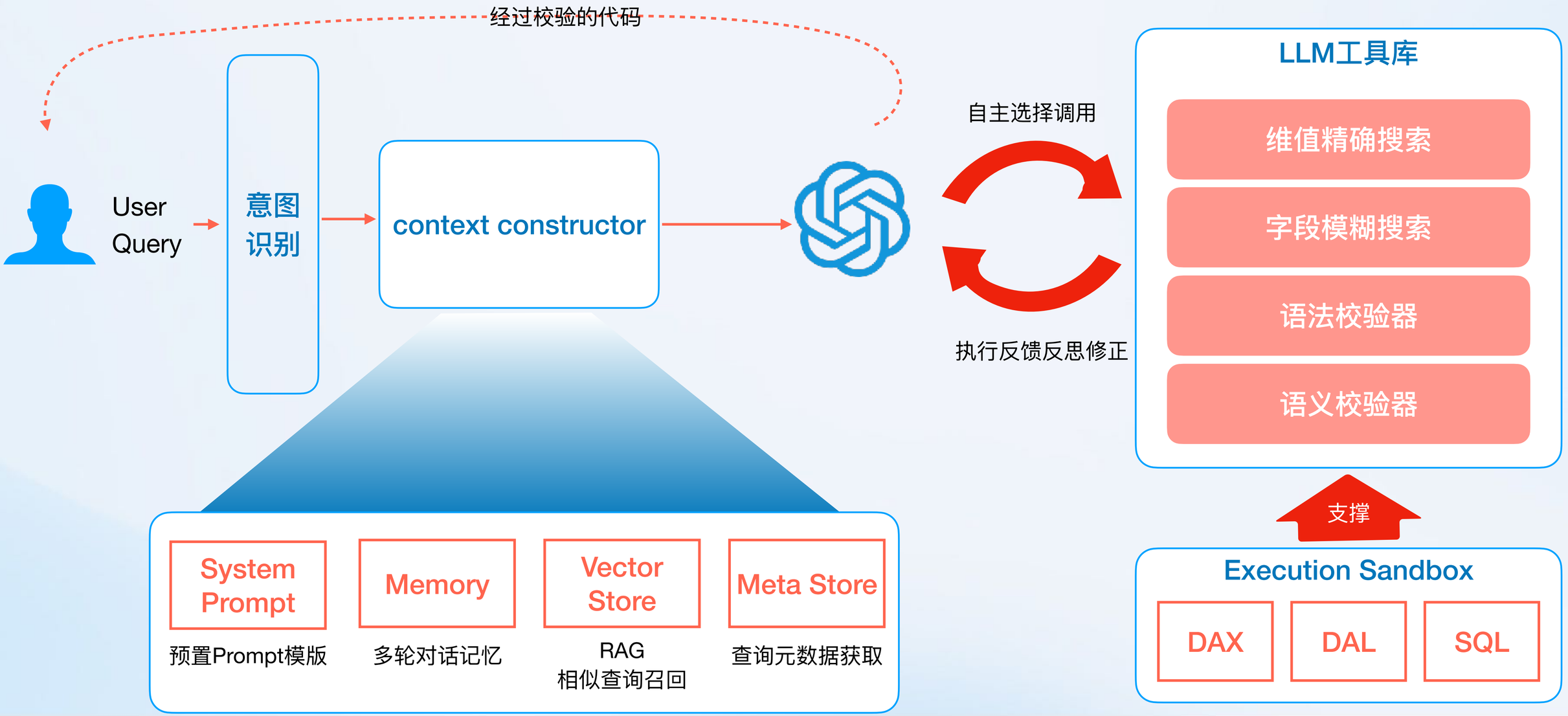
训练效果



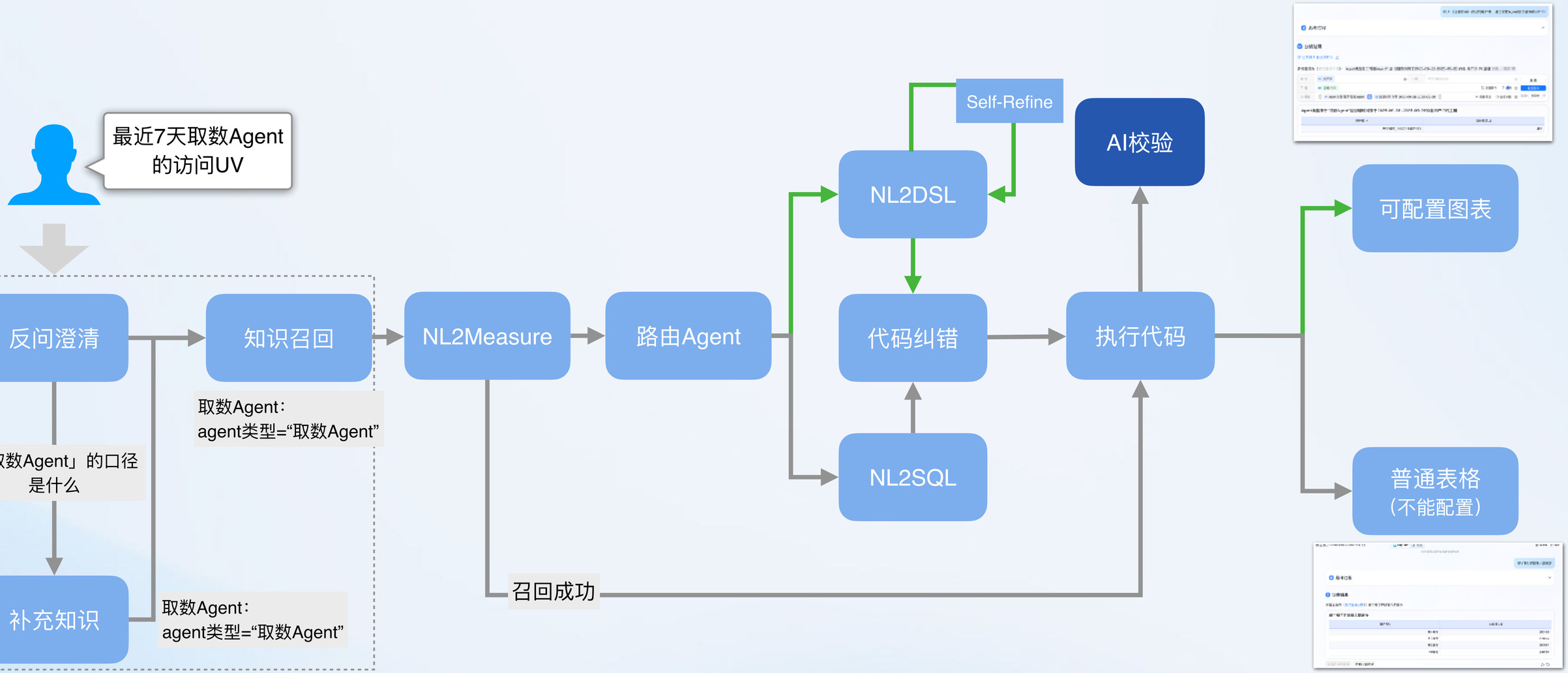
https://arxiv.org/pdf/2305.11206

通过多种语料混合，增加语料的多样性，提升语料的质量，可以提升模型训练效果

在推理阶段，通过Self-Refine提升NL2DAL准确率，正确的反馈可以帮助模型修正结果



三、如何对大模型生成的结果进行校验



三种校验方式，帮助用户判断大模型生成的结果是否可信

	SQL人工校验	AI校验	产品配置人工校验
优点	通过代码逻辑可以直接判断结果是不是OK	普通用户就可以根据这个校验结果判断是否可用	普通用户可以通过配置来判断和预期是否一致
缺点	需要人为校验，需要较高的专业度	会出现误判	覆盖率不高（复杂查询难以还还成配置）

最终这三种方式都会提供给用户，用户根据实际情况进行选择，多方校验能保证结果更准确。

AI校验，用另外一条路径去校验当前结果是否正确

大模型能否找出自己的错误？类似问题是，体育老师不懂数学，没有老师教的话，自己不会知道题目哪里算错

1.采用能力强的大模型进行校验

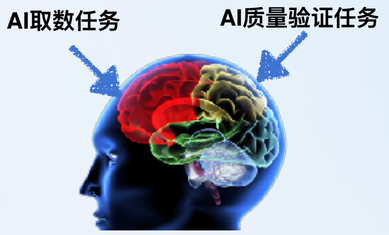
采用DeepSeek R1能力强的大模型作为校验模型
避免弱的模型本身有缺陷，强模型会更可信

2.异构知识避免大模型在两个任务中共享相同的模型缺陷

类似换一种解题思路，通过取数SQL和关系代数（执行计划）来校验与用户诉求是否匹配。避免大模型有缺陷，导致text2dal有问题反过来再通过dal来验证也可能有问题，共享相同的大模型缺陷

3.双路径验证

SQL和关系代数（执行计划）都是通用知识大模型本身对其都有很好理解，多一个路径来检查



使用大脑不同的知识区域，进行思考

大模型提示词

用户输入的诉求 + 数据集信息 + 领域知识 + 取数SQL + 关系代数（执行计划）

用户输入：
评测集中评测题数量是多少?

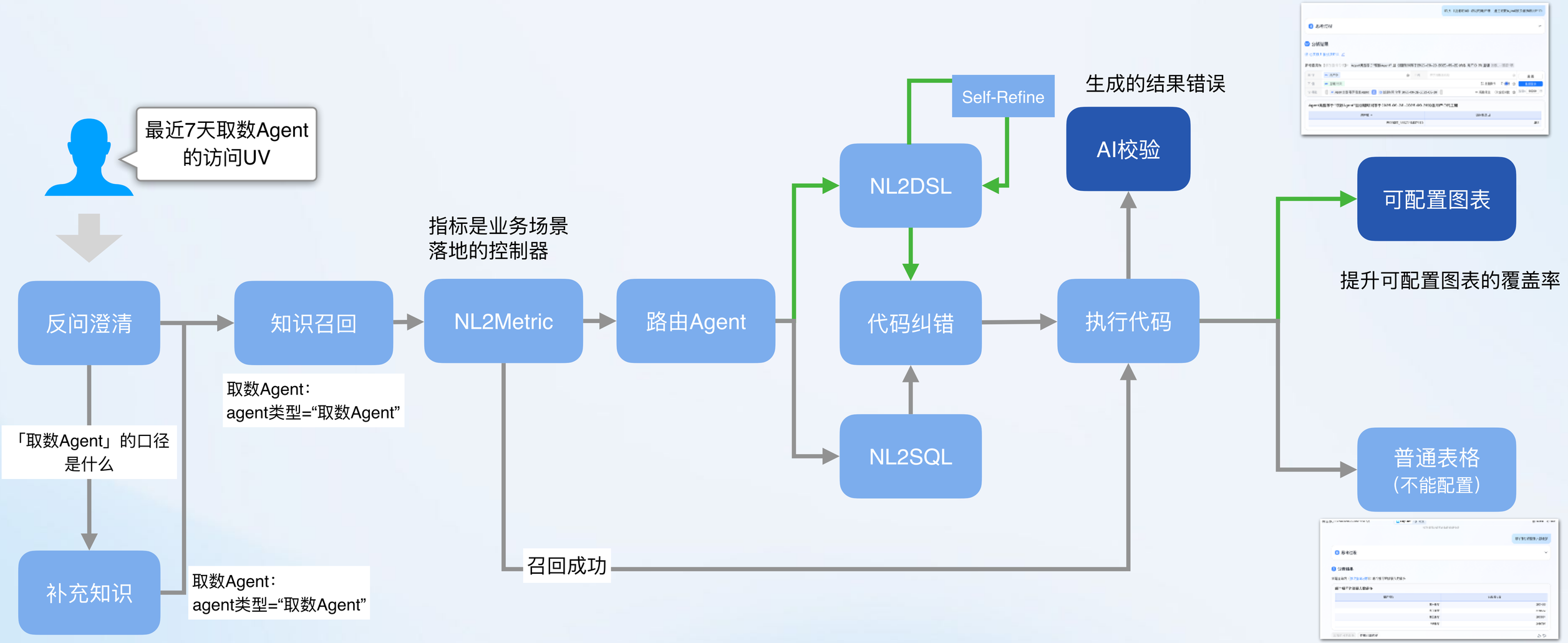
```
SELECT cast(`table0`.`column005` AS double)
FROM `alipaytech_dev`.`TMP_xlsx_668608_data175`
GROUP BY  cast(`table0`.`column005` AS double)
limit 50000
```

```
Project[id: 8]
|   Fields: [取数评测集评测记录.评测总题数]
|   取数评测集评测记录.评测总题数:= $0
└ Aggregate[groupBy: $0, id: 7]
    |   Fields: [取数评测集评测记录.评测总题数]
    └ Project[id: 6]
        |   Fields: [取数评测集评测记录.评测总题数]
        |   取数评测集评测记录.评测总题数:= Cast($5, DOUBLE)
        └ TableScan[table: 取数评测集评测记录, id: 3]
                Fields: [取数评测集评测记录.评测轮次, 取数i
```

大模型判断
是否满足用户取数诉求

目前内部评测集效果，校验准确率91.7%，线上真实情况评测效果，校验准确率88%，并还有优化空间可继续提升

四、当大模型生成的结果错误时，用户如何继续操作



当前有三种情况，都会导致用户无法通过自然语言完成生成结果的工作

复杂场景通过自然语言描述效率低

比如留存率、字段封箱

大模型生成结果有幻觉

结果验证是一个难点

领域知识覆盖率低导致错误

导致大模型无法根据有限信息进行判断

那么用户如何去确定模型生成的结果是正确的?

将自然语言转化为图表配置，用户根据具体的配置对结果进行判断，另外可以交互式操作

用户通过自然语言输入

转化图表配置，用户根据配置判断生成是否有问题，并进行相应调整

结果展示区

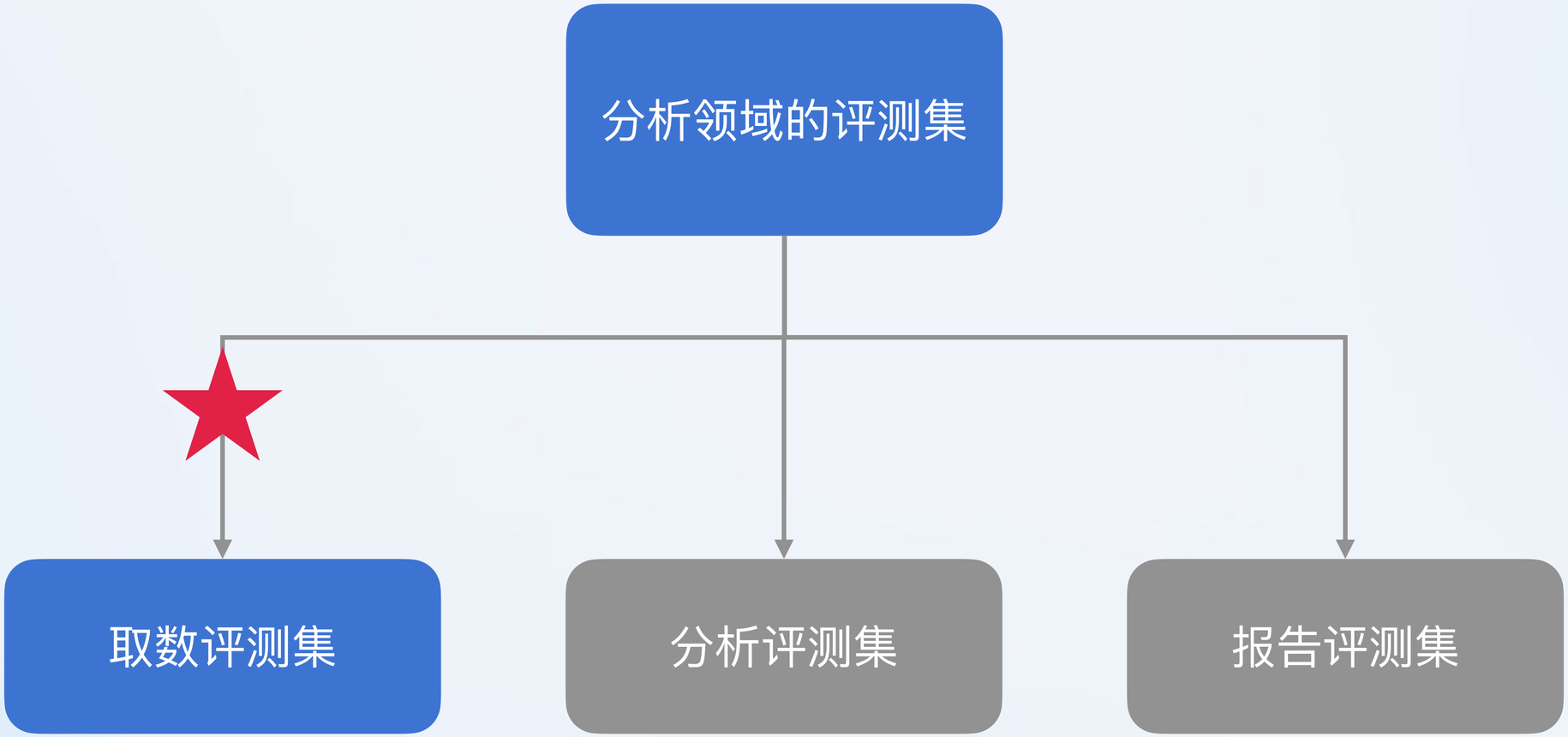
对结果进行校验

即使自然语言生成的结果不对，用户仍然可以根据生成的结果进行相应的调整，用户不需要从头开始

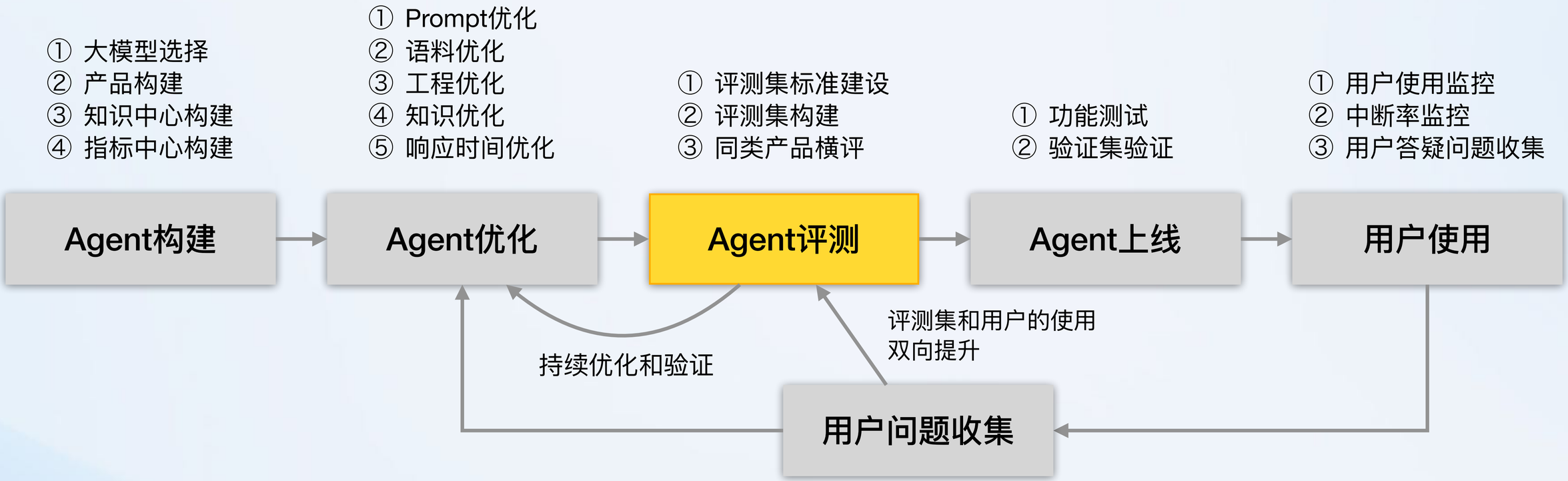
03

评测集构建

构建评测集，正确而漫长的路，需要长时间的积累



完善的评测集可以实现Agent的快速高效迭代，没有评测集就是盲人摸象



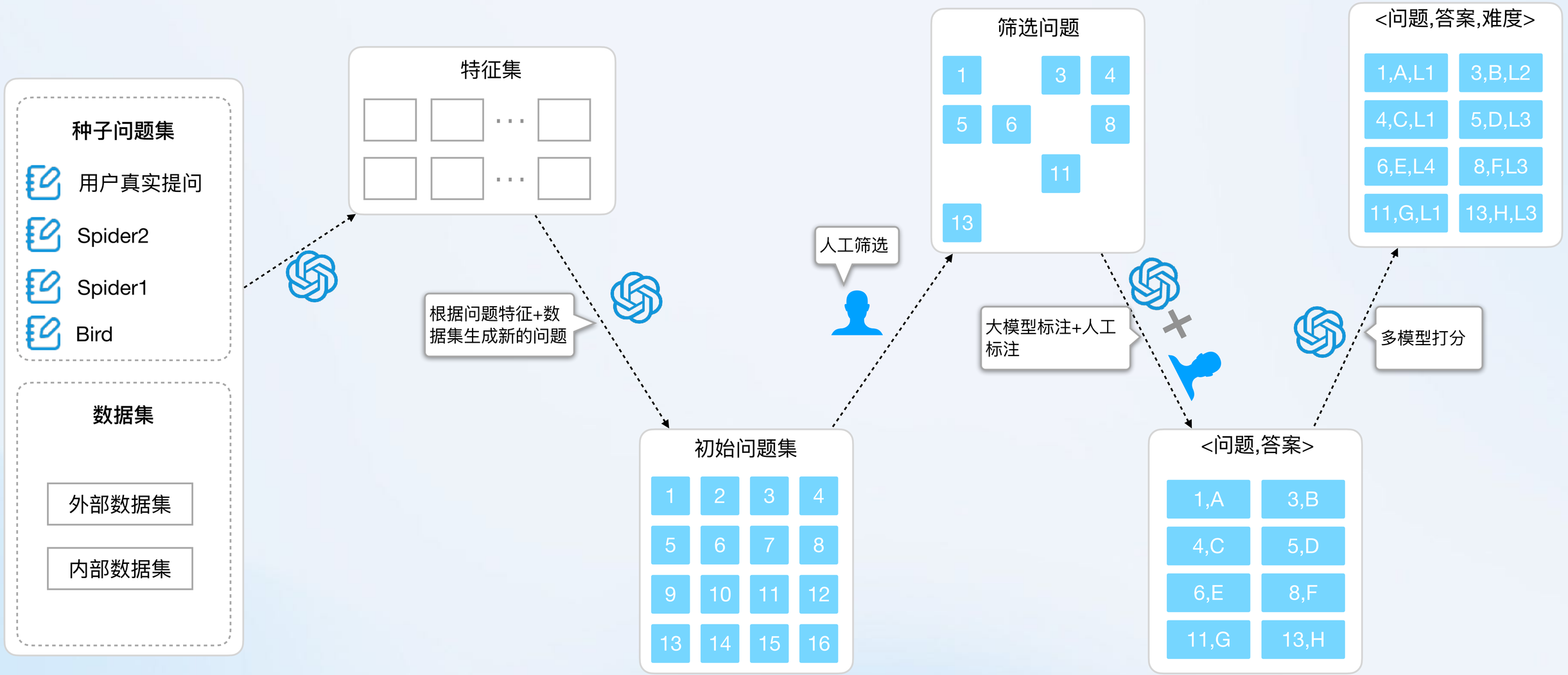
一个好的评测集能指导Agent迭代的方向，Agent的优化可以一直围绕评测集进行迭代。

业界的评测集方案，英文居多，且数据集非常标准，和真实场景不匹配

业界方案	出品方	出品时间	优点	缺点
WikiSQL	Salesforce Research	2017	出品时间早	单一领域、SQL简单
Spider 1.0	耶鲁大学	2018	有难度分级	简单、面向查询
CSpider	西湖大学	2019	中文	翻译的Spider
TableQA	追一科技	2020	中文，表数量多(6000个)	单表、数据量小
DuSQL	百度	2020	中文、带有知识	数据集未公开、未知
BIRD-SQL	香港大学、阿里巴巴	2023	覆盖面广、数据量大、带有知识	标准数据库、英文
Spider 2.0	香港大学、Salesforce Research	2024	覆盖比较全	题目偏难

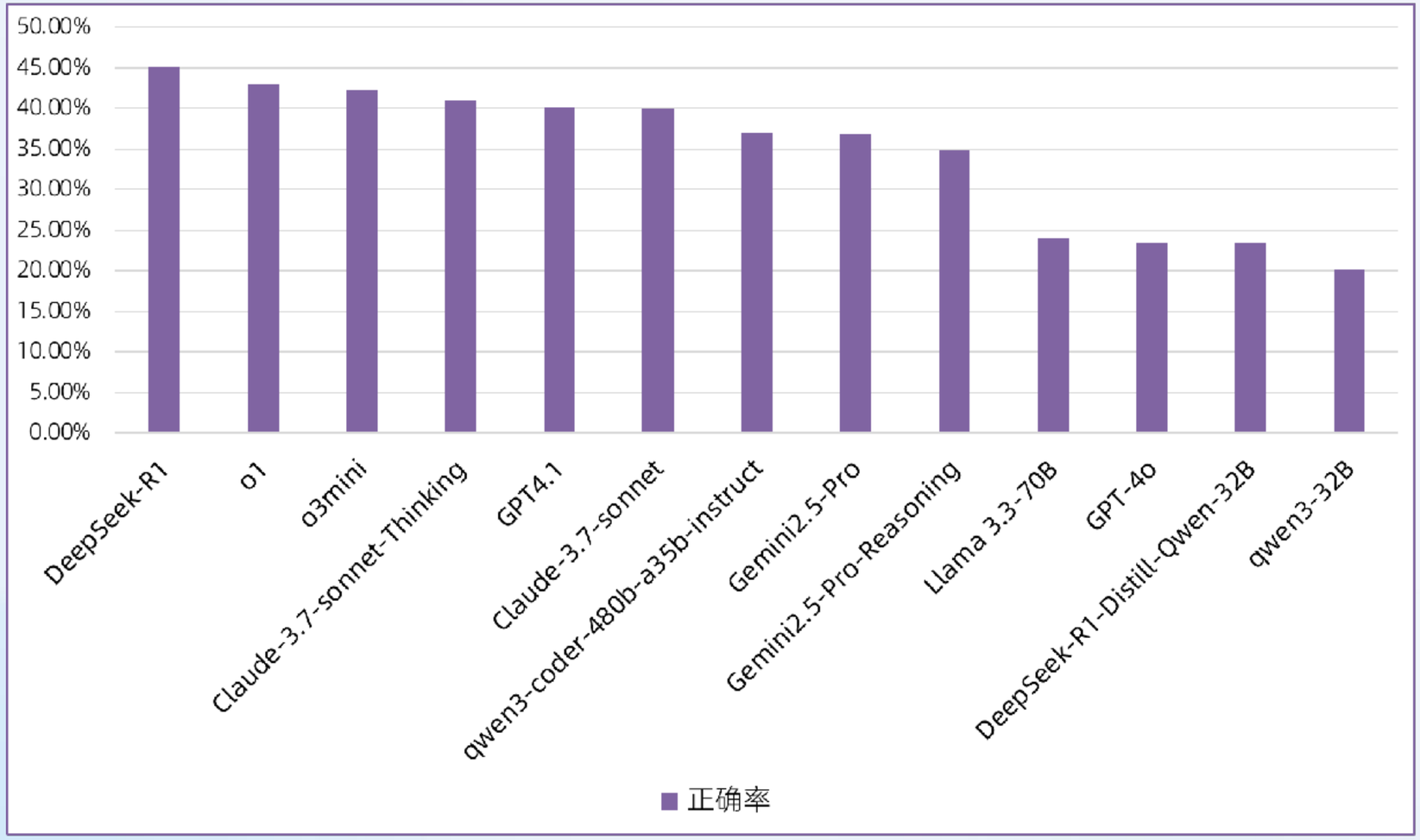
真实场景下的数据集会有各种各样的情况，比如列名不标准，有领域知识等，目前公开的评测集都没有覆盖

评测集问题生成的整体流程：① 生成问题 ② 确定难度



高质量的评测集的关键在于覆盖率，能够随着Agent的发展而发展，不是一个静态的过程

评测集在各大模型的跑分结果，考虑模型的发展，评测结果整体偏难



模型名称	正确率
DeepSeek-R1	45.2%
o1	43.0%
o3mini	42.2%
Claude-3.7-sonnet-Thinking	41.0%
GPT4.1	40.2%
Claude-3.7-sonnet	40.0%
qwen3-coder-480b-a35b-instruct	37.0%
Gemini2.5-Pro	36.8%
Gemini2.5-Pro-Reasoning	34.8%
Llama 3.3-70B	24.0%
GPT-4o	23.4%
DeepSeek-R1-Distill-Qwen-32B	23.4%
qwen3-32B	20.2%

评测集及评测框架已经开源，大家可以进行自行评测

评测体系构建及应用

对外开源	• 域外：500道题（已开源） • 域内：226道题（待开源）

对内应用	• DeepInsight各Copilot智能体评测 • 端到端的评测产品化构建

产出2篇专利

- 《一种基于SQL计算特征和语义表达特征的标注方法》
- 《一种基于SQL计算特征和语义表达特征的评测集生成方案》

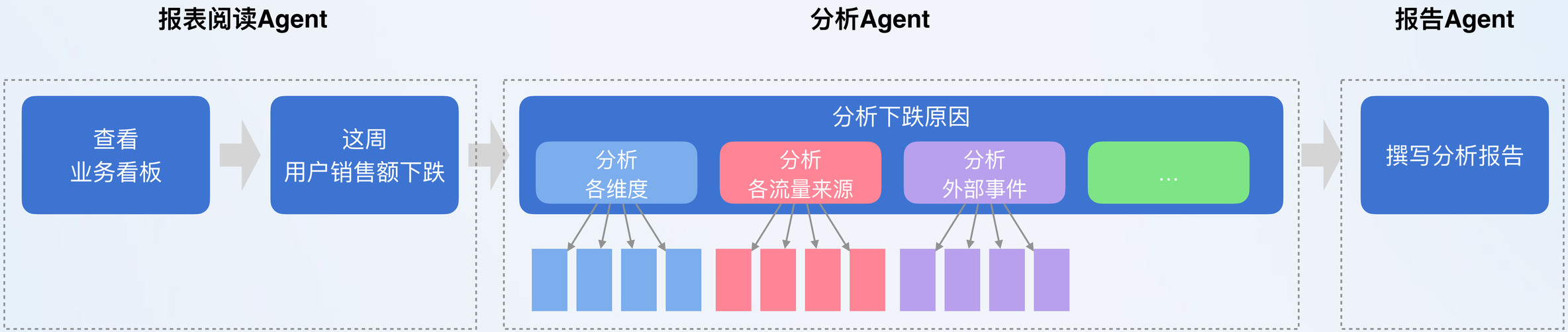
开源建设

- 开源地址：https://github.com/eosphoros-ai/Falcon
- 论文《Falcon: A Chinese Text-to-SQL Benchmark for Enterprise-Grade Evaluation with SQL- and Semantics-Aware Annotations》（待发表）
- 蚂蚁评测集使用说明文档：https://www.yuque.com/alanchen-nd2og/kbfv26/olmdxmoe7s6uogcg?singleDoc#

04

长程任务探索

具体场景中，长程任务是一系列简单任务组合的结果



① 下跌的阈值如何确定依赖经验校验

① 分析的数据不全，比如分析维度、外部事件，那需要重新加工
② 没有相关分析模式，需要人工探索
③ 数据不规范，需要大量的人工经验介入

① 按照一定的分析模式+数据组织文档

需要完成长程任务，核心仍然取决于单个Agent的正确率，需要不断打磨每一个Agent的正确率

DeepInsight端到端数据分析Agent产品——DataSage

全新产品

DataSage

产品定位：新一代数据分析Agent，你身边的数据分析专家。

- Data（数据）+Sage（智者、专家、古希腊智慧女神雅典娜别称）
- 直接传递“数据领域的智慧专家”形象
- 寓意产品能像“智者”一样提供洞见
- 产品提供端到端数据分析能力，需求执行全流程AI自主，就像你身边的可靠数据分析专家

项目实战

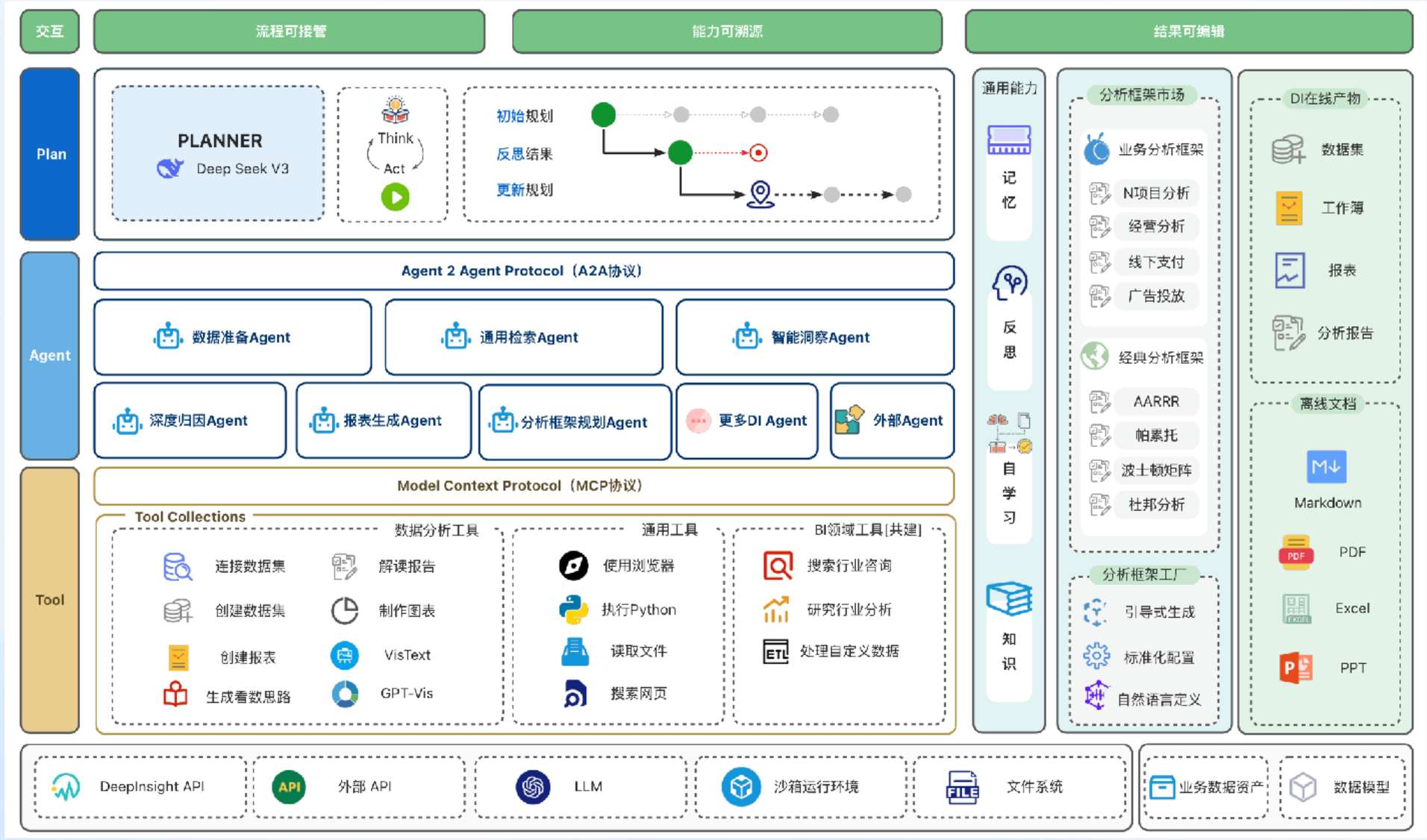
BI + DataSage

Transforming Data into Intelligent Insights

蚂蚁集团商业智能BI团队+DeepInsight团队联合共建

从最需要BI智能化的场景出发，解决最迫切、最高频的BI痛点问题，从实战中淬炼产品

长程任务的技术实现： Planner + MCP + A2A协议



要落地，ReAct更好还是Workflow更好?

案例：如何通过任务拆解实现一个归因分析及生成最终的分析报告

分析思路越来越多?

分析思路越来越复杂?

专业用户：
高管/业务同学

Q：4月13对比3月13的支付成功率指标下跌进行归因分析，并产出分析报告

匹配到BI/解决方案供给的分析框架

以支付类：商户支付成功率归因分析报告为例

①融合了分析框架，得到用户满意的带用户业务知识、以及分析框架的执行规划

DataSage以产品+业务共建方式进行落地，探索长程任务的最佳实践

DataSage 既能接入DeepInsight已有数据分析能力，又能快速集成外部工具/智能体

快速的实现业务场景落地

DeepInsight

数据集 | 工作簿 | 报表 | 自助分析 | ...

报表制作Copilot | 自助分析Copilot | 报表Copilot | ...

深度分析 | 图表制作 | 取数 | 问答 | ...

创建数据集 | 编辑图表 | 性能诊断 | ...

DataSage

BI、蚂蚁内部、互联网

A2A

行研报告Agent | ERP Agent

... | 文件管理Agent

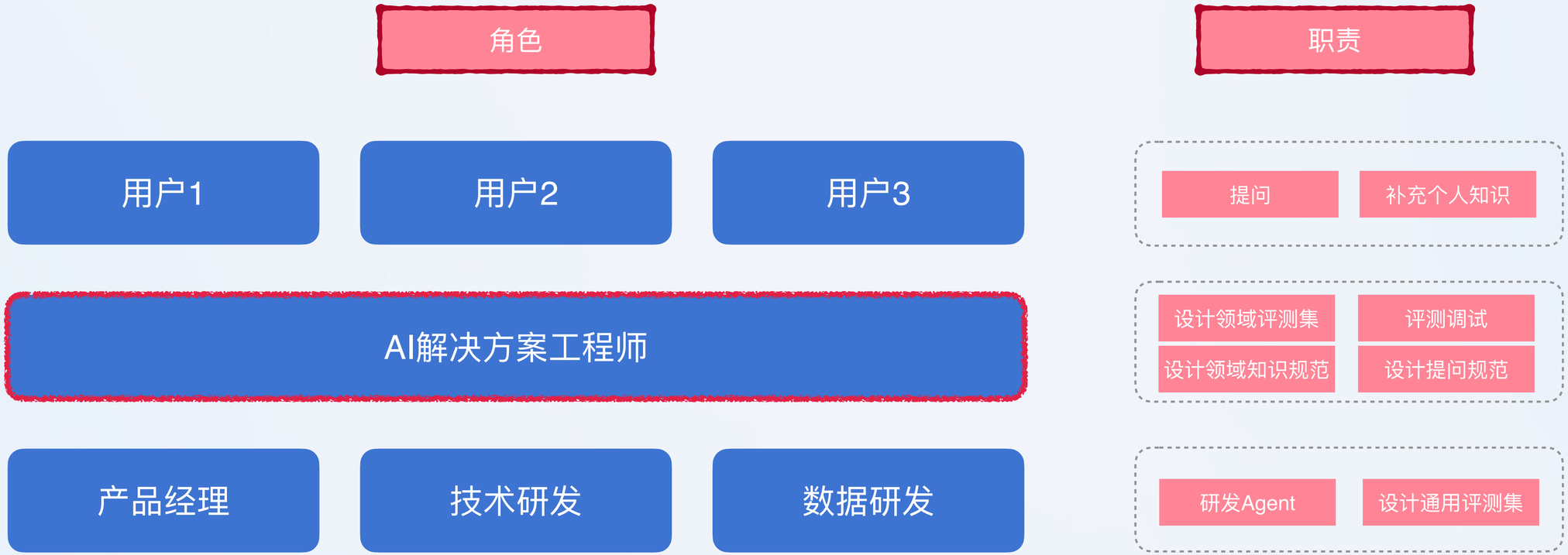
MCP

... | 搜索工具 | PPT生成工具

BI分析工具 | BI分析思路

BI沉淀小工具

在AI+BI这套新的范式下，需要一套新的生产关系进行配套



通过AI解决方案工程师来弥合业务和大模型之间的GAP，让大模型更好的服务用户

QCon
全球软件开发大会

InfoQ 极客传媒

Q & A

有兴趣可以微信交流