

AgentLoop[★]

建设企业级 Agent 进化数据飞轮实践

Agent 持续进化 · 让智能体越用越聪明

演讲人 · 夏明

阿里云 · AgentLoop 产品负责人

从 Demo 到生产：Agent 需要一个能持续进化的数据飞轮

“

Agent 不是写完 Prompt 就上线的一次性产品，
而应是一个能持续观测、持续评估、持续优化的
闭环进化系统。

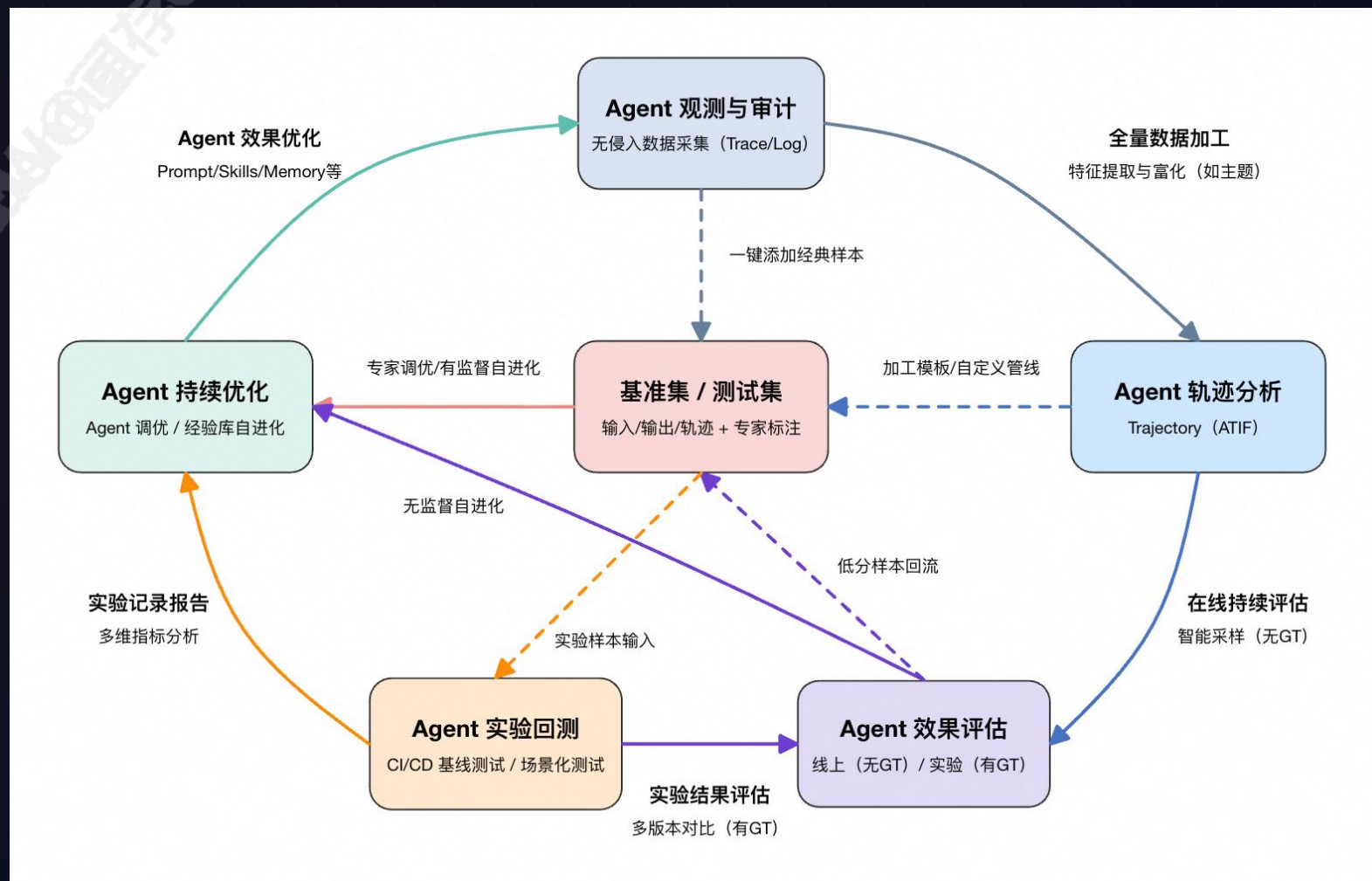
—— 演讲核心观点

-- 本场分享 · 主线

AgentLoop MVP 闭环： 五环驱动持续进化

数据采集 · 资产沉淀 · 效果提升
→ 正向增强循环

- 1 Agent 观测与审计
- 2 Agent 轨迹分析
- 3 Agent 效果评估
- 4 Agent 实验回测
- 5 Agent 持续优化



Agent 观测与审计

LoongSuite 开源可观测套件 · 4 语言零代码插桩,全栈 Agent 数据采集

1 主流框架覆盖度 OTel 原生 · 开源自动插桩 · 4 语言全栈 · 去重后约 45 个框架

Python22 框架

LangChain

LangGraph

LlamaIndex

OpenAI Agents

AgentScope

Qwen-Agent

Agno

Java10 框架

Spring AI Alibaba

Spring AI 4.x / 5.x

OpenAI Java SDK

MCP

Go16 框架

Eino

Coze-Studio

LangChainGo

Google ADK

Volcengine Ark

Dify · MCP

Node/PilotBeta6+ 框架

Qoder

Claude Code

Cursor

Codex

QwenCode

OpenCode

- 零代码改造自动插桩,业务无侵入

• 覆盖国产生态:AgentScope/Qwen-Agent

• AI 编码 Agent 系统性采集方案

零代码路径 Go 编译时 AST Java 字节码增强 Python OTel Agent JS/TS Hook eBPF 探针

2 GenAI 语义规范 · 字段级深度采集 三层规范融合 · 55 个 GenAI 字段覆盖最全

OTel GenAI社区开源标准,含阿里贡献的 step / MCP Span 扩展

LoongSuite Schema采集层扩展: session / turn / cost 等专属字段

AgentLoop产品侧数据契约,面向后端存储与消费统一对齐

84%有效字段覆盖率 · LoongSuite Pilot 领先,竞品最高仅 51%

Pilot84%

Arize51%

LangSmith51%

Langfuse31%

独有字段

ReAct step.id 推理还原

finish_reasons 截断检测

messages_delta 存储↓90%

Agent 轨迹分析 Trajectory

原始 Trace/Log → 结构化 ATIF · Pipeline 自动化数据管线 · 多模块联动



ATIF 结构化格式

FORMAT

Agent Trajectory
InterchangeFormat · 统一轨迹
协议



Pipeline 数据管线

AUTO

过滤 / 去重 / 采样 / 聚类 /
LLM 语义加工自动化流水线



语义检索

VECTOR

基于轨迹语义向量，秒级召回
相似执行链路



分析与聚类

CLUSTER

自动归类相似轨迹模式，沉淀
典型行为画像



联动观测归因

PERF

性能热点 → 并发瓶颈

COST

Token黑洞 → 无效调用



联动审计攻击链

SECURITY

PII 泄露 / 高危行为 / 攻击链
全路径审计与检测



联动轨迹评估

JUDGE

作为 Agent-as-a-Judge 评估
输入，量化轨迹质量



联动数据集

DATASET

清洗轨迹沉淀进基准集 /
测试集，资产复用



Trajectory 轨迹飞轮 · 让数据 **越跑越多、质量越用越高** · 联动观测 / 审计 / 评估 / 数据集 四大下游

数据中枢：基准集与测试集

"Dataset 是 Agent 的题库 – 持续刷题，越练越强"

01 / SOURCE

多路数据来源

— CLASSIC · 经典三路

观测 / 审计数据
来自环节 1 · Trace / Log

Trajectory 轨迹数据
来自环节 2 · ATIF

评估低分结果
来自环节 3 · Bad Case

— OTHER · 更多导入路径

+ 历史数据集

+ 公开数据集

+ 业务系统数据

02 / PIPELINE

数据清洗管线

六道工序 · 沉淀 Ground Truth

过滤

去重

特征提取

语义富化

标注模块 · ANNOTATION

LLM 标注

人工标注

03 / DATASET

题库 · Ground Truth

Agent 的"四类题型"

业界经典分类：Benchmark · Adversarial · Regression · SFT/DPO

● 评估基准集 GOLDEN
全面衡量 Agent 综合能力

● Bad Case 集 HARD
攻防对抗 · 边界场景

● 回归测试集 REGRESSION
CI/CD 门禁 · 确保不退化

● 后训练数据集 POST-TRAIN
SFT/RLHF 微调素材

下游用途 → 实验评估 · 持续优化 · 回归门禁

Agent 效果评估

有 GT / 无 GT 双通道覆盖全场景

| 有 GT 评估

- ✓ 基于基准集回归评估
- ✓ Ground Truth 对比打分
- ✓ 量化准确率 / 一致性

适用场景

版本回测、门禁拦截

| 无 GT 评估

- 🕒 线上实时评估（无需标注数据）
- 🕒 Agent-as-a-Judge 深度打分
- 🕒 回放轨迹 + 工具复核 + 多步推理

适用场景

线上巡检、主动报警

★ **AgentLoop** = Agent-as-a-Judge → 评估器本身就是 Agent，比 LLM-as-a-Judge 更接近人类专家评估效果

Agent 实验回测

CI/CD 质量门禁 · 让变更必过回测



实验计划

锁定变量与环境配置模板

确保多次迭代一致性



Playground

双模式：Agent 端到端测试 / LLM 单轮调用

配置 Agent 端点或模型+提示词，绑定数据集一键运行



实验记录

完整快照：推理结果 + 评估得分 + 工程指标

支持横向对比分析



基准回测门禁

变更前自动回归测试

拦截 80%+ 变更故障上线

80%+ 变更故障拦截率

100% 实验记录可追溯

数据反馈回基准集沉淀

Agent 持续优化

数据驱动 · 从经验试错到实验验证



Prompt / Skill 调优

高质量评估器 + 数据集实验回测 → 专家调优 / AI 辅助调优
替代经验驱动



记忆库 Memory

跨会话理解用户背景与偏好
越用越懂你

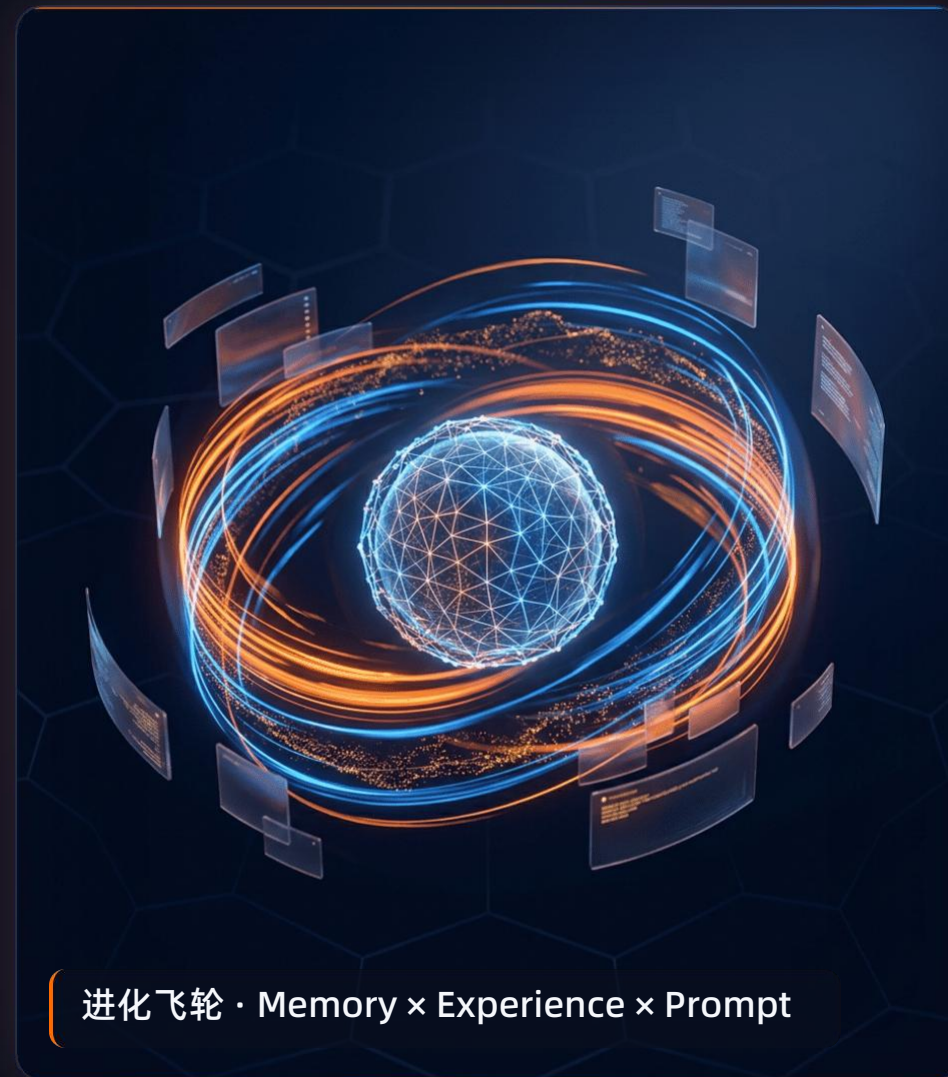


经验库 Experience

从历史执行中学习最佳实践
自动检索注入运行时

优化完成 → 回到环节 1 观测验证 → 闭环循环

环节 5 · 持续优化



进化飞轮 · Memory × Experience × Prompt

全链路闭环 · 企业接入进化飞轮

MVP 五环 · 从 Day 1 到飞轮自驱



自动化成熟度演进 · 从人工运维到 Agent 自进化

飞轮启动 · 数据回灌持续进化



未来的壁垒：谁的飞轮转得快，Agent 就会越聪明

“当 Agent 具备自进化能力，企业 AI 的竞争壁垒将从「谁的模型好」转向「谁的**进化飞轮转得快**」。”

—— AgentLoop 团队

数据驱动 · 不靠经验 → 飞轮越转 · 数据越多 → 数据越多 · Agent 越聪明

谢谢聆听

一起构建 Agent 进化数据飞轮

演讲人 · 夏明

阿里云 · AgentLoop 产品负责人



扫码进入 AgentLoop 控制台
agentloop.console.aliyun.com