

QCon
全球软件开发大会

从云原生迈向智能原生： 蚂蚁中间件进化之路

章耿（余淮）

蚂蚁集团 AI 中间件负责人





自我介绍



章耿（余淮）

- 蚂蚁集团 AI 中间件负责人
- 目前主要负责应用框架、中间件架构、AI 中间件等方向
- 10+ 年中间件工作经验，深耕基础架构领域，致力于通过推进技术架构演进，提高开发人员的研发效能

目录

- 01 智能原生的挑战
- 02 中间件“进化”思考
- 03 蚂蚁 AI 中间件体系
- 04 总结与展望

📍 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

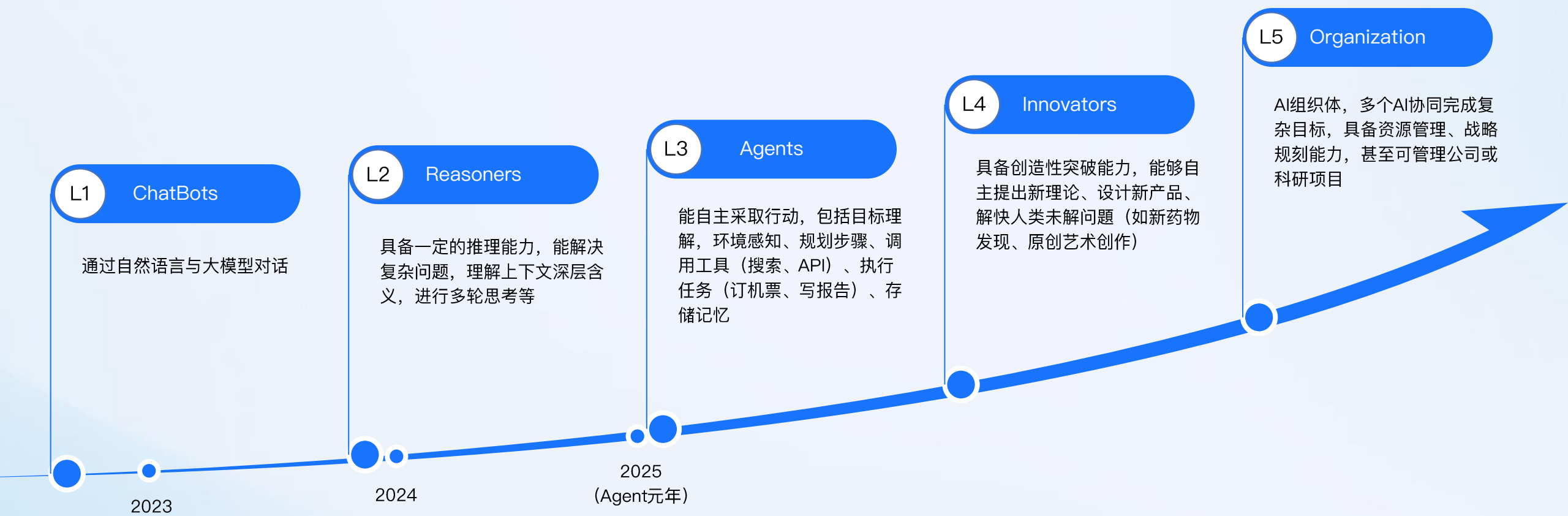


查看会议

01

智能原生的挑战

AI 应用演进趋势



蚂蚁应用架构演进



智能原生应用

智能原生 AI Native

智能原生应用是指围绕人工智能技术为核心进行设计、构建和运营的一类应用，它旨在充分利用和发挥AI的理解、生成、逻辑和记忆能力。

智能原生架构使应用能够具备动态适应、自主优化和数据驱动的核心特征，从而在复杂多变的环境中提供持续智能的服务。

≠ 应用+AI

传统云原生应用

智能原生应用

交互方式

页面/表单/按钮
被动反馈

自然语言/多模态
主动协作

服务实体

微服务

Agent（智能体）

业务逻辑

确定性
静态逻辑
无状态

概率性（泛化）
动态逻辑
有状态

关键存储

关系型数据库

向量数据库

核心算力

CPU

GPU（大模型）

智能原生时代的挑战

新的服务实体

新基础设施

新用户

不确定性

Agent
Tools

多模态
RAG
Memory

非技术同学
大模型

上下文工程
评测
状态管理

研发效率

安全

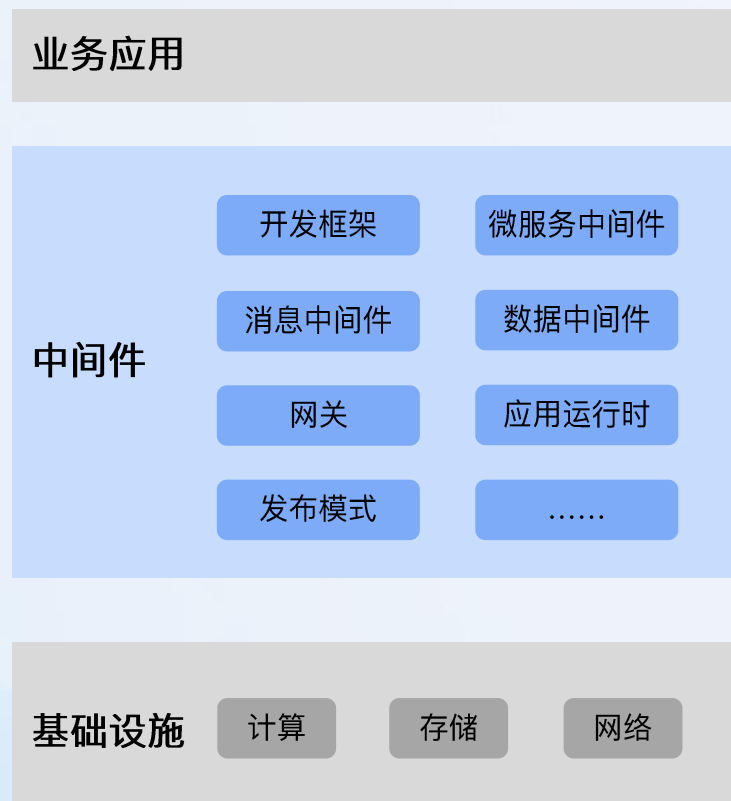
高可用

成本

02

中间件“进化”思考

传统中间件主要解决的问题



解耦
+
粘合

抽象

沉淀通用能力

沉淀最佳实践

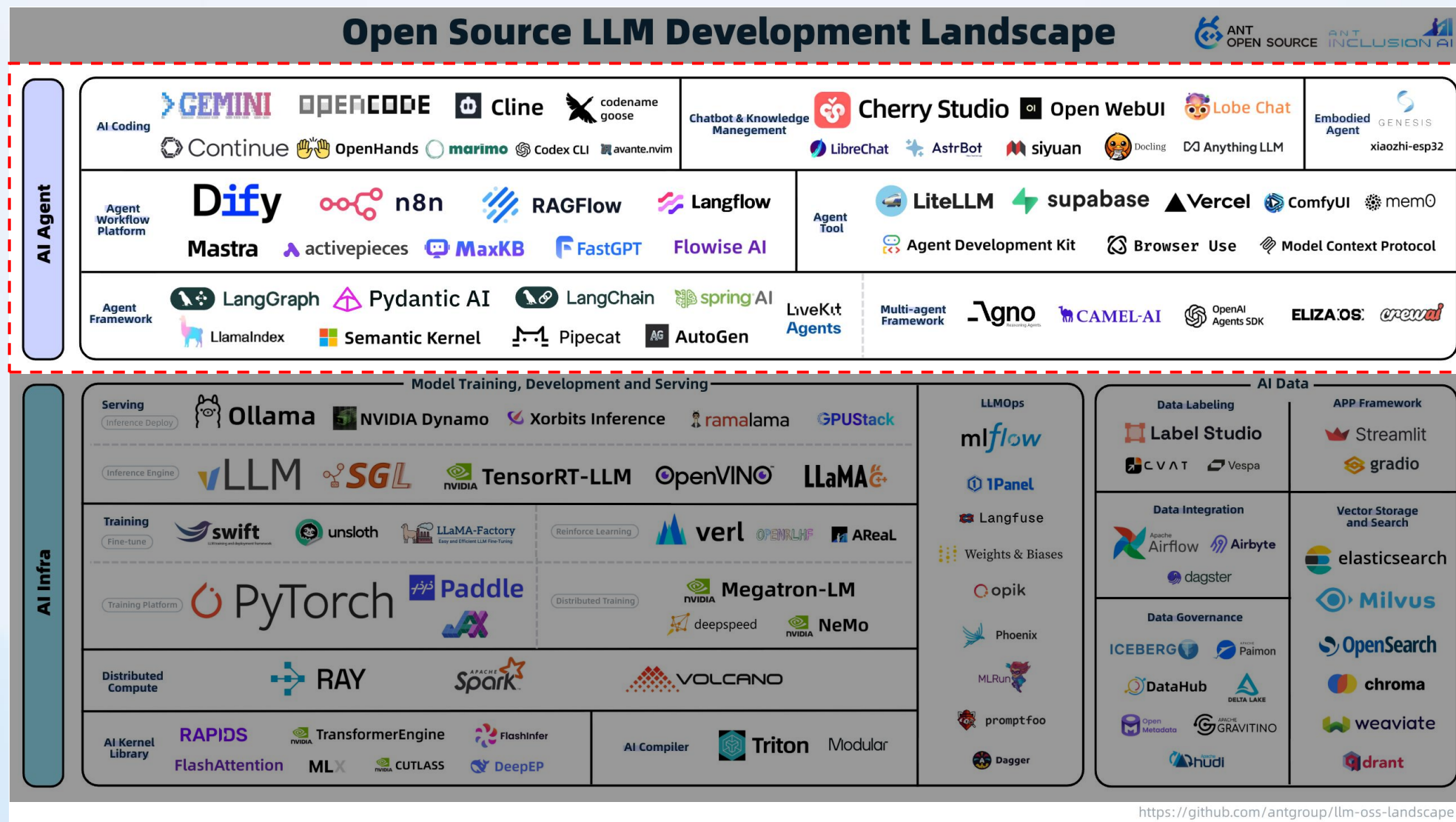
降低研发门槛

提升研发效率

推进技术规模化应用

架构统一演进

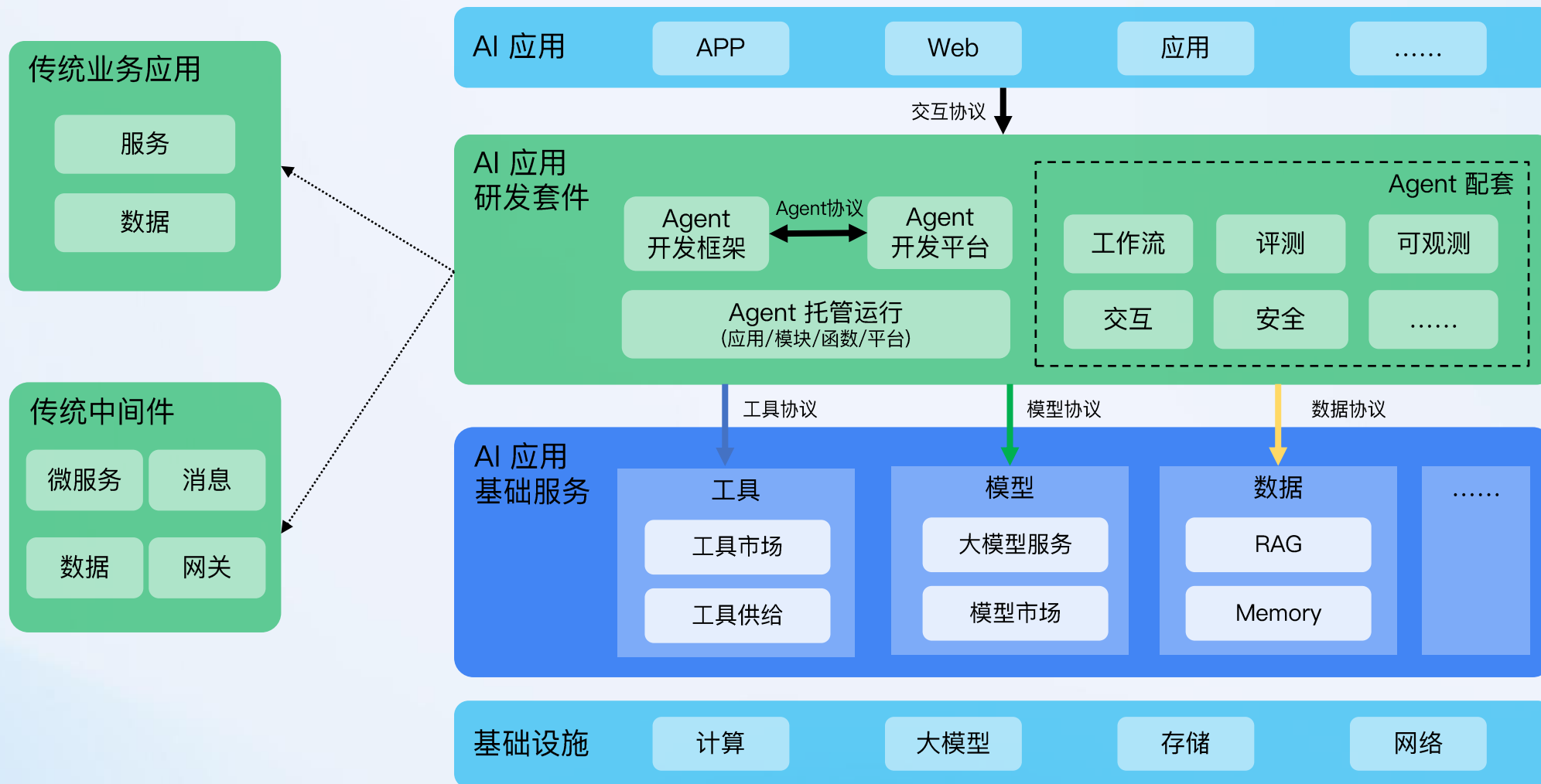
AI 应用研发技术栈全景图



AI 中间件需要解决哪些问题



AI 中间件建设设想



中间件选型决策



- ✓ 同时服务 Java 和 Python 开发群体



- ✓ 拥抱开源，结合内部体系深度自研

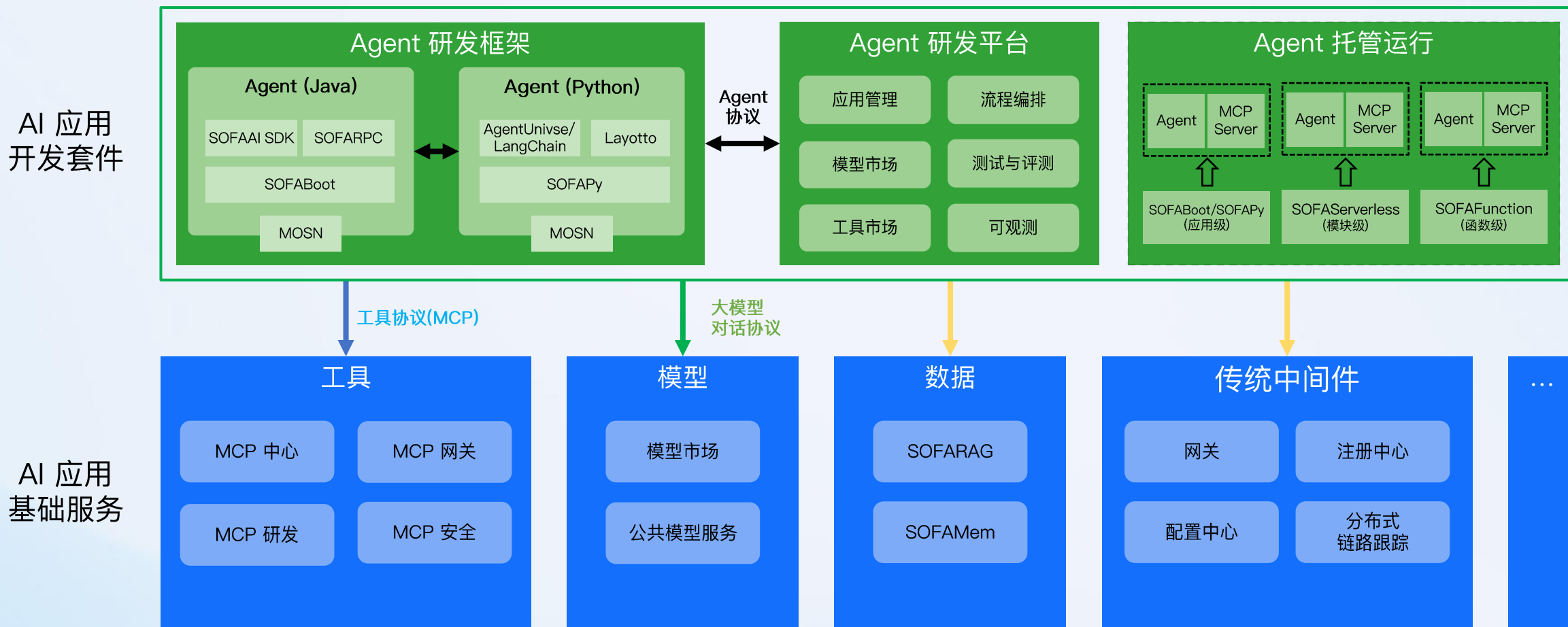


- ✓ 边建设边推广

03

蚂蚁 AI 中间件体系

当前蚂蚁 AI 中间件产品体系





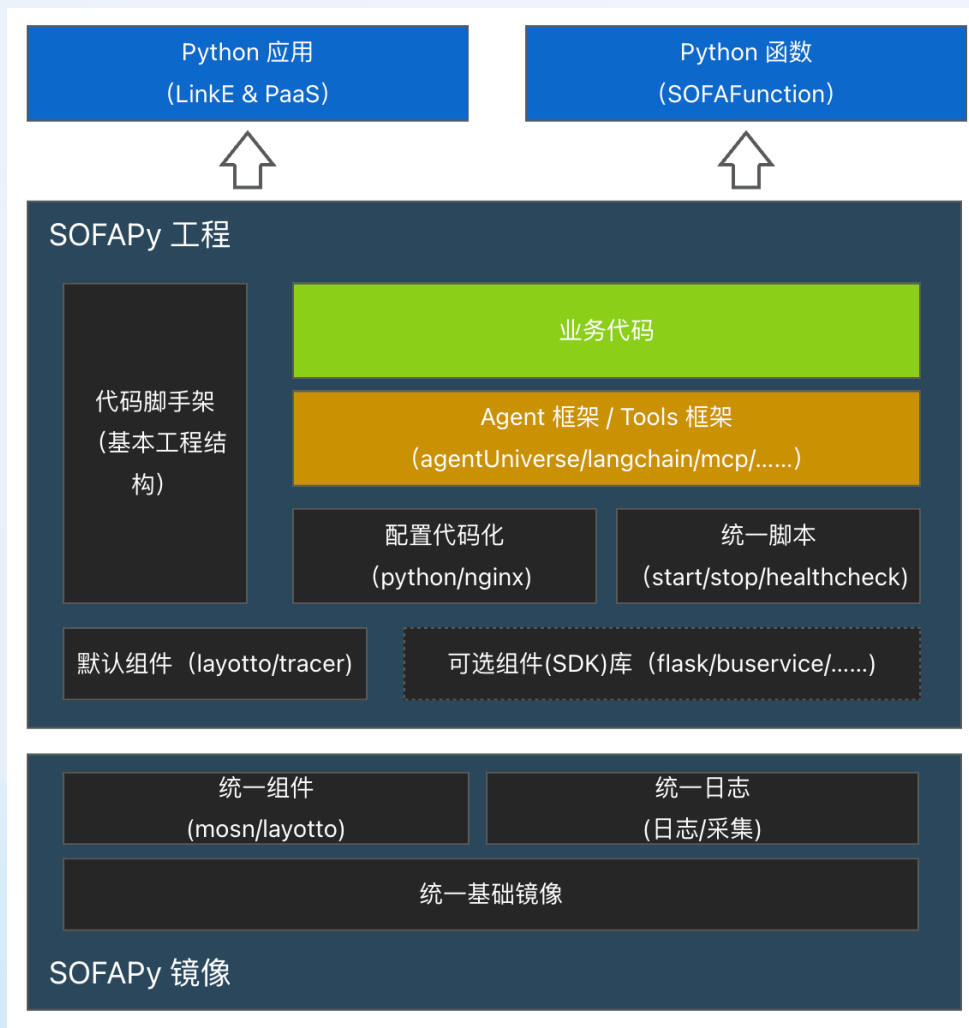
Agent: 研发框架 SOFAAII SDK



SOFAAII SDK 是基于开源的 Spring AI 产品进行构建的 AI Agent 应用框架。它的目标是让基于 SOFA 生态的 Java 应用能够快速、灵活的使用 AI 能力构建 AI Agent。

- 完全适配 SOFA 体系，研发运维上手成本低
- 提供内部和外部多种大模型服务对接能力
- 支持通过 MCP 协议、Function Calling 使用工具
- 内置通用的 Agent 设计模式，例如CoT、意图识别、Multi-Agent、ReAct 等
- 支持应用、模块和函数等部署模式
- 内置安全、可观测、全链路流式等基础能力

Agent: 研发框架 SOFAPy

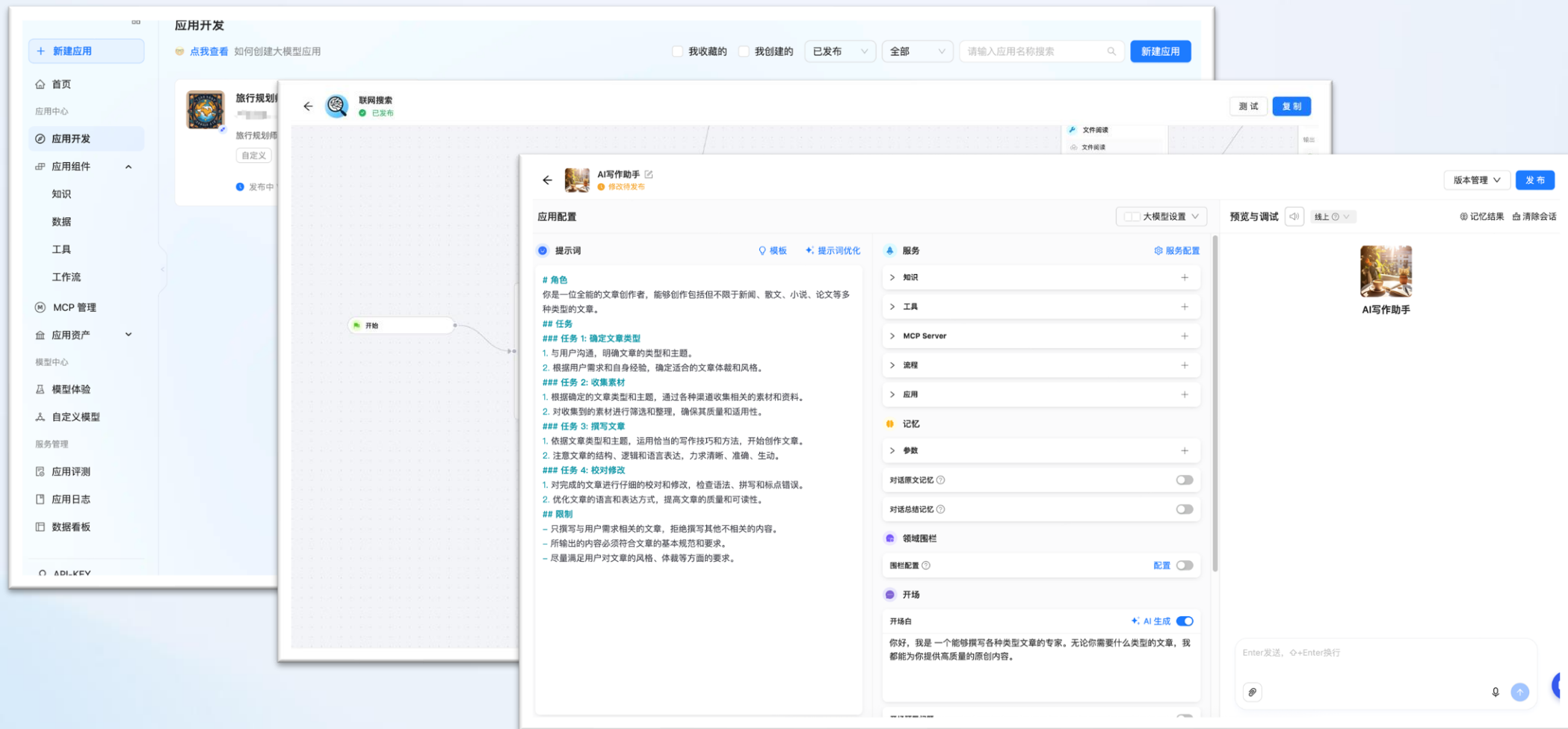


SOFAPy 是 Python 语言的应用开发框架, 旨在帮助 Python 开发者像 SOFABoot 一样快速开发内部应用, 无需关心复杂的中间件、PaaS 等底层能力。

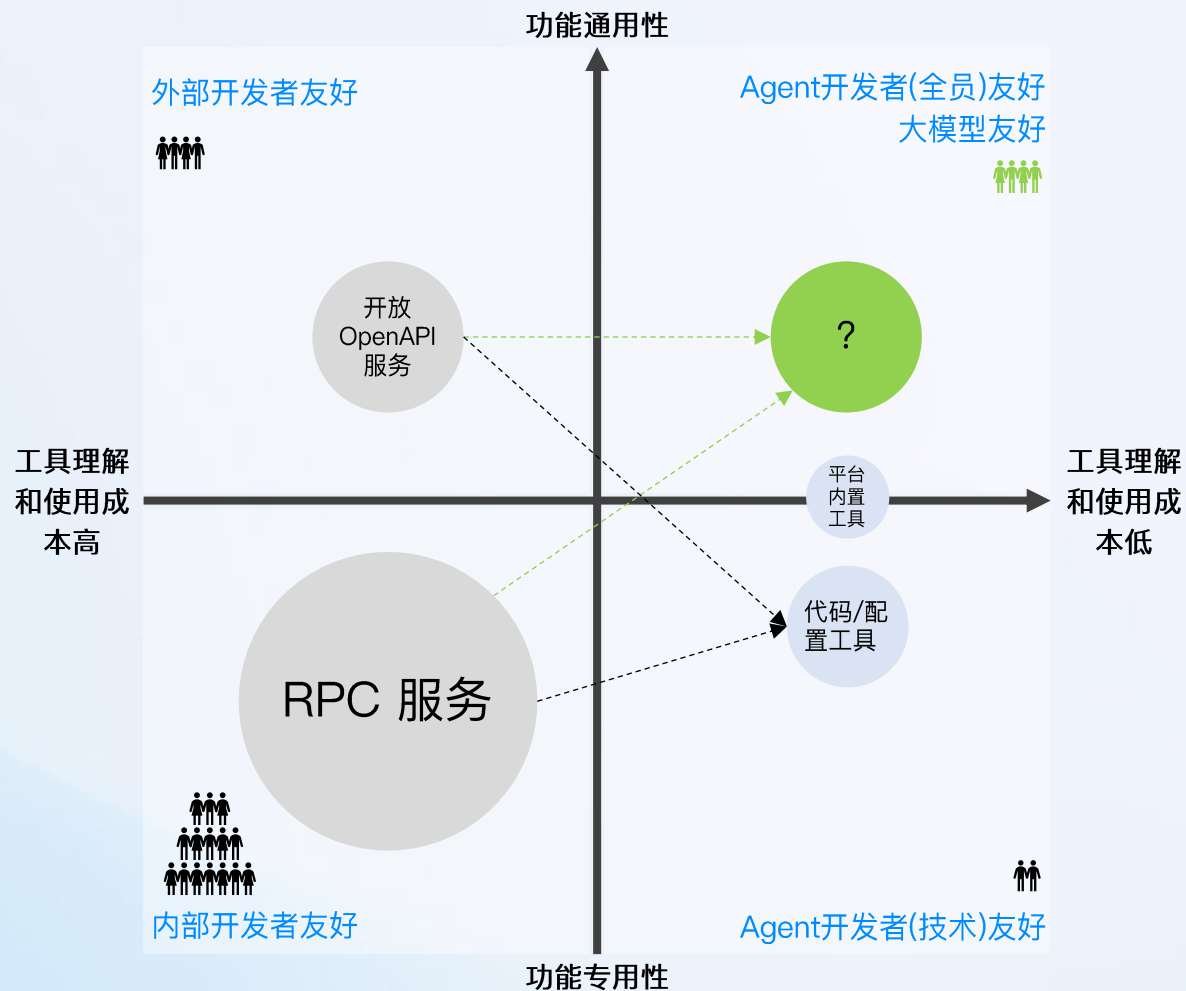
- 原生适配研发运维平台, 对齐 SOFABoot 研发运维生态能力
- 可集成 AgentUniverse、LangChain 等 Agent 开发框架
- 基于 MCP 组件提供 MCP Server/Client 能力
- 基于 Flask 组件提供 Web 托管能力
- 内置 Layotto 等基础组件, 可快速访问各种 SOFA 中间件
- 内置 Tracer 组件, 默认接入蚂蚁可观测体系



Agent: 研发平台



工具：MCP 场景与挑战



MCP 带来的改变

工具定义的标准化

生产关系的解耦

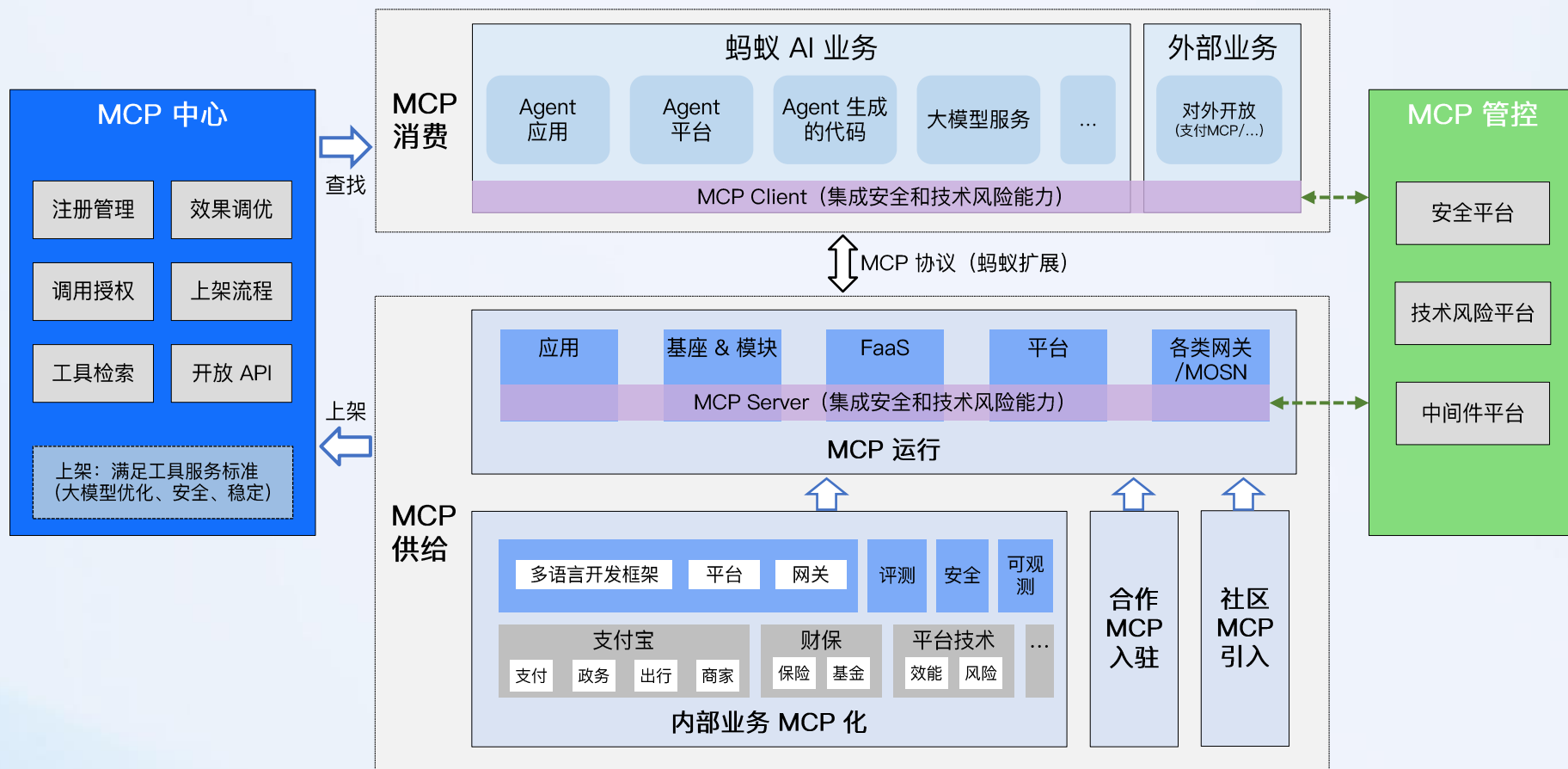
MCP 带来的挑战

新流量入口

基建需长期建设

最佳实践待摸索

工具：MCP 基础设施



数据：SOFARAG



数据：SOFAMemory



传统中间件适配 AI 场景

微服务

全链路 Trace 集成

多模态全链路流式

Agent/Tools 注册中心

Agent 配置中心

网关

大模型服务网关

Agent 互通网关

MCP 互通网关

消息

异步/批量推理

Agent 长程协作

函数平台

MCP Server 托管

沙箱托管

04

总结与展望

“进化”心得

深入业务

从业务中来的，到业务中去，需深入 AI 业务解决业务痛点，从解决问题中沉淀最佳实践，而不是拿着锤子找钉子。

拥抱开源

相信社区的速度和集体智慧，拥抱开放的标准和能力，积极推动标准化、工程化、组件化。

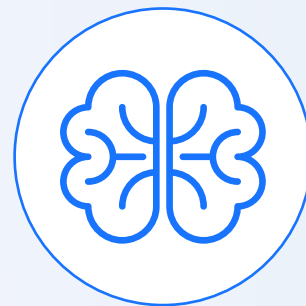
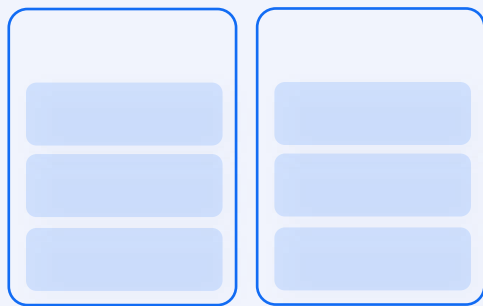
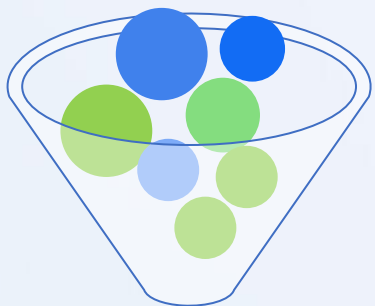
长期主义

AI 应用研发范式日新月异，不断演进，尚未收敛，中间件也需要持续迭代，持续演进。

守正创新

拒绝 AI FOMO，传统中间件依然是核心，不能为了拥抱 AI 放弃传统中间件的投入。

趋势展望



短期

各子领域依然会涌现新的AI中间件组件，部分组件会随模型能力增强被替代。

中期

AI中间件在不同子领域都会逐渐组件化、标准化，AI中间件体系逐渐完善。

长期

AI中间件会成为AI应用的骨架和神经中枢，连接模型、工具、数据。

📍 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



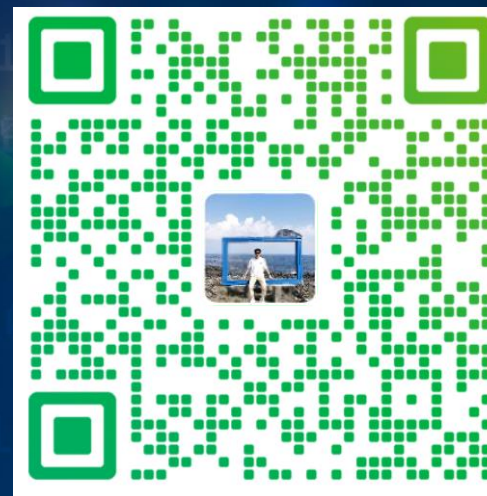
参会咨询



查看会议

The best way to predict the future is to create it.

预测未来最好的方式就是去创造它！



欢迎加微信好友交流加入蚂蚁中间件

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software

