亚马逊云科技

中国峰会

2026年6月23日-24日 上海 · 世博中心

亚马逊云科技
中国峰会

取之有度，用之有节-破解 Agentic 应用 Token 爆炸难题

吕琳

亚马逊云科技 数据专家架构师技术负责人

于超

聚云科技 售前技术总监

今天你养龙虾了吗?

OpenClaw 为什么会成为
2026年最火爆的 AI Agent 项目?

- 从“只会说”到“能动手”
- 极低的使用门槛
- 强大的可扩展能力
- 开源生态快速扩张

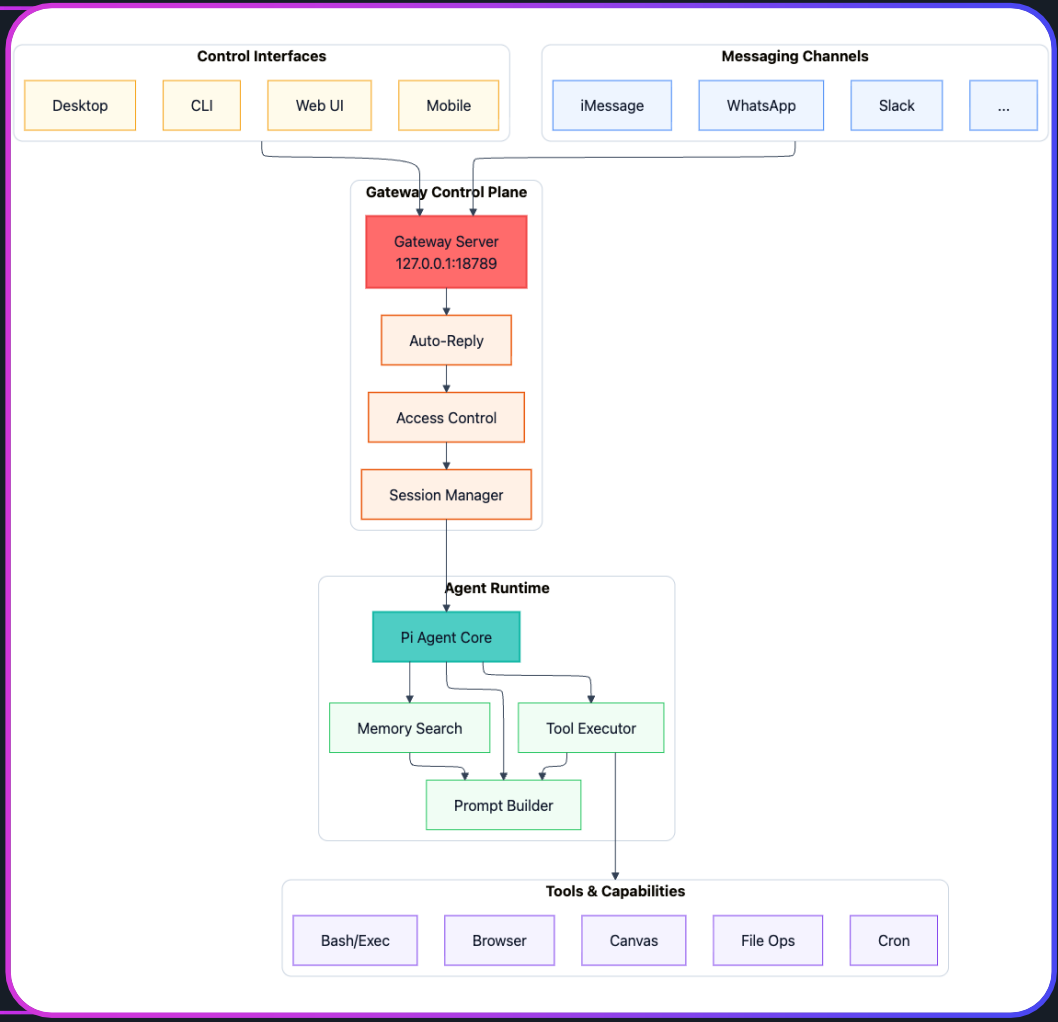
带来一个新的问题:

Token 爆炸

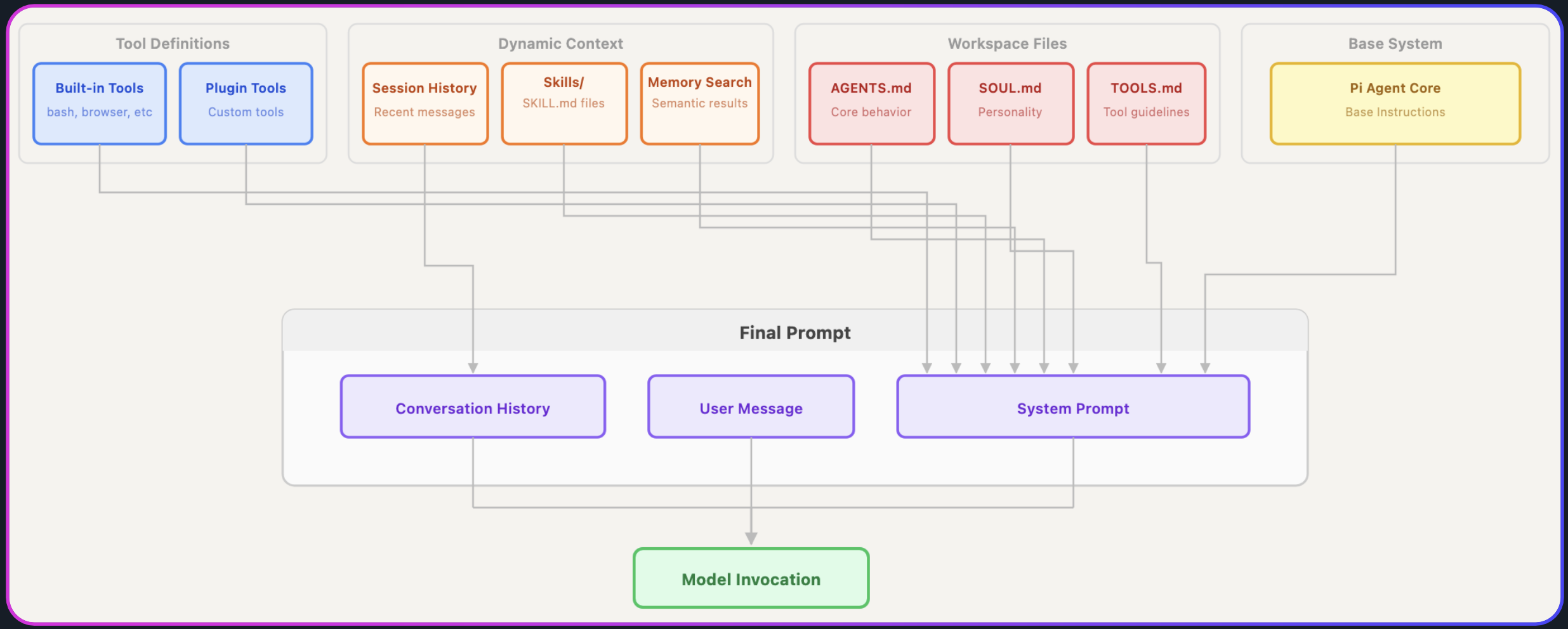
实验环境到生产的拦路虎

OpenClaw 核心组件

1. **Channel Adapters**：将不同消息平台的差异抽象化，提供统一的消息接口，可以对接不同消息平台。
2. **Control Interfaces**：提供多种方式与 Gateway 交互方式
3. **Gateway Control Plane**：消息路由中心，更是安全边界、状态管理器和协调者的统一体。
4. **Agent Runtime**：
   - 解析会话。
   - 组装上下文
   - 流式调用模型并执行工具调用
   - 将更新的状态持久化到磁盘



OpenClaw System Prompt



常见的 Token 爆炸原因

黑盒型爆炸
—
可观测性

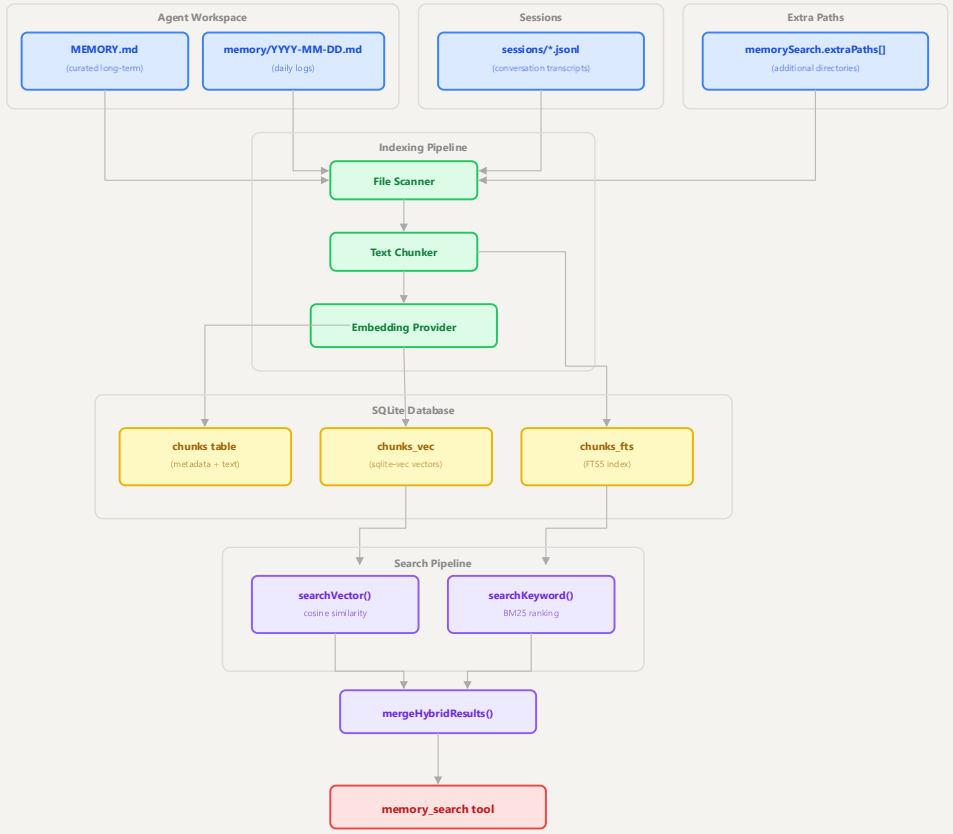
重复型爆炸
—
记忆管理

注入型爆炸
—
Skill 管理

OpenClaw 如何管理记忆

OpenClaw 的记忆管理系统以本地文件（Markdown）+ SQLite 索引为核心，并结合向量嵌入（Vector Embedding）和全文检索（FTS）来提升检索效率。

- **MemoryIndexManager**
- **混合搜索**
- **记忆同步与更新机制**
- **Dreaming系统**



AI Agent 的记忆管理最佳实践

分层记忆和记忆策略

上下文工程和动态加载

向量存储选型

命名空间和跨会话共享

多种向量存储适配不同场景

Amazon DynamoDB

适用场景:
高并发键值存储场景，支持毫秒级读写和无缝扩展

特点:
- 毫秒级响应延迟
- 无服务器自动扩展
- 全托管键值存储

Amazon OpenSearch

适用场景:
需混合搜索(语义+关键词)、大规模向量数据、已有Elasticsearch 技术栈

特点:
- 混合检索能力
- 丰富分析功能
- 极致成本优化

Amazon S3 Vector

适用场景:
大规模向量归档 (PB 级)、成本敏感型 AI 应用、冷数据存储

特点:
- 超低存储成本
- 多租户强隔离&无限扩展
- 混合架构集成（AOS)

Aurora PostgreSQL

适用场景:
熟悉 PG 技术栈、需要混合查询(向量+结构化数据)、多租户 SaaS 应用

特点:
- PostgreSQL 生态兼容
- 统一数据模型
- 企业级可靠性

Neptune Analytics

适用场景:
GraphRAG 应用、需要可解释性的 AI 推理

特点:
- 图原生推理
- 超高性能
- 一体化查询

Elasticache/ MemDB

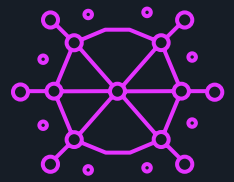
适用场景:
实时推荐、语义缓存、高并发向量收获（万级 QPS)

特点:
- 极致低延迟
- 实时索引更新
- 显著成本节约

多态融合的全景 Serverless 数据底座，实现按需使用与自动驾驶级的弹性免运维

S3 Vectors

空间隔离



每个索引提供独立的向量维度(dimension)和向量空间距离(distance)配置

每个索引可以单独配置 SSE (S3/KMS)的加密方式

性能隔离

目前已知的服务上限（尤其是向量的注入和查询）都是按每索引来设置的上限

索引和索引之间性能相互独立（通过内部程序已验证），可横向扩展

安全隔离

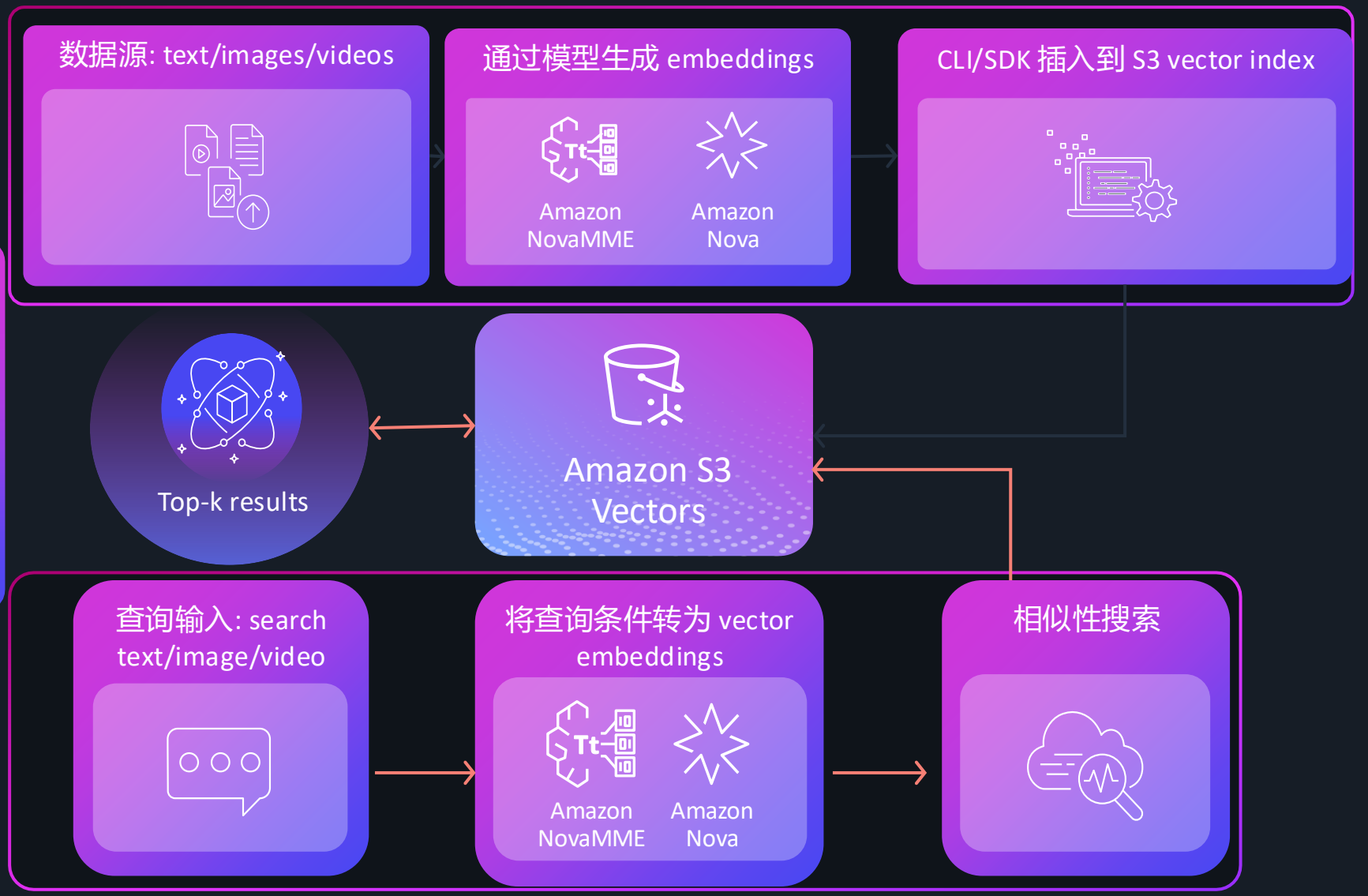
每个index都有独立的 ARN

每个 index可以单独配置 IAM policy 控制每个 API 的权限

S3 Vectors

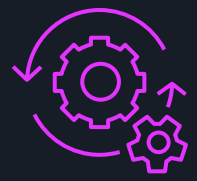
示例架构

相似性检索
NovaMME + S3 Vectors



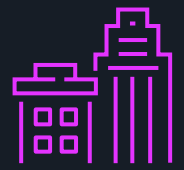
AgentCore Memory: 构建具有上下文感知的 Agent

简化记忆系统管理



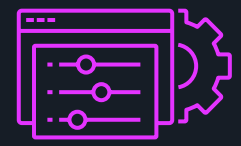
- 只需几行代码即可存储和检索记忆。
- 自动完成向量嵌入、存储、记忆整合和反思过程。

企业级安全



- 在基于命名空间 (Namespaces) 的加密存储中，按用户、项目或业务单元划分内存。
- 在安全的 VPC 环境中，保持数据隔离且易于检索。

深度定制化

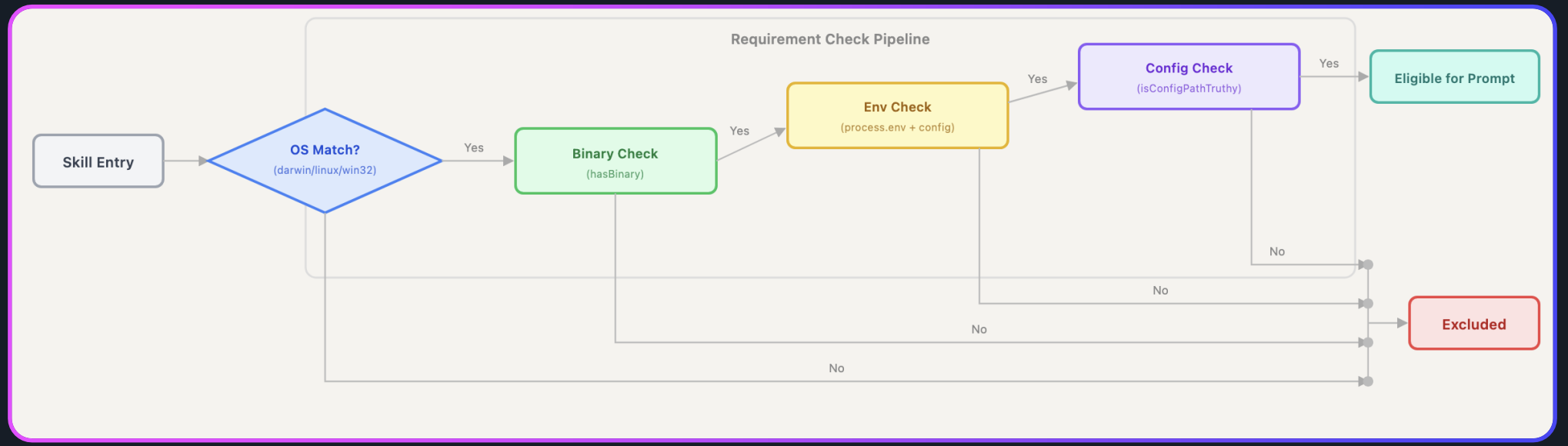


- 通过跨 Session 和 跨 Agent 的短期和长期记忆功能，实现行业领先的准确性。
- 支持 [1] 使用预定义的提取策略，或 [2] 使用任何模型创建自定义逻辑。

OpenClaw 如何管理 Skill

在 OpenClaw 中，Skills（技能）是扩展 Agent 能力的核心机制。从实现上看，OpenClaw 的技能系统主要包括：

- 声明式定义
- 通过多来源加载
- 渐进式披露
- 动态注入和快照



AI Agent 的 Skill 管理最佳实践

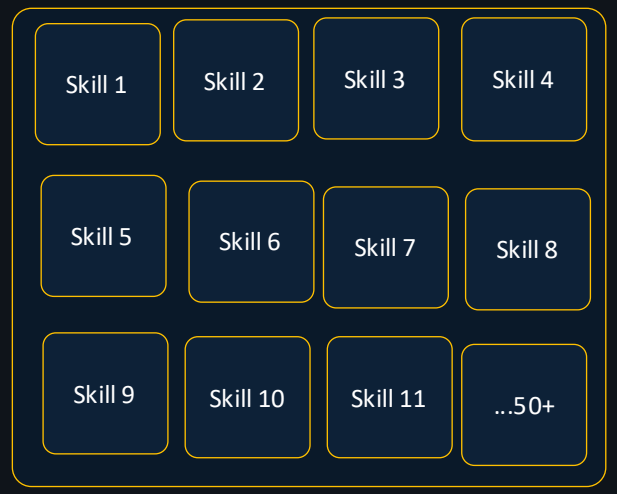
技能检索
和
按需加载

渐进式
披露

技能组织
和
多租户管理

技能按需召回：告别“Token 刺客”

痛点：“填鸭式“ Token 消耗巨大

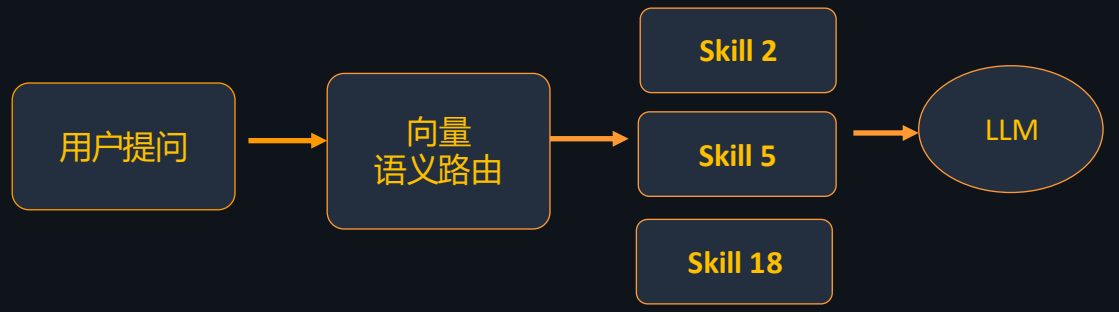


企业 Agent 挂载 50+ Skills，每次提问全量注入所有工具说明

→ API 账单爆炸
→ 更长的 prompt 拖慢大模型响应

解法：动态技能智能路由

50+ -> Top-N Skills



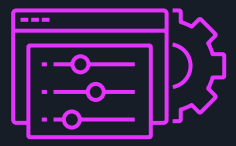
只挑出最相关的 Top-N Skills 喂给模型

成效：又快又省

~90%	~50ms	3-5
Token 节省	**缓存命中延迟**	**精准 Skill 匹配**
提示词消耗从 ~2000 锐减至 ~200	运行时透明拦截，用户完全无感	告别 50+ Skill 盲目全量注入

Amzon Agent Registry：在整个组织中发现、管理和重用 Agent、Tool、Skills

所有 Agent 的统一平台

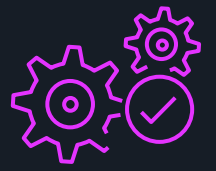


- **统一管理**：所有 Agent、Tool、MCP Server 和 Skills 的单一可信数据源——无论它们在何处构建或托管
- **统一注册**：通过指向 MCP或A2A Endpoint 进行注册，或通过控制台、SDK 或 API 手动注册
- **协议灵活**：原生支持 MCP 和 A2A，具有灵活的自定义架构

查找已存在的内容

- **混合搜索**：关键词过滤和语义搜索支持自然语言查询
- 多方式：可通过 AgentCore控制台、API 以及作为来自 Kiro、Claude Code 的 MCP Server
- OAuth 支持：让团队无需 IAM 凭证即可构建自定义发现 UI

管理发布内容



- **审批工作流**：草稿 → 待审批 → 已审批，之后 Agent 才可被发现
- **完整的生命周期跟踪**：包括版本控制、弃用和审批流程钩子
- IAM 访问控制：管理谁可以注册、发现和使用

OpenClaw 的可观测性

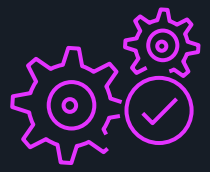
基础层面
/status 查询

单纯
结果展示

缺乏执行链路
细粒度
追踪能力

AgentCore Observability：全面掌握 Agent 性能状况

确保质量与可信度



- 全面掌握所有 AgentCore 服务的 Agent 执行情况与运营指标
- 加速调试和质量审计
- 快速检测问题并评估性能趋势

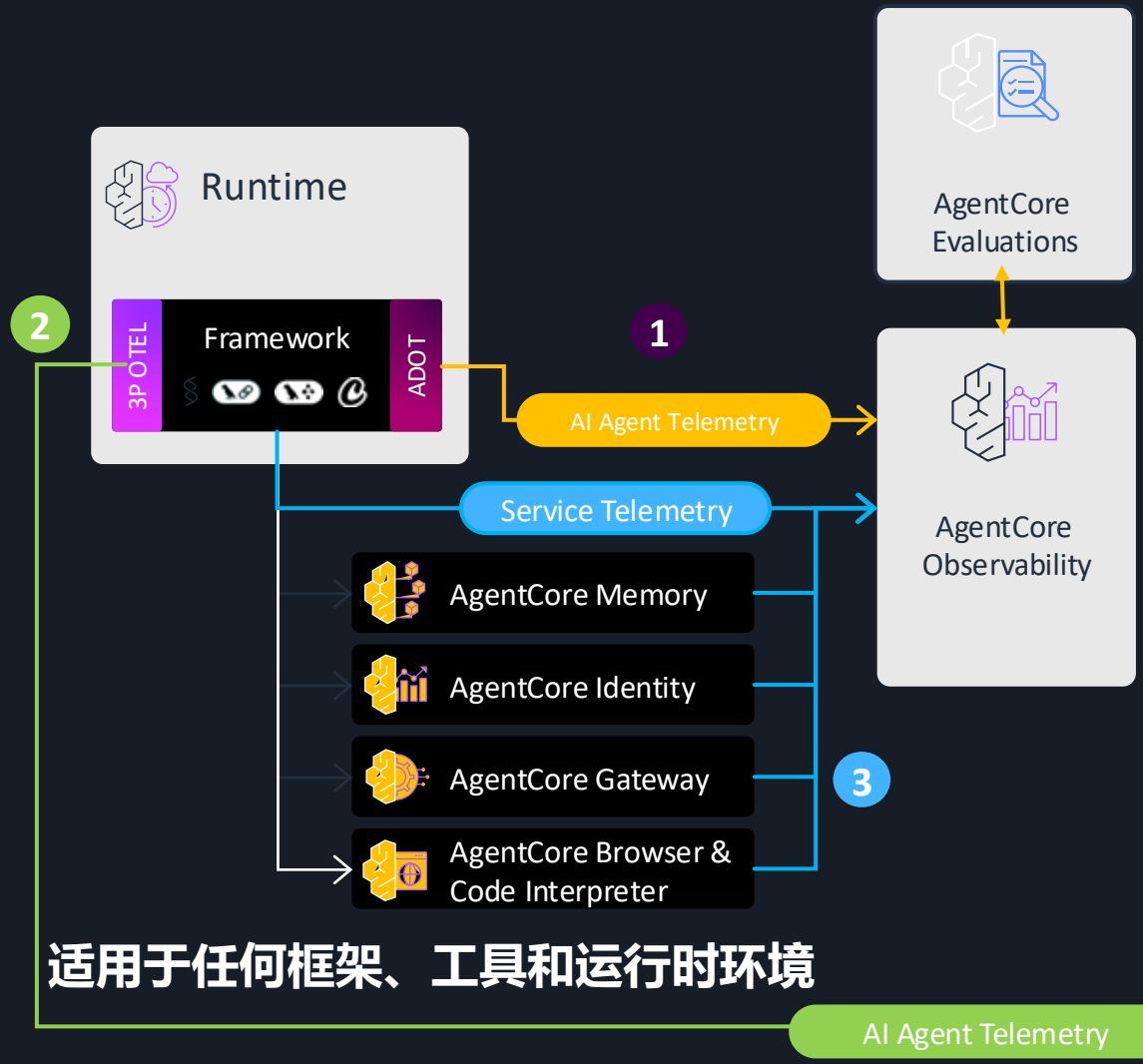
加速产品上市时间

- 通过实时控制面板集中显示 Agent 的运行状况和性能，以加快问题解决速度
- 只需极简的可观测性基础设施配置
- 通过在 Agent 追踪信息中加入自定义属性和元数据，使其与业务成果的关联更为紧密

支持与您自选的各类可观测性工具进行集成

- 兼容 OTEL 的遥测技术可与多种可观测性工具集成，包括 CloudWatch、 Dynatrace、 Datadog、 Arize Phoenix、 LangSmith 和 Langfuse
- 充分利用现有的可观测性技术栈

AgentCore Observability

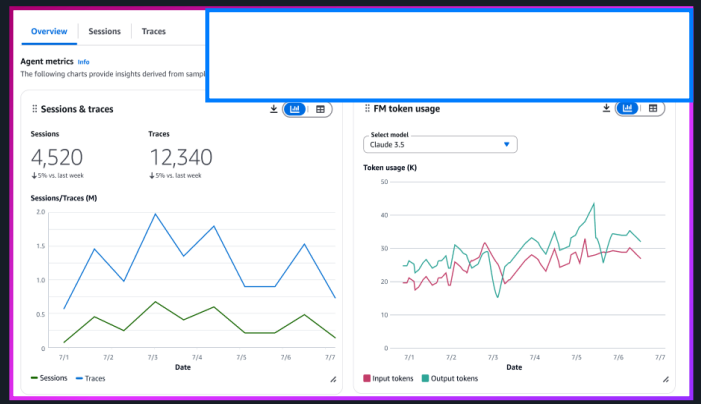


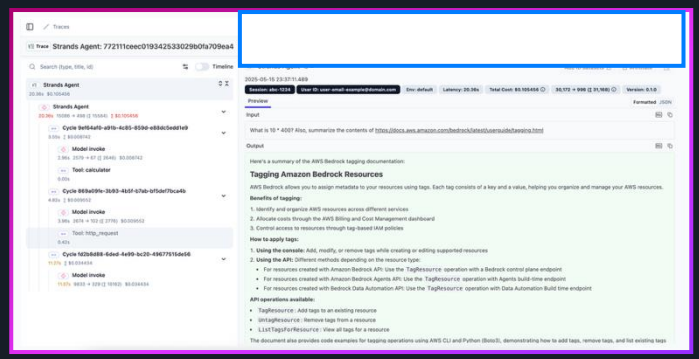
1 使用 AgentCore Observability 收集 AI Agent执行轨迹

2 也支持使用第三方可观测性工具收集 AI Agent 的执行 trace。

3 （可选）启用 AgentCore 可观测性服务级 trace

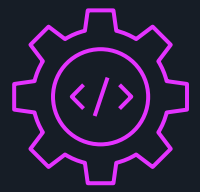
Automatic linking of AI Agent & Service generated OTEL





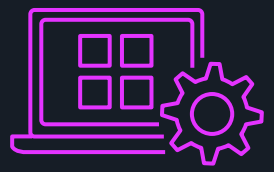
AgentCore Evaluations：基于真实世界性能提升 Agent 质量

实时质量智能



- 使用13个内置评估器对实时交互进行采样和评分
- 从真实世界用户行为中获得可操作的洞察，持续改进智能体性能。
- 提供按需 (On-Demand) 和在线评估 (Online) 两种评估模式

自定义业务评分



- 使用自定义评估器为特定用例构建定制的质量评估。
- 查看各种粒度级别（整个Trace、最终响应、单个步骤）。

无基础设施开销

- 从 Amazon CloudWatch 中可用的日志进行开箱即用的评估。
- 消除构建基础设施和管理运营复杂性所需的数月工作。

企业　人

业务产出

AI 成熟度等级: 1.试点 → 2.生产 → 3.规模化

	业务价值	衡量标准
Agents 专用 Agents（业务/垂直领域） 通用 Agents（如软件开发、客户交互中心、知识工作者）	作为"数字员工"为公司工作	• 能否及时有质量完成工作 • 成本是否合理
Agentic 平台 运行环境 / 记忆 / 工具（浏览器） / 工具（代码运行） / MCP网关 / 技能 / 多 Agent 协同 规则 / 评测 / 可观测性 / 安全 / 治理 / ...	Agent 所需的通用环境、工具、规则、评估、治理等必要组件	能否支持成百上千个 Agents 的开发、部署、管理、迭代
数据和知识 RAG / 向量数据库 / 知识库 / 数据管道	为 Agent 提供相关数据	• 检索准确性 • 数据新鲜度 • 覆盖范围
模型 Claude / GPT / Gemini / Grok / Nova Seed / Qwen / GLM / DeepSeek / MiMo / ...	以 API 调用方式为 Agents 提供"智能服务"	• 智能水平 • 速度 • 成本 • 上下文窗口
AI 基础设施 GPUs / AI 专用芯片	为模型推理提供算力	• 吞吐量 • 每小时成本（初始采购成本、能源效率） • 可靠性

安全

效果

性能

成本

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

从 OpenClaw 到 EasyClaw：理论如何变成真实运行的客服 Agent

理论基础

- 🧠 Memory — 分层记忆，避免重复 Token 消耗
- 🔧 Skill — 按需召回，90% Token 节省
- 📚 Knowledge Base — 向量化知识，精准检索
- 👁️ Observability — 四层穿透式成本归因

生产实践

- 👥 万级用户 — 真实服务规模
- ⏰ 7×24
- 💬 飞书群并发 — 多群实时运行
- 🚀 零停机 — 持续稳定运行中

客服，是 AI 员工最好的第一岗位

四个天然条件，让客服成为 Agent 落地风险最低、价值最快显现的场景

高频重复问题

80% 的问题来自 20% 的知识点，天然适合知识库覆盖

Knowledge Base

有标准答案

产品规格、价格、政策有权威来源，无需创造性发挥

RAG 精准检索

多系统查询

查订单、查库存、建工单，天然需要多个 Skill 协作

Skill 按需召回

风险边界清晰

不知道就转人工，天然支持 Human-in-the-loop

Observability 兜底

四个条件，四个技术支柱——客服是天然匹配的最佳验证场景

龙虾客服运行架构

六层架构：OpenClaw 理论在客服场景的完整落地

用户 · 飞书 / 企微 / Web

↓

Agent Runtime · EasyClaw

↓

判断层 · 意图识别 + 风险边界

↓

知识层 · Better KB + OpenSearch

↓

行动层 · Skill（CRM / 工单 / 库存）

↓

可观测层 · EasyClaw 监控看板

理论概念	龙虾客服落地
🧠 Memory 短期	当前会话上下文
🧠 Memory 长期	历史工单 + 客户偏好
🔧 Skill 按需召回	CRM / 工单 / 库存 3个Skill
📚 Knowledge Base	Better KB 4类知识分层
🔍 Vector DB	OpenSearch 向量索引
👁️ Observability	Token 用量逐对话可见

知识库不是文档仓库，是客服上岗手册

Better KB + OpenSearch 如何让龙虾客服做到不越界、不乱答

📄 FAQ 标准问答

高频问题的权威答案，命中即返回，不走大模型推理
节省省 Token

📋 Rules 风险规则

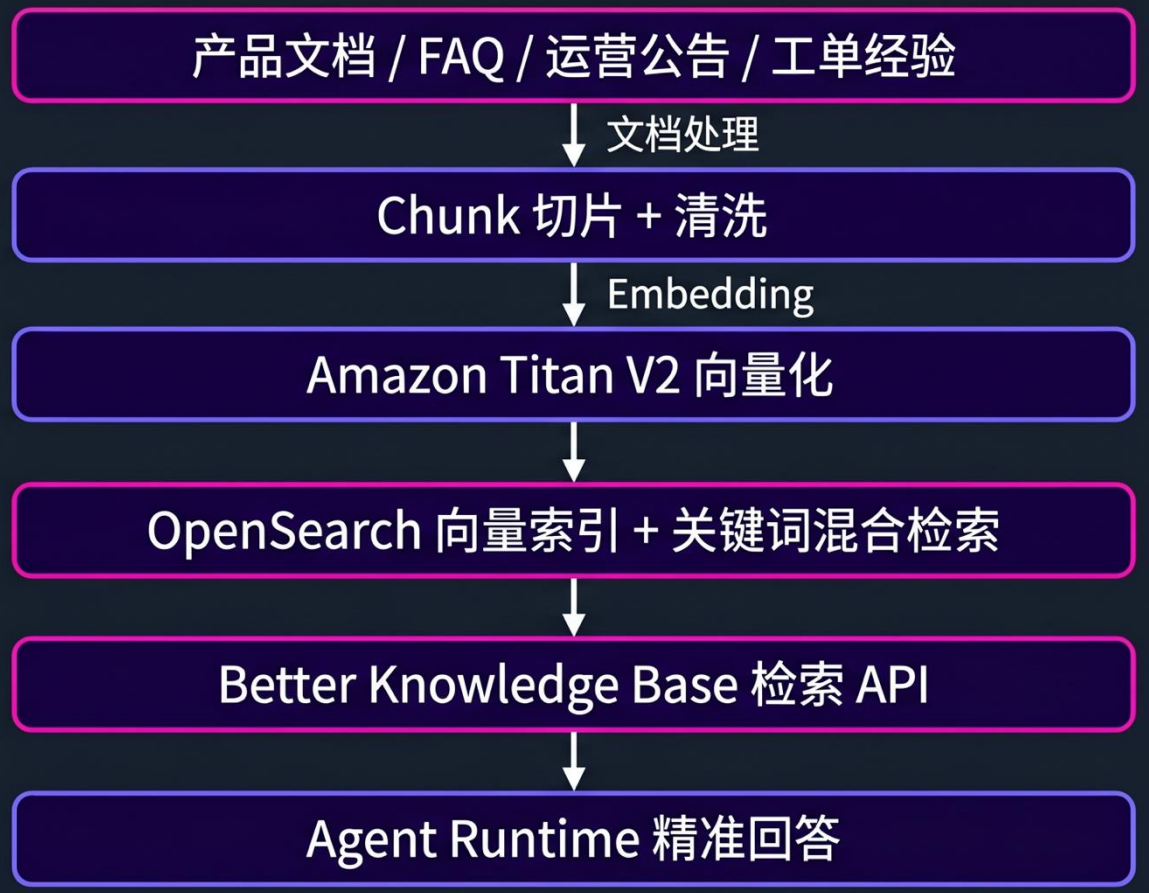
哪些话不能说，哪些操作要转人工，在检索前过滤

📣 Updates 最新政策

促销规则、库存调整，带时间戳，最新优先

📦 Cases 历史工单

真实解决过的案例，作为 RAG 的补充知识源



风险收口 + 人机协同：Agent 的边界与成长

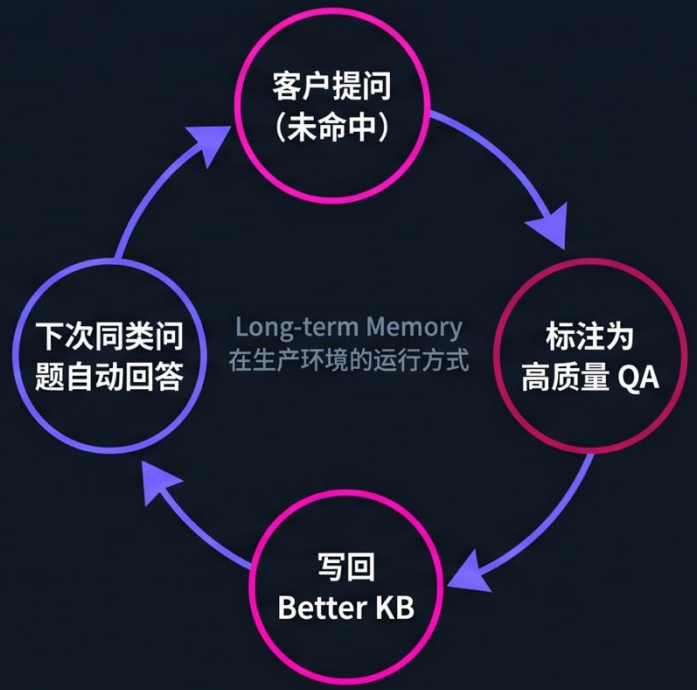
好的客服 Agent，知道什么时候该回答，什么时候该让人来

风险收口

- **敏感词触发** 价格争议、投诉升级 → 自动转人工，附带上下文摘要
- **置信度不足** 知识库检索相似度低于阈值 → 标注「我不确定」，转人工确认
- **需要执行动作** 退款、换货 → Skill 调用需要人工审核节点

「Agent 最大的价值不是回答所有问题，而是知道什么时候不回答」

Memory 持续进化



取之有度，用之有节：龙虾客服如何控制 Token

不把所有知识、历史和工具说明塞进 Prompt，而是按当前问题动态取用

全量注入，看似简单，实际很贵

常见做法

- 所有知识文档一次性塞入 Prompt
- 客户历史记录全量带入
- 所有工具说明全部注入
- 风险规则混在普通知识里
- 每轮对话都重复消耗 Token

Prompt 变长 · 响应变慢 · 成本升高 · 越界风险增加

按需取用，才是真正的「用之有节」

龙虾做法

- 当前问题只召回相关 FAQ / Rules / Responses
- 短期 Memory 只保留当前会话关键上下文
- 长期 Memory 只取必要的客户偏好和历史工单
- Skill 只注入本轮需要的 CRM / 工单 / 库存能力
- Token 用量逐对话可见，持续优化召回策略

更低 Token · 更快响应 · 更少误答 · 更容易治理

Token 控制不是简单少给模型内容，而是让 Agent 在正确时间取正确的信息

典型案例

EasyClaw.work 龙虾管家，我们自己的龙虾客户实践

龙虾管家

飞书群智能客服专家，服务万级以上用户

万级以上
服务用户规模

7×24
全天候自动服务

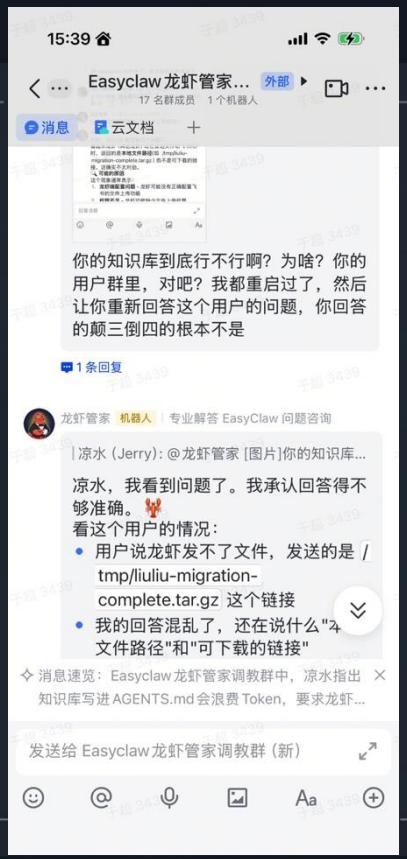
多群
同时接入飞书群

- 产品咨询FAQ精准回答，口径统一不走样
- 超范围问题自动转人工，不乱答
- 多群同时服务，一虾多用
- 主人私聊更新FAQ，群里立即生效

响应速度↑90%　人工成本↓60%　客户满意度↑40%

1 深夜解答

2 众人调教



亚马逊云科技
中国峰会

Thank you