

FAQ 问答文档

基于 2026-11-06 直播课堂学员提问按 12 个主题归类；答复结合 Amazon Bedrock AgentCore 官方能力补充，并引用课程资料中的 AgentCore 架构图。

📞 方案咨询 / POC 免费试用专线：1010 0966

📖 **官方中文文档：**https://docs.aws.amazon.com/zh_cn/bedrock-agentcore/latest/devguide/what-is-bedrock-agentcore.html

1. 产品定位与获取

Q：AgentCore 到底是什么、由哪些能力组成、和其它 Agent 产品/框架有何差异、如何获取与上手？是否开源、有无源码？

A：Amazon Bedrock AgentCore 是一个企业级“智能体平台”，用于安全、规模化地构建、部署与运营 AI Agent，核心主张是“任意框架 + 任意模型 + 免基础设施运维”。它由一组可独立或组合使用的模块化服务构成：Harness（托管 Agent Loop）、Runtime（Serverless 运行时）、Memory（记忆）、Gateway（工具/API 网关）、Identity（身份认证）、Code Interpreter（代码沙箱）、Browser（云端浏览器）、Observability（可观测/追踪）、Evaluations（评测）、Optimization（自动优化/A-B）、Policy（基于 Cedar 的权限策略）、Registry（Agent/工具治理目录）、Payments（智能体支付）。其中，中国区域目前提供 Runtime、Code Interpreter、Browser、Identity、Gateway、Observability；其余能力目前在海外区域可用，具体以中国区域官网为准。

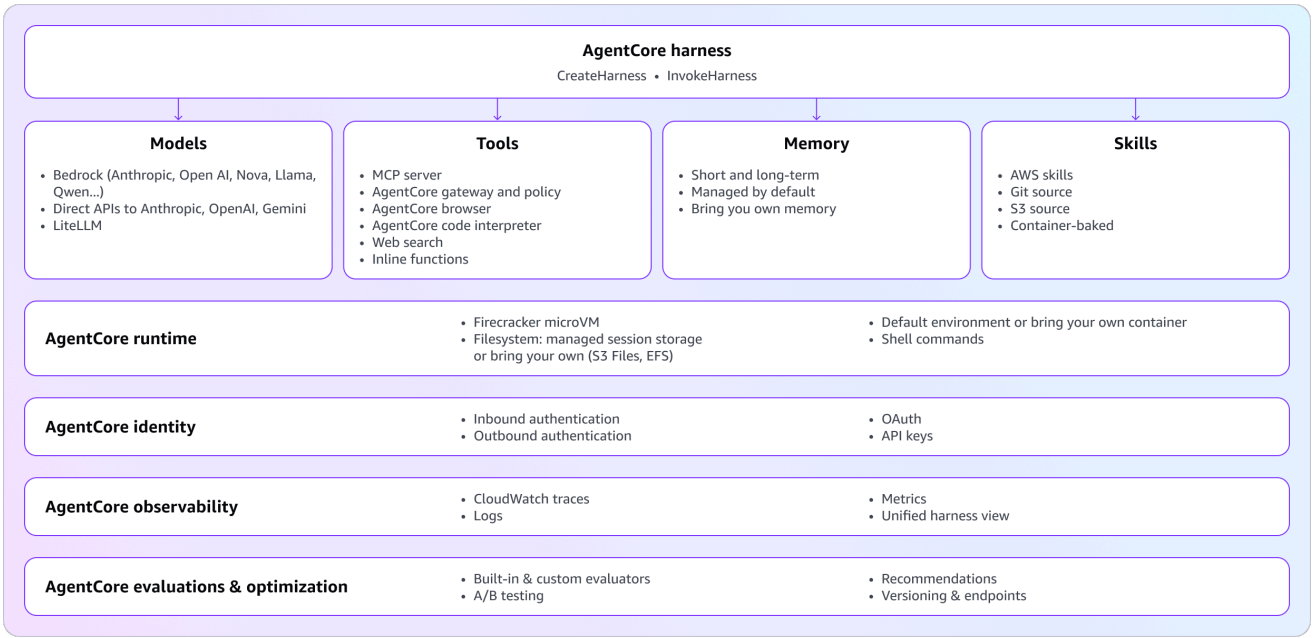


图 1 · AgentCore 能力全景：Harness（Models / Tools / Memory / Skills）+ Runtime / Identity / Observability / Evaluations & Optimization（来源：Agent 速造局课程资料）

AgentCore 与开源 Agent 框架是互补而非替代关系：它不绑定模型、不锁定框架、免运维、全托管——可直接承载 CrewAI、LangGraph、LlamaIndex、Strands、Google ADK、OpenAI Agents SDK 等任意框架。它是托管云服务，服务本身不开源；配套的 AgentCore SDK 已在 GitHub 公开，可查看源码并用于本地开发与部署。获取方式：拥有亚马逊云科技云账号后登录控制台即可使用，无需下载客户端；可用 Harness 低代码/无代码方式点选模型与工具快速构建 Agent。

2. 部署·Runtime·Harness

Q：Runtime 与 Harness 有什么区别？“4 行代码上云”本质是什么、运行在什么环境、从本地到生产覆盖 Agent 生命周期的哪些环节？

A：两者是分层关系：Runtime 是“安全的 Serverless 运行时”，底层为隔离的 MicroVM，提供快速冷启动、会话级隔离、内置身份、以及对多模态/多 Agent 的支持；你把自己的 Agent 镜像（任意框架、任意模型）部署上去即可运行。

Harness 则是在运行时之上的“托管 Agent Loop”——一次 API 调用即可内联声明 model + system prompt + tools，由平台自动处理编排、工具执行、记忆管理与响应生成，并一键集成 Memory / Gateway / Browser / Code Interpreter / Observability。

因此拆开的意义是：需要完全自定义 Agent Loop 的团队用 Runtime 跑自有镜像；希望“配置即 Agent”快速搭建的团队用 Harness。“几行代码上云”指的是用 SDK/CDK 把本地 Agent 打包成镜像并部署为 Runtime，平台接管编排、扩缩容、会话隔离、网络、记忆等，覆盖从本地开发到云端运行、扩展、观测的完整生命周期；仍需处理的通常是业务侧的 Agent Loop 逻辑、截断/超时策略与依赖打包。

3. 私有化·合规·数据驻留

Q：是否支持私有化/本地化部署？数据存放在哪里、是否出境、如何满足合规与数据保护？中国区与 Global 有何差别、是否有中文资料？

A：AgentCore 是云端托管服务，不提供“装到企业自有机房”式的本地化私有部署（自建整套托管能力复杂度很高）。但它运行在“你自己的亚马逊云科技账号”内，支持数据加密与网络隔离，并沿用亚马逊云科技的安全与合规能力，帮助企业落实数据保护要求；个人/企业无需自购硬件。

部署区域上，客户可按业务需要选择中国区域或海外区域。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。中国区域目前提供 Runtime、Code Interpreter、Browser、Identity、Gateway、Observability，其余能力目前在海外区域可用。官方提供的技术说明文档可查看更多信息：📖 官方中文文档：https://docs.aws.amazon.com/zh_cn/bedrock-agentcore/latest/devguide/what-is-bedrock-agentcore.html

4. 安全·权限·身份

Q：多租户如何隔离、企业知识库与隐私数据如何保护？身份、权限、越权、审批、审计与提示注入如何管控？

A：隔离是平台级能力：Runtime 中每个会话（Session ID）获得“专属 MicroVM”，拥有独立的计算、内存与文件系统；会话结束后整个 MicroVM 被销毁、内存被清理，从机制上消除跨会话/跨租户数据残留，再叠加按用户 ID 隔离上下文与记忆。

身份方面有 AgentCore Identity，兼容 Amazon Cognito、Okta、Microsoft Entra ID、Auth0 等现有 IdP，可把用户身份与授权范围传递到调用链（无需迁移用户或重建认证）。权限/越权由 AgentCore

Policy（目前在海外区域可用）提供确定性管控——用自然语言或 Cedar 策略语言定义“哪个 Agent 能用哪些工具、在什么条件下执行什么动作”，并在 Gateway 处于每次工具调用前拦截校验；Gateway 还能做内容过滤防止敏感数据外泄。

审计与排障由 Observability 提供（OpenTelemetry 标准、逐步追踪执行路径）。对于“审批绑定具体工具/参数/对象、变更后自动重新审批”“提示注入导致越权”等深度诉求，可结合 Policy + Identity + Observability 组合设计，建议与技术架构师沟通。其中针对提示注入：Policy 在模型之外、于 Gateway 处对每次工具调用做确定性校验——即使提示注入影响了模型判断，也无法越权调用未授权的工具。

5. 记忆系统 Memory

Q：Memory 的短期/长期记忆如何划分、能存多久？能否跨会话/跨 Agent/跨模型共享、如何自动提取与纠错、是否消耗 token?

A：AgentCore Memory 是为 Agent 工作负载专门设计的记忆服务，明确区分两类：短期记忆用于多轮对话的上下文，长期记忆跨会话持久保存，并支持在多个 Agent 之间共享记忆库、从历史经验中学习。关键机制上要分清两层：Runtime 会话内的“运行态”是临时的，随会话（MicroVM）终止即被清除；真正需要跨会话留存的数据应写入 Memory。

记忆以 namespace 组织，可按用户/Agent 维度隔离，也可自行设计 namespace 实现跨会话/跨 Agent/跨模型共享。长期记忆由平台从对话中“策略化提取”（而非原样堆存），检索可结合语义与元数据（如创建/更新时间戳）；因此面对“新老偏好冲突”可通过时间戳加权与更新策略处理，并能记录来源以支持纠错。

短期原始事件 7–365 天可配；长期记忆记录独立 TTL，最短 31 天、默认不过期。提取与总结属平台

后台能力，使用时 Agent 按需检索；实时记忆流会产生相应的调用/计费开销。

6. 工具·Gateway·集成

Q：如何接入业务系统与内网？Gateway/MCP 相比直连 API 的价值是什么？能否即插即用、用 Skill 定义任务、接自有模型、支持跨云、语义发现工具？

A：AgentCore Gateway 可用几行代码把你的 REST API、Lambda 函数、现有服务转换为 MCP 兼容工具，也能直接对接既有 MCP Server，并内置 Salesforce、Zoom、JIRA、Slack 等常见集成；连接内网系统支持 VPC 部署。

相比直连 HTTP API，Gateway/MCP 的增量价值不只是协议统一，更包括：凭证托管、权限治理（配合 Policy 在调用前拦截）、以及语义化工具发现（混合语义+关键词检索，帮 Agent 从大量工具中找到合适的）——注意“语义最接近”不等于“适合执行”，因此要叠加 Policy 做权限/参数/风险约束，避免“查询订单”误选到“取消订单”。你可以用 Skill 定义企业 SOP 定向任务。

跨云/跨网最基本的保障是网络与数据链路可达，复杂集成建议联系客服（电话 1010 0966）。

有状态 MCP 托管在 Runtime 时，其 session 生命周期与 Runtime 会话的超时对齐需在设计时明确划分责任。

7. 多 Agent 协作·A2A

Q：多 Agent 协作与 A2A 能做哪些场景？主/子 Agent 如何调度、异构框架如何互通、何时该上多 Agent、跨公司协作如何落地、长任务一致性怎么保障？

A：AgentCore Runtime 原生支持多 Agent 与 A2A 协议，主 Agent 可编排基于 Strands、LangGraph、CrewAI、Google ADK 等异构框架构建的子 Agent，传输用 JSON-RPC/A2A，异构框架

间的状态/消息差异通过统一的上下文传递与 schema/适配层来对齐。

架构选型应从单 Agent → 固定工作流 → 多 Agent/Federated A2A 逐级评估，复杂度、成功率、延迟、成本与排障难度都要权衡——不是 Agent 越多越好。

跨公司协作可实现，但需提前约定账号与权限共享规则。对长任务与分布式一致性（如 MCP 同步调用成功、随后 A2A 委托超时）：应借助 Runtime 的异步/长时会话（通过 /ping 的 HealthyBusy 维持）、幂等设计与 Workflow 中的审核/补偿步骤来避免“恢复后重复执行”，并用 Policy/熔断限制单个工具或子 Agent 异常的横向扩散。Registry 可作为跨组织发现与治理 Agent/工具的目录。

KIRO

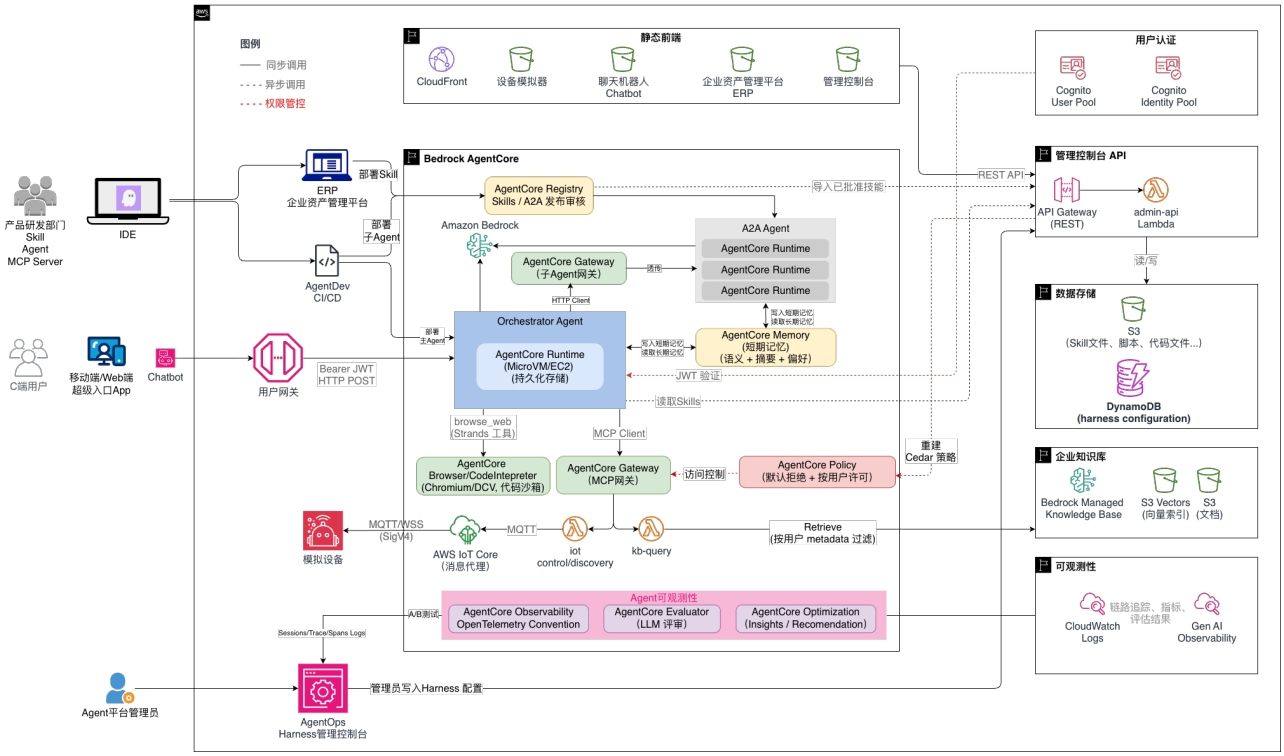


图 2 · 基于 AgentCore 的企业级多智能体方案架构：Orchestrator + 子 Agent、Gateway、Memory、Policy、Registry、Observability 全景（来源：Agent 速造局课程资料）

8. 性能·可靠性·长任务·幻觉

Q：延迟、并发、长任务续跑与模型幻觉如何保障？出错如何排查、能否自动监控/审核、能否自我迭代优化?

A：延迟优化多管齐下：选更快的模型（如轻量决策模型）、优化 Agent Loop（简单/复杂意图分路）、利用 Runtime 的快速冷启动与会话复用（同一 Session 多次 Invoke 复用上下文）。

长任务已上线。Runtime 两种计算类型及计费差异：MicroVMs（Serverless 秒级按量、I/O 等待时 CPU 缩容至零）最长 8 小时；Instances（账号内 EC2 + 管理费，目前在海外区域可用，中国区域仅支持 MicroVMs）最长 14 天，异步长任务通过 /ping 返回 HealthyBusy 维持存活；需配合幂等与断点续跑设计，避免“任务恢复了、业务却重复执行”（如已创建订单的去重判断）。

可靠性/幻觉方面：可在 Workflow 中加审核与一致性校验步骤兜底，用 Memory + RAG 补足业务知识；更系统的做法是用 AgentCore Evaluations 对会话/trace/span 做自动化评测（成功率、边界场景、输出稳定性），用 Observability（OpenTelemetry）追踪每一步执行路径定位 badcase，再用 Optimization 基于 trace 自动给出 system prompt/工具描述的改进建议并以 A-B 流量切分验证——这正是“Agent 自我迭代优化”的产品化路径（强化学习/自进化相关内容下一期课程 11/13 周五 14:00–15:00 介绍）。完全可以构建自动监控群消息/系统并自动接单的 Agent。

9. 模型支持

Q：支持哪些大模型、能否自定义或接自有模型？国内可用哪些模型？

A：AgentCore 不绑定模型。在海外区域，Bedrock 提供上百种模型，可通过 model ID 便捷切换；常见模型包括但不限于 Google Gemini、Anthropic Claude、Amazon Nova、Meta Llama、Mistral，也可接入企业自有/自定义模型端点。

中国区与 Global 可用的具体模型清单存在差异（受区域与合规影响），建议按目标区域确认，登录控制台，可直接查看亚马逊云科技云服务中提供的模型（全球不同区域）。

10. 商业化·收费·试用

Q：如何收费（包年包月还是按量）？有无试用额度？相比自研能省多少人力？

A：按用量计费，非固定包年包月套餐：Runtime、Code Interpreter、Browser 按实际使用的计算资源秒级计费；Gateway、Memory 等按调用量或存储量计费，详见官方定价页。企业级应用可申请 POC 试用额度——可拨打 1010 0966 咨询。

相比自研，AgentCore 以免运维、全托管把会话隔离、身份、记忆、工具网关、观测、评测等“脏活”产品化，可显著降低开发与运维投入（具体节省量依场景而定）。

11. 行业场景·应用案例

Q：有哪些行业落地场景与模板（电商、传统生产贸易、客服、知识库等）？有无认证考试、课程是否有操作/试用环节与课后资料？

A：AgentCore 适用于客户支持、流程自动化、数据分析、编码助手等自治 Agent 场景，并针对电商（选品、制图、广告视频、自动化运营）、传统生产贸易、独立站客服、企业知识库等企业级应用提供参考示例、架构定制等方案——欢迎拨打 1010 0966 由专家答疑设计。

企业知识库可结合 Memory + RAG + Gateway（接内部数据/系统）+ Identity/Policy（按用户控制可见范围）搭建。

课程方面：直播后会发布课堂笔记总结重点；目前没有专门的 Agentic 认证，相关内容已纳入 AWS Certified Generative AI Developer – Professional 专业级认证。

12. 其他·个人 vs 企业

Q：个人 Agent 与企业生产级 Agent 有何不同？自建是否必须基于这套底座、如何共享给团队、存量 Agent 迁移成本如何、能否检测重复建设？

A：企业级 Agent 的核心差异在于强管控与可治理：需要管理员在模型与工具层做安全防护，自主定义运行时环境与对外暴露的工具清单——推荐做法是把员工使用的个人助手、编码助手等 Agent 统一部署到 AgentCore Runtime，由企业集中管控模型、工具与权限。AgentCore 提供免运维、托管的 Agent 基础设施（会话隔离、身份、记忆、工具网关、观测、评测等），让你无需从零自建底座即可快速构建，面向 Agent 开发者与低代码构建用户。

多人使用的最佳实践是每用户一个独立实例、以 Session ID 区分。存量迁移上，因 Runtime 兼容任意框架与模型、并支持 MCP/A2A 协议，存量 prompt/自研脚本通常以“打包成镜像 + 接入 Gateway 工具”的方式平滑过渡，工作量依现状而定。

对“Agent 重复建设”，AgentCore Registry 提供集中编目、语义+关键词检索与发布/审批治理，有助于发现并复用已有 Agent/工具、减少重复造轮子与统一审计。

注：前述 Amazon Bedrock 及 AgentCore Harness、Memory、Policy、Evaluations、Optimization、Registry、Payments 等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。