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

125

构建企业级AI Agent 安全与合规方案实践

李阳

高级 AI 安全架构师

亚马逊云科技

戎巍

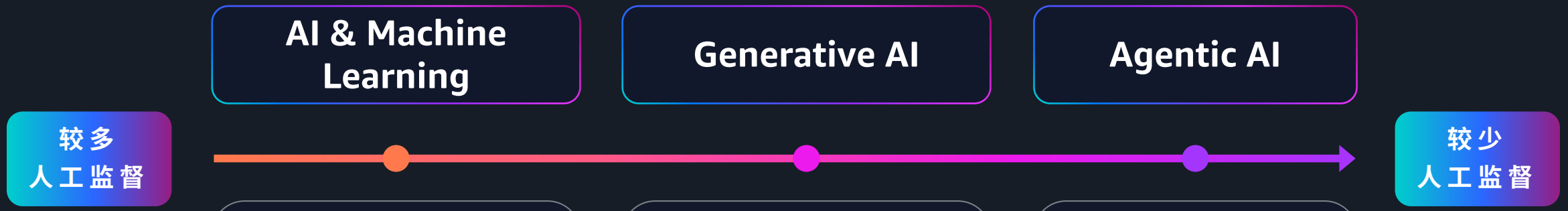
AI 产研负责人

龙腾出行

目录

企业级 Agent 威胁模型

人工智能的发展与演变



*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。

生成式 AI 应用的风险和挑战

行业框架

OWASP Top 10 for LLM/ NIST AI RMF

A list of the most critical vulnerabilities found in applications utilizing LLMs

法律法规

EU AI Act
ISO 42001
Responsible AI

Promoting the safe and responsible development of AI as a force for good

企业数据

Data Privacy and Security

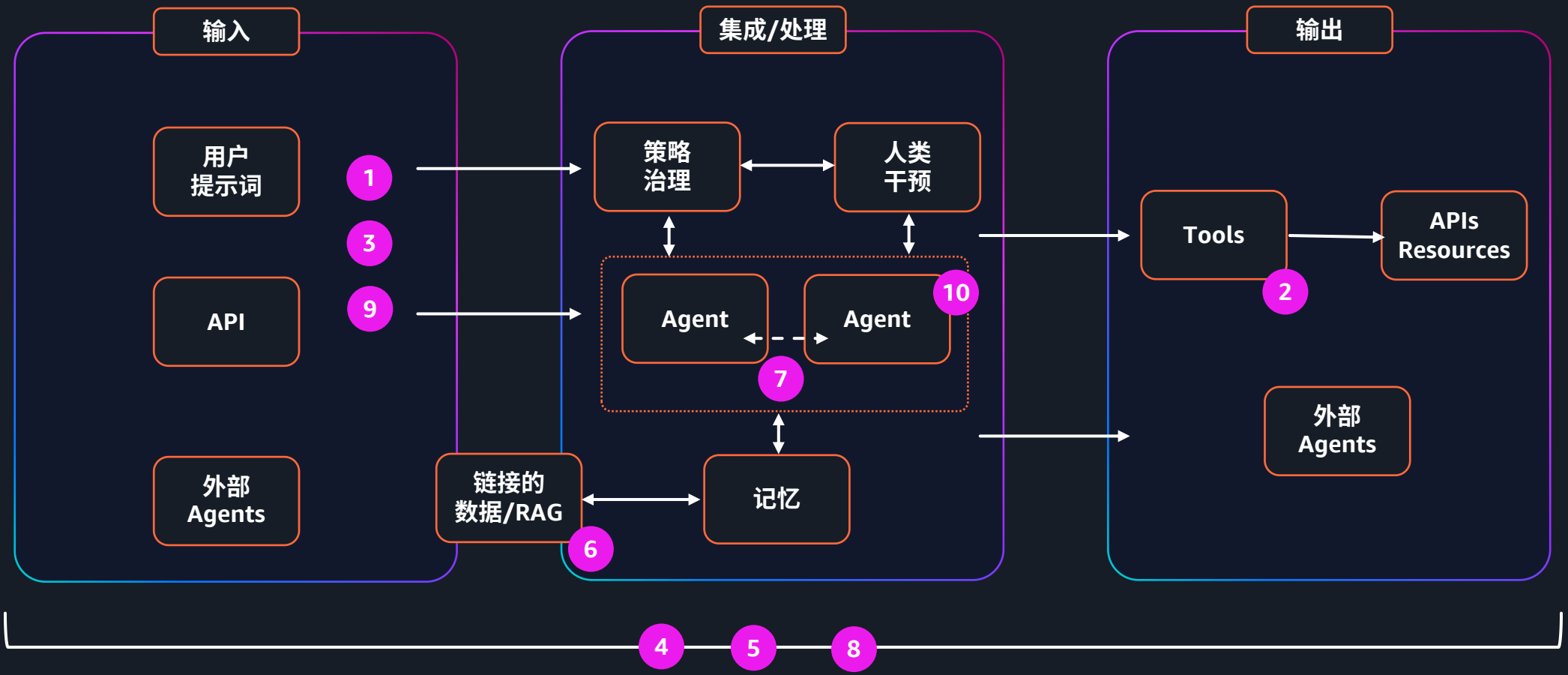
Protect enterprise sensitive data and personal information, data exposure, privacy and residency

*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。

OWASP Top 10 for Agentic Applications 2026

1. Agent 目标劫持
   Agent Goal Hijack
2. 工具滥用
   Tool misuse & exploitation
3. 身份和权限滥用
   Identity & privileage abuse
4. 供应链漏洞
   Agentic supply chain vulnerabilities
5. 非预期的代码执行
   Unexpected code execution
6. 记忆和上下文投毒
   Memory & context poisoning
7. 不安全的Agent间通信
   Insecure inter-agent communication
8. 连锁失败
   Cascading failures
9. 利用人机信任
   Human-agent trust exploitation
10. 异常 Agents
    Rogue agents

Agentic Top 10 安全威胁点



ASI01: Agent Goal Hijack
ASI02: Tool Misuse & Exploitation
ASI03: Identity & Privilege Abuse
ASI04: Agentic Supply Chain Vulnerabilities

ASI05: Unexpected Code Execution (RCE)
ASI06: Memory & Context Poisoning
ASI07: Insecure Inter-Agent Communication

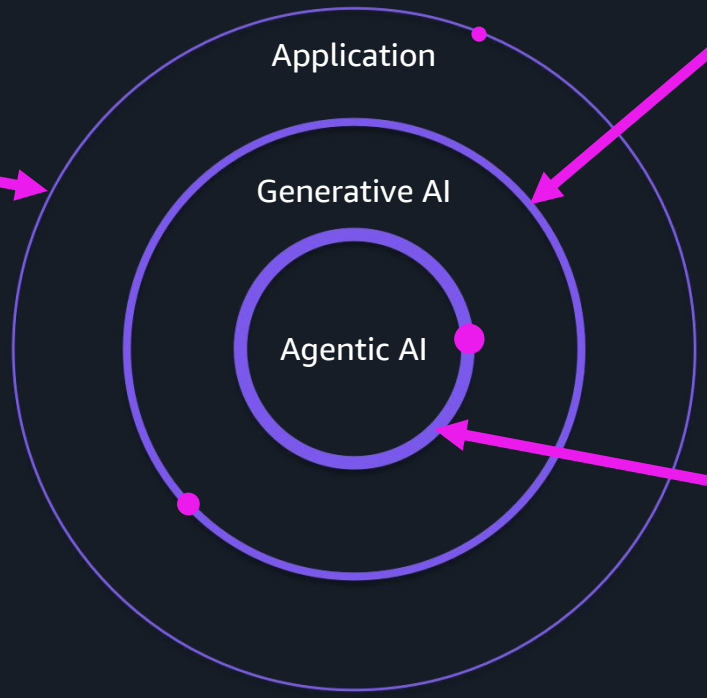
ASI08: Cascading Failures
ASI09: Human-Agent Trust Exploitation
ASI10: Rogue Agents

*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。

AI Agent 安全的纵深防御

通用应用安全

- 网络边界安全
- 身份与访问控制
- 威胁检测与事件响应
- 基础设施保护
- 应用保护
- 数据保护



生成式 AI 安全

- 基础模型安全
- 模型托管服务安全
- 防注入攻击
- 不安全内容过滤
- 基础安全升级

Agentic AI 安全

- 统一身份和鉴权
- 端到端授权
- 防止工具操纵攻击
- 防止间接数据投毒

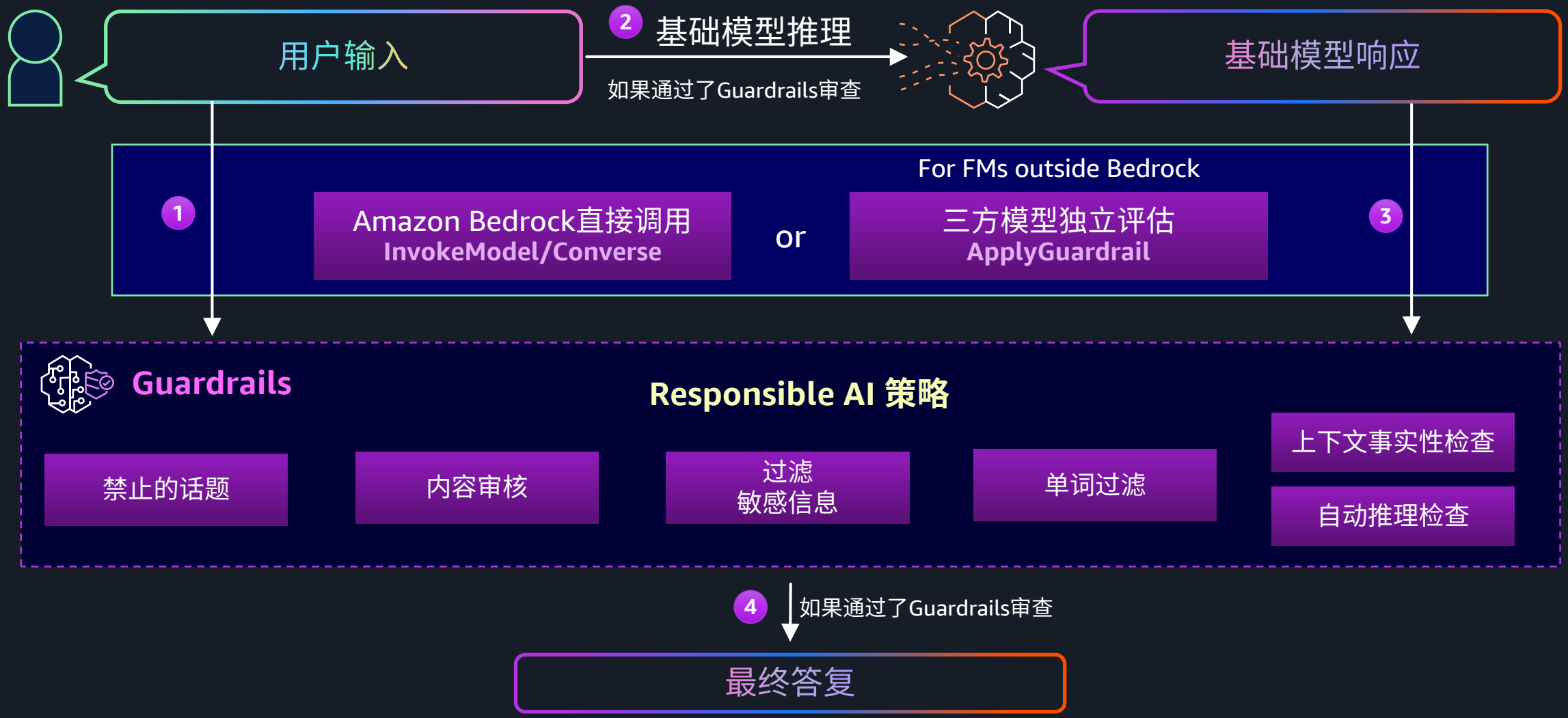
*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

生成式上下文过滤与合规

OWASP Top 10 for Agentic Applications 2026

1. Agent 目标劫持
   Agent Goal Hijack
2. 工具滥用
   Tool misuse & exploitation
3. 身份和权限滥用
   Identity & privileage abuse
4. 供应链漏洞
   Agentic supply chain vulnerabilities
5. 非预期的代码执行
   Unexpected code execution
6. 记忆和上下文投毒
   Memory & context poisoning
7. 不安全的Agent间通信
   Insecure inter-agent communication
8. 连锁失败
   Cascading failures
9. 利用人机信任
   Human-agent trust exploitation
10. 异常 Agents
    Rogue agents

Amazon Bedrock Guardrails



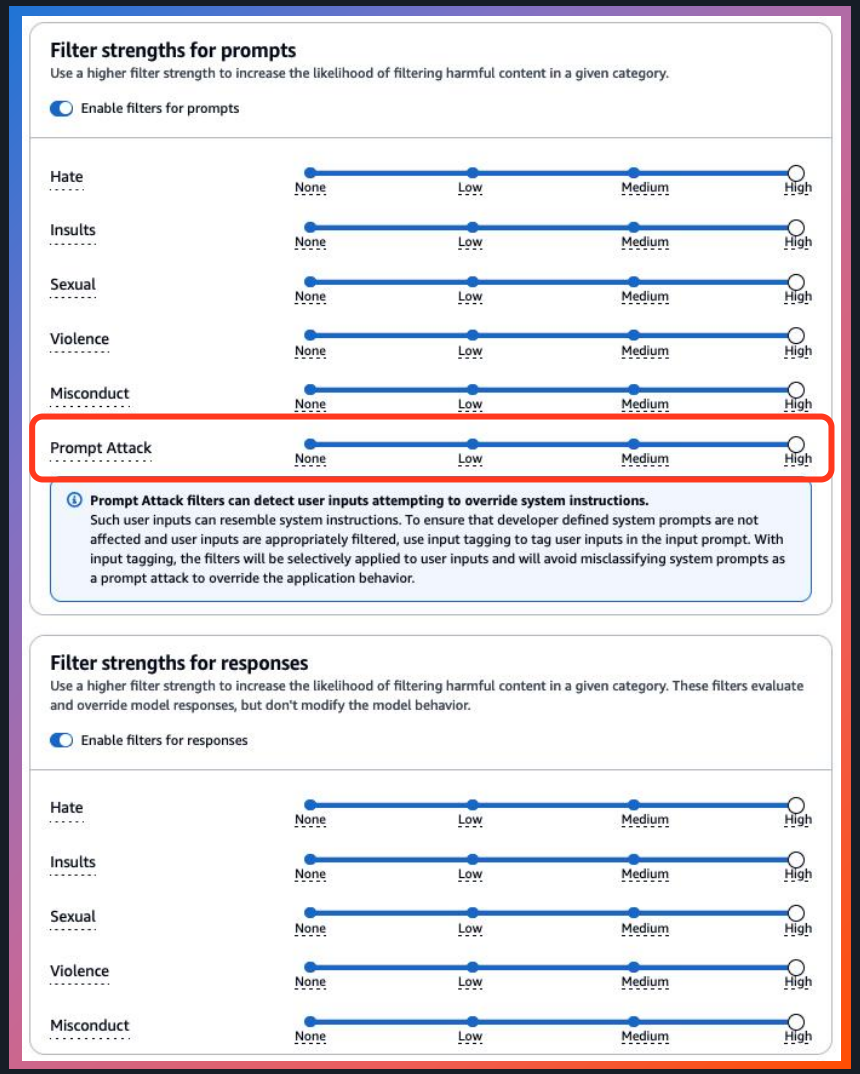
*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

内容过滤器

配置阈值以不同程度过滤内容

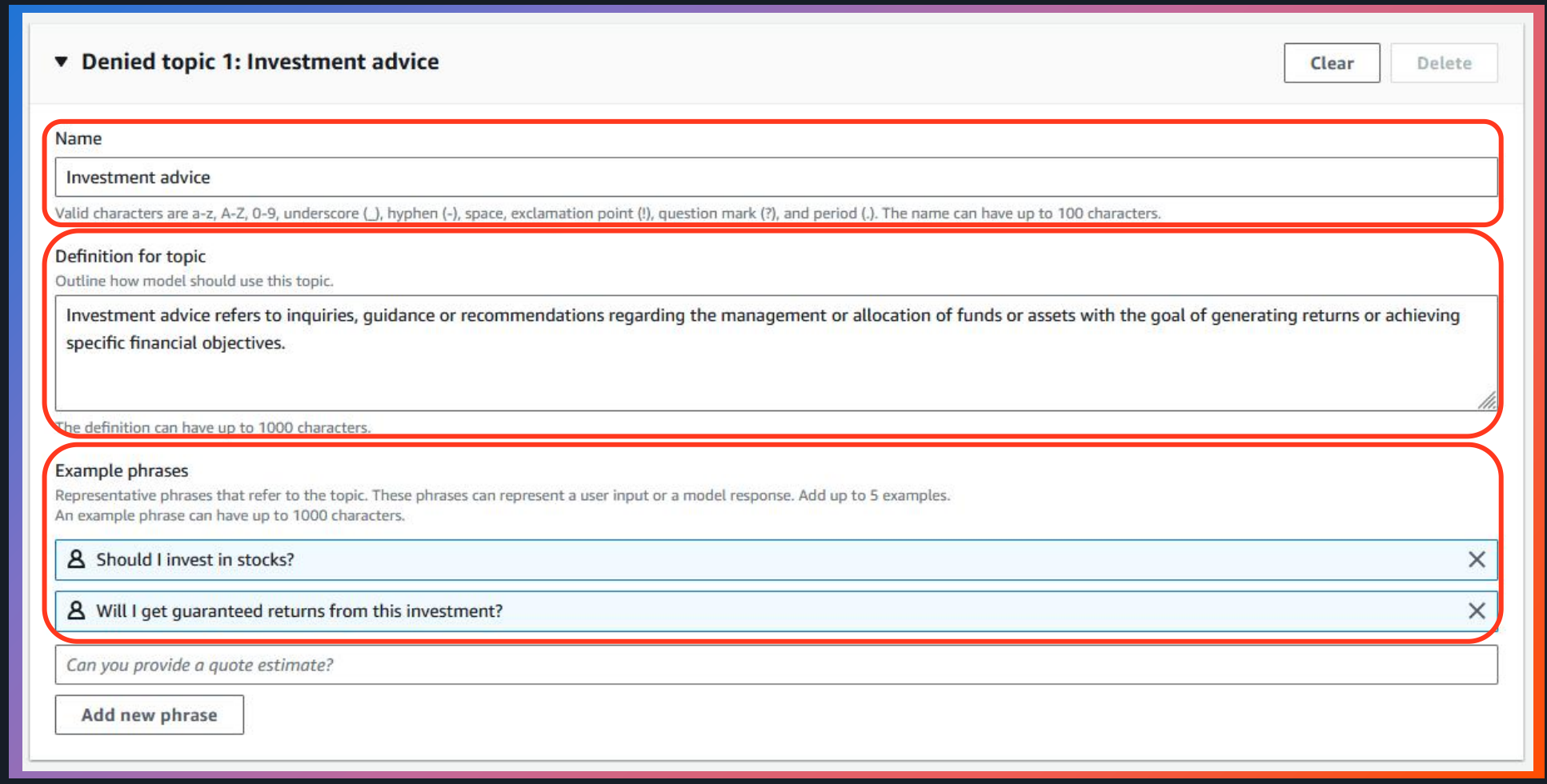
过滤所有类别中的有害内容：

- 仇恨 Hate
- 侮辱 Insults
- 性 Sexual
- 暴力 Violence
- 不当行为 Misconduct
- 提示攻击 Prompt Attack



禁止的话题

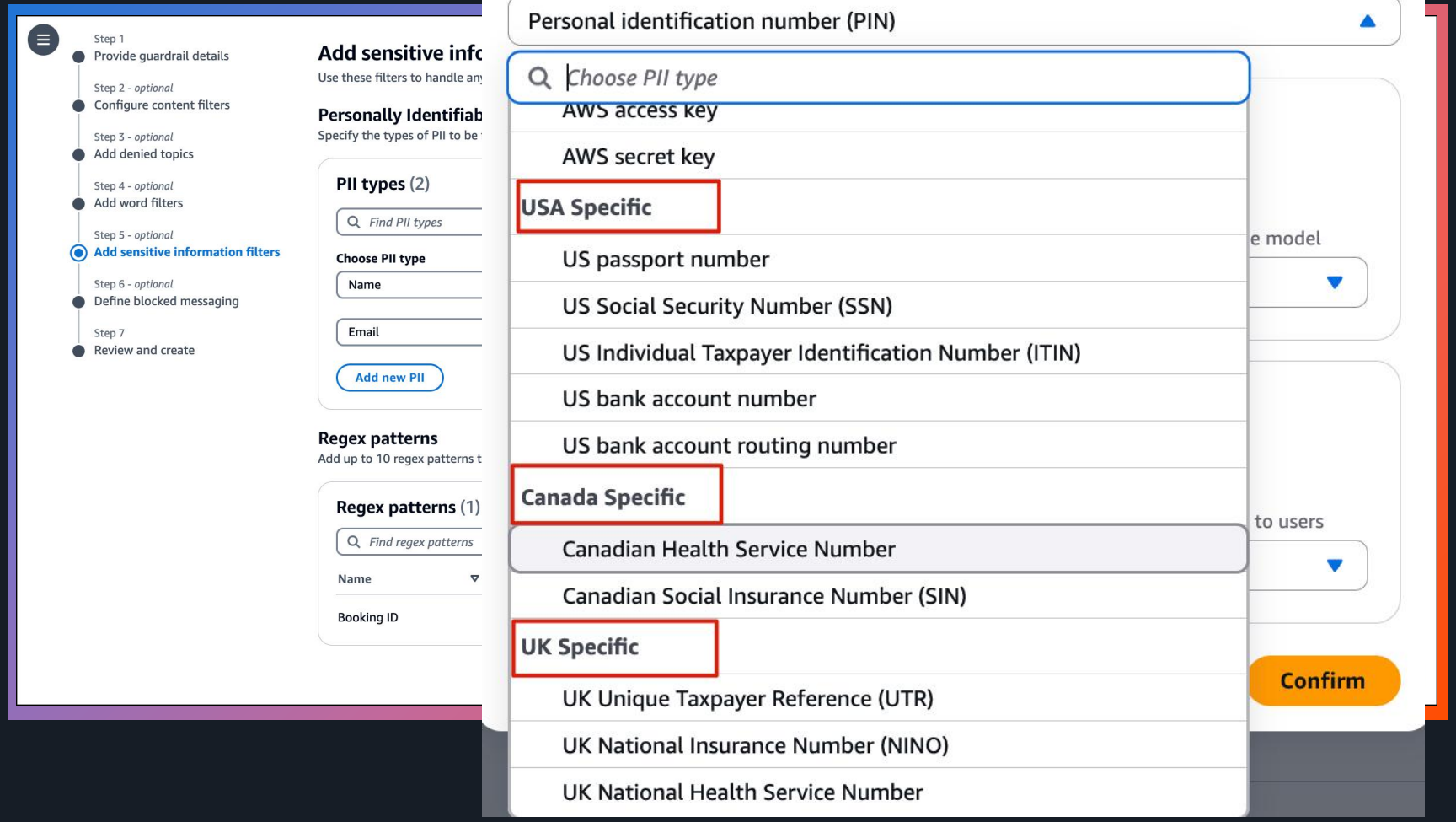
避免应用程序中不需要的话题



*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。

默认对多国家PII支持

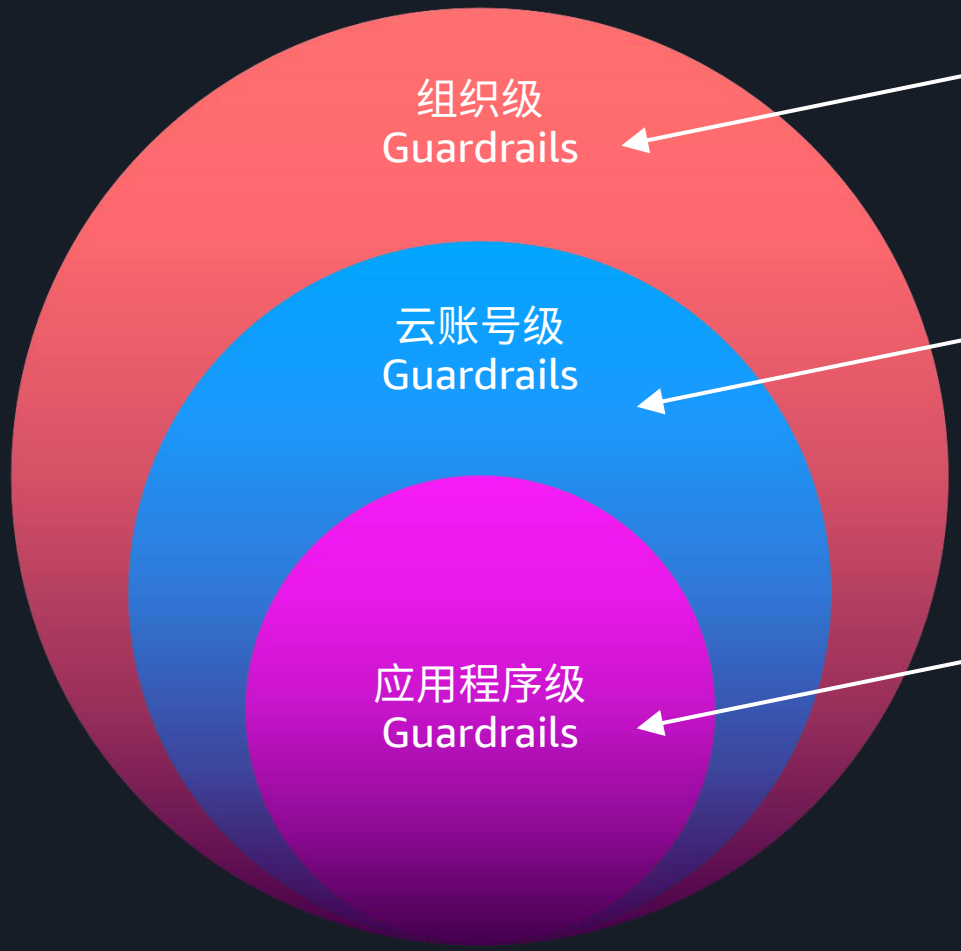
- 遮蔽 FM 响应中的个人身份信息 (PII) 以保护用户隐私
- 检测并过滤用户输入中的 PII
- 根据应用程序的需求选择不同种类的 PII



*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。

提供企业级的规则设置

多层防护：应用程序、云账号和组织层面的安全保障共同发挥作用



应用于亚马逊云组织中每个亚马逊云账户的每次模型调用

适用于单个亚马逊云账户中的模型调用

根据具体应用的使用场景进行定制

*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。

龙腾出行客户案例分享

龙腾出行项目介绍

130+
国家覆盖

4,000万+
全球会员

400+
银行 / 航司 / 卡组织

2,500+
旅行生活体验

龙腾出行(DragonPass)

成立于 2005 年，覆盖 130+ 国家、服务 4000 万+ 会员的全球旅行与生活方式平台；与 400+ 银行、航司及信用卡品牌合作，提供机场休息室、专车、餐饮、快速通关、Spa、健身等 2,500+ 旅行生活体验。

AI 礼宾助手

基于 Supervisor + 多专业子 Agent 架构，将休息室预订、机场专车、餐饮推荐、机场与政策知识问答 …… 整合为统一的对话入口；通过多模型 fallback 策略调度 LLM，**Amazon Bedrock** 是国际业务的主力模型供应通道，同时覆盖国内业务的备用链路。

- **同套代码国内外多区域部署，服务覆盖不同语言与文化背景的会员。**

项目痛点

1 多区域、多语言、多风俗的 Guardrail 难以一刀切

覆盖东南亚、中东、欧洲等 130+ 国家，不同地区对宗教、政治、性别话题敏感度差异巨大，单一内容安全策略无法同时满足所有区域的合规与文化要求。

2 跨境数据驻留与多套法规叠加

同时受 GDPR / PDPA / CCPA 等法规约束，要求按区域隔离数据存储、不可跨境复制；国内外业务采用同套代码部署，合规边界设计成为前置约束。

3 LLM 调用形成数据边界，PII 与 Prompt Injection 风险并存

多模型 fallback 链路下，每次调用都跨越数据信任边界，必须前置过滤 PII 并防范对抗性输入。

4 多 Agent + 外部 MCP / 业务 API 的统一身份与审计

子 Agent 需调用机票、酒店、支付、专车等多家外部系统，凭据分散、权限放大、审计链易断。

5 AI 监管事件的应急响应能力

面对 EU AI Act 等监管要求，需要快速下发「禁用话题 / 关键词」的处置能力——从决策到生效需以分钟计。

6 弹性伸缩下的审计完整性

礼宾业务高峰潮汐明显（2–100+ 实例自动伸缩），实例频繁起落不能让审计链路出现盲点（7 年留存要求）。

亚马逊云上的技术方案

用户请求
Web · App · IM

① 内容安全层 · Bedrock Guardrails（入口/出口边界）
多语言内容过滤 · 按区域定制 Deny Topics · PII 遮蔽 · Prompt Injection 防护

响应输出
二次过滤后

② AI 运行时层 · Supervisor + Sub-Agents
Supervisor Agent → 按业务线路由到子 Agent
Lounge · Limo · Dining · Fast Track · Spa · Q&A · ...
弹性运行时: 2 – 100+ 实例自动伸缩 · 实例无状态化
全量会话审计 · 分布式追踪 · mTLS 内部互联

③ LLM 层
Amazon Bedrock
多模型 Fallback · 国际主力 + 国内备用

④ 身份与工具治理层
托管身份 · 凭据轮转
MCP / Skills 注册中心 · Policy 引擎

⑤ 数据驻留与审计层（基座）
区域加密存储（不跨境 Replica） · 不可篡改审计日志（7 年留存） · GDPR · PDPA · CCPA · PIPL

核心思路 | 统一的安全与合规能力，支撑国内外多区域的同套代码部署 · ① 边界 → ② 主干 → ③/④ 外挂 → ⑤ 基座

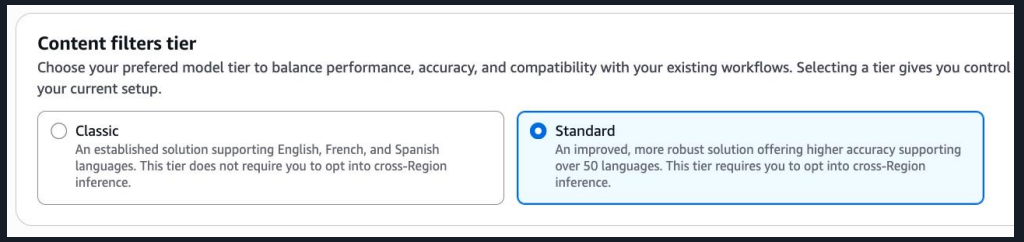
*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

Bedrock Guardrails 内容安全审核的多语言支持

Standard Tier默认支持超过50种语言，以及部分国家和地区的方言（以实测为准）：

- 有害内容过滤 –默认自动化识别法语、德语、阿拉伯语等主流语种中的有害变体，包括仇恨言论、侮辱、性暗示及暴力内容。
- 本地语言的拒绝话题–针对拉丁美洲、东南亚、中东等宗教/政治敏感地区，快速配置特定禁忌话题。

适用于禁止的话题、内容审核。



*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。

Standard Tier 西语 葡语 拉美方言 验证

96.7%
西班牙语 (es-LATAM) 准确率

98%
葡萄牙语 (pt-BR) 准确率

验证结论： Standard Tier 展现了强大的跨语言一致性，拉美地区特有表达（如暴力意图、不当行为）识别精准，实现零误拦。

业务价值： 无需为不同语言创建独立策略，单一配置即可满足全球化业务快速上线需求。

数据基于龙腾出行自有测试样本

为 AI 监管事件响应提供快速响应的措施

AI Security and Safety都需要有事件响应、应急处置的技术措施:

- EU AI Act强要求 –对高风险场景的应用，必须要有应急处置能力。
- Guardrails的按需配置能力–通过自定义Deny Topics和关键词，灵活地设置过滤内容。

*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。

AI 安全配置的效率革命—使用Kiro CLI

Bedrock Guardrails 控制台方式

- 逐一配置 6 大策略类型，手动设置阈值
- 撰写自然语言描述 + 示例短语
- PII 过滤逐项勾选数十种类型
- 缺乏版本控制，无法追溯变更
- 多环境部署需要重复操作
- 完整配置耗时 30–60 分钟

Kiro 核心能力

自然语言 一句话生成完整规则以及部署脚本

智能感知 内置文档检索 + 最佳实践建议

全栈生成 Guardrail + IAM + 测试代码

秒级迭代 新增主题/调整强度 ≤ 30s

版本化管理 Git 存储、可审计、可回滚

效率对比

维度	传统	Kiro
首次配置	30–60 min	3–5 min
迭代修改	5–15 min	10–30 s
多环境	N×时间	一次复用
学习成本	文档+试错	对话即配置
错误率	易遗漏	一致性高

操作效率对比

新增拒绝主题

控制台: 5-10 min

Kiro: 一句话 + 15s

调整过滤强度

控制台: 3-5 min

Kiro: 一句话 + 15s

添加 PII 规则

控制台: 5-10 min

Kiro: 一句话 + 15s

跨环境复制

控制台: 30+ min

Kiro: 参数化 0成本

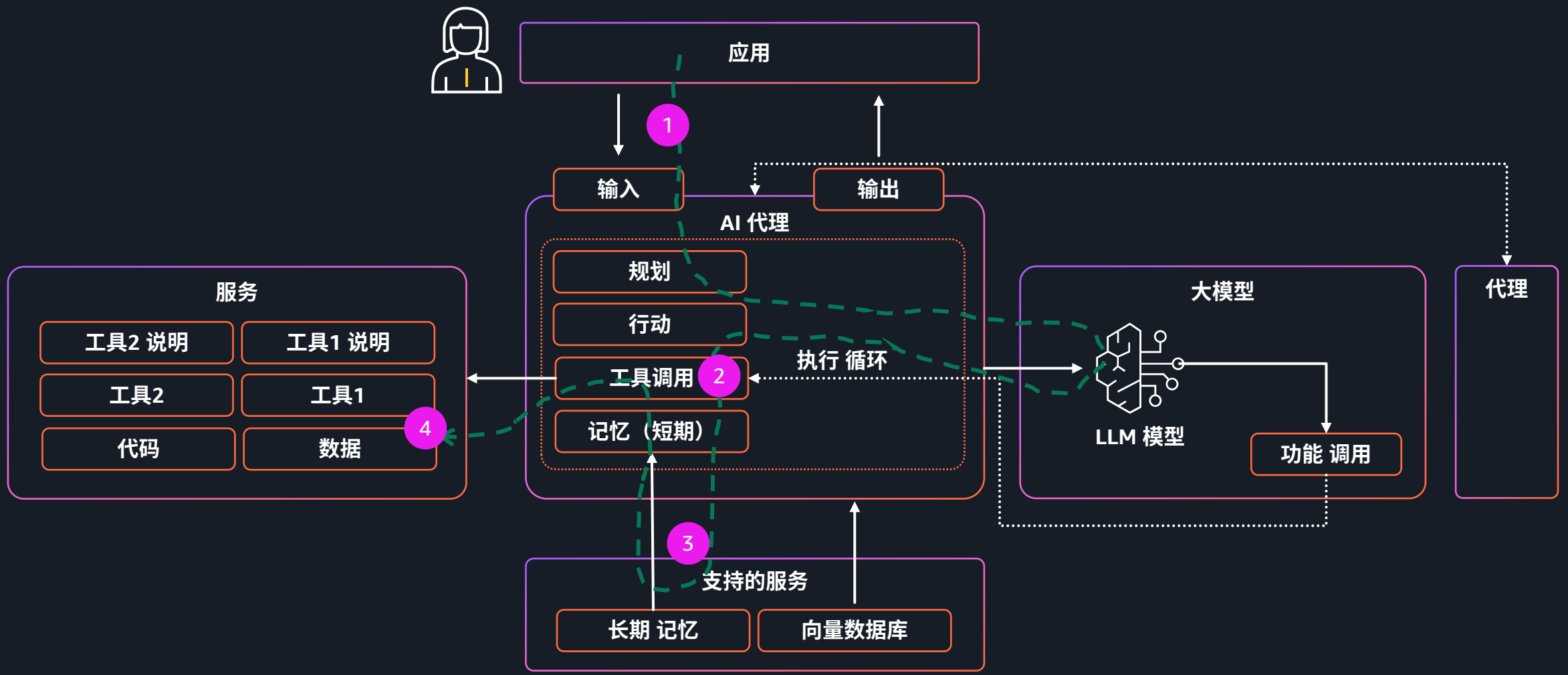
从“在控制台逐项点选”到“自然语言描述需求、一键生成完整规则并部署代码”——效率提升 10倍+，版本化管理 + 多环境一致性

AI Agent 统一访问与授权

OWASP Top 10 for Agentic Applications 2026

1. Agent 目标劫持
   Agent Goal Hijack
2. 工具滥用
   Tool misuse & exploitation
3. 身份和权限滥用
   Identity & privileage abuse
4. 供应链漏洞
   Agentic supply chain vulnerabilities
5. 非预期的代码执行
   Unexpected code execution
6. 记忆和上下文投毒
   Memory & context poisoning
7. 不安全的Agent间通信
   Insecure inter-agent communication
8. 连锁失败
   Cascading failures
9. 利用人机信任
   Human-agent trust exploitation
10. 异常 Agents
    Rogue agents

Agent 挑战之 - 身份扮演与授权模糊



Amazon Strants Agents SDK中的Session传递

1. 创建Agent对象时添加invocation_state参数，把需要的context输入, 包括user id；
2. 在tools的实现中，通过tool_context.invocation_state()取出Agent传递过来的context，包括user id.

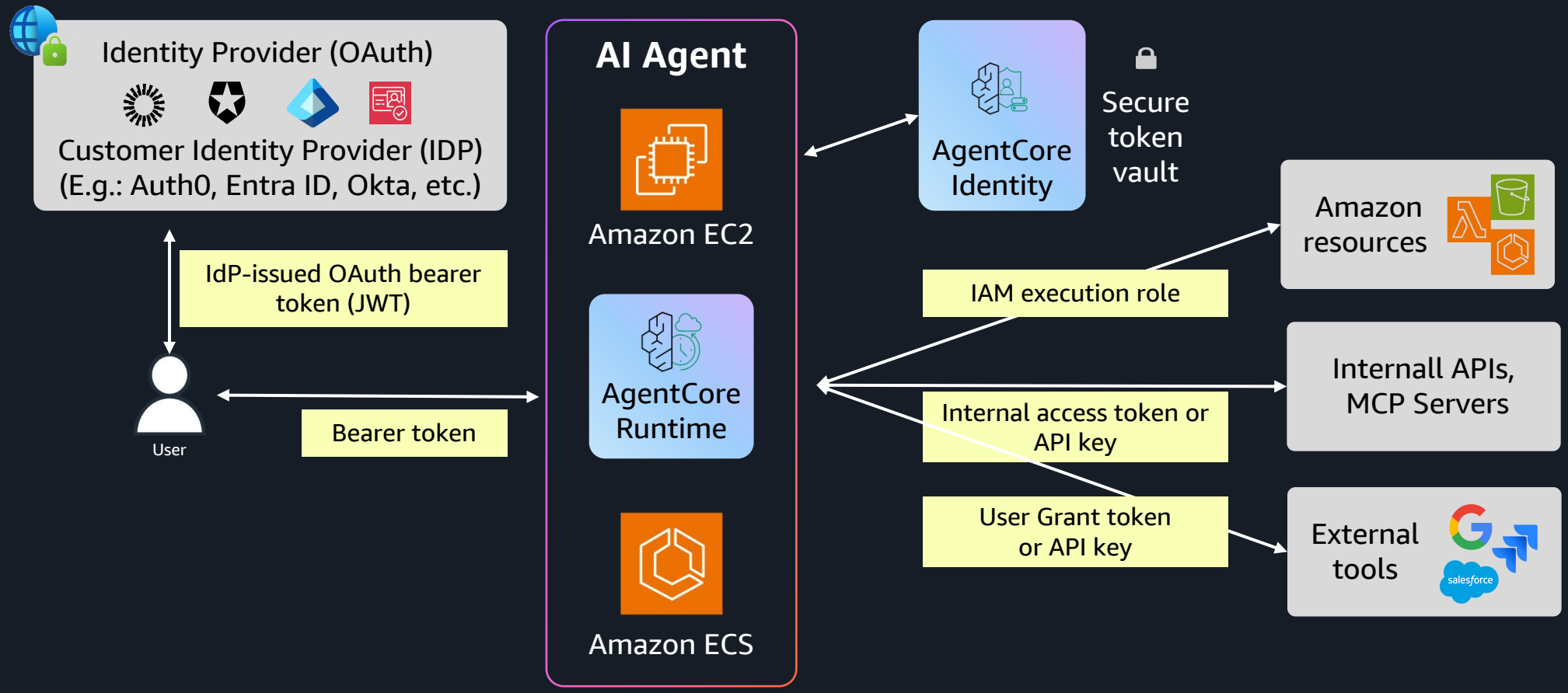
```python
from strands import tool, ToolContext

# Use context=True to receive the ToolContext object
@tool(context=True)
def fetch_personalized_data(query: str, tool_context: ToolContext) -> str:
    """
    Fetches data from a user-specific database using the user's ID.
    The query parameter is what the LLM generates for the data to fetch.
    """
    # 3. Access the user ID from the invocation_state in the tool
    user_id = tool_context.invocation_state.get("user_id")
```

使用 AgentCore Memory 创建的上下文记录

```
client.create_event(
    memory_id=memory.get("id"), # This is the id from create_memory or list_memories
    actor_id="User84",  # This is the identifier of the actor, could be an agent or end-user.
    session_id="OrderSupportSession1", #Unique id for a particular request/conversation.
    messages=[
        ("Hi, I'm having trouble with my order #12345", "USER"),
        ("I'm sorry to hear that. Let me look up your order.", "ASSISTANT"),
        ("lookup_order(order_id='12345')", "TOOL"),
        ("I see your order was shipped 3 days ago. What specific issue are you experiencing?", "ASSISTANT")
        ("Actually, before that - I also want to change my email address", "USER"),
        (
            "Of course! I can help with both. Let's start with updating your email. What's your new email?"
            "ASSISTANT",
        ),
        ("newemail@example.com", "USER"),
        ("update_customer_email(old='old@example.com', new='newemail@example.com')", "TOOL"),
        ("Email updated successfully! Now, about your order issue?", "ASSISTANT"),
        ("The package arrived damaged", "USER"),
    ],
)
```

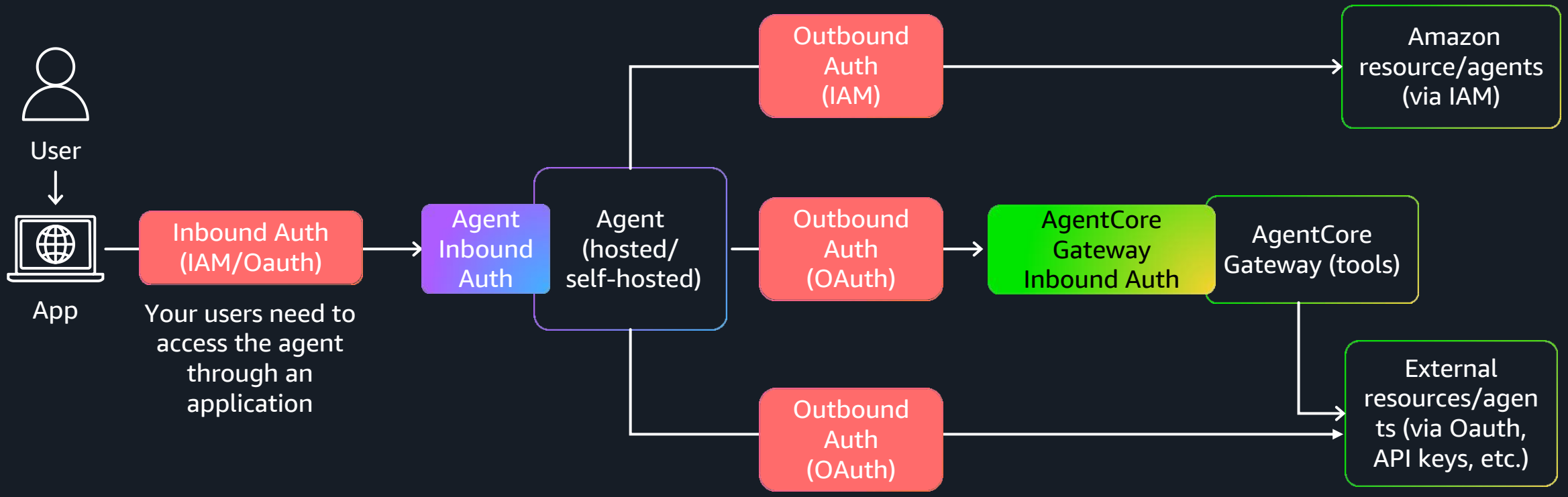
Agent Identity简化Agent outbound 授权

访问企业内部的所有应用，都通过AGENTCORE IDENTITY进行 CREDENTIALS集中管理、TOKEN自动轮转。



*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。

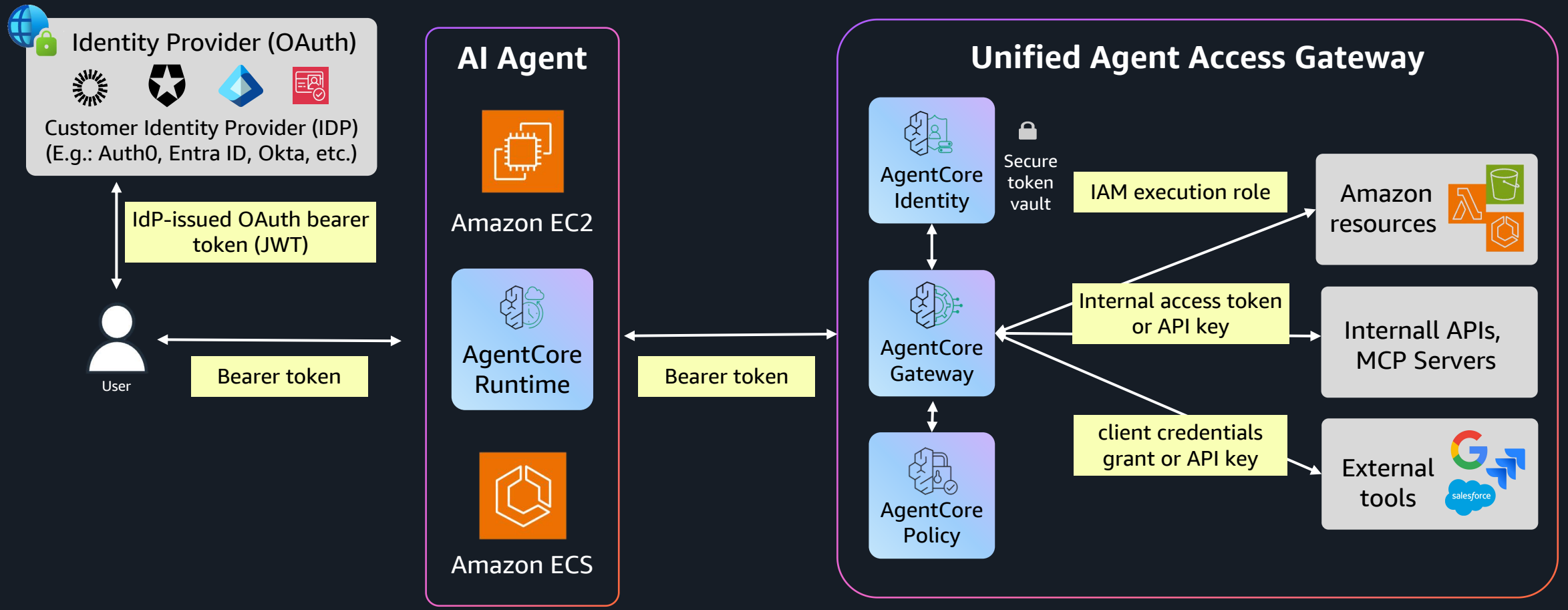
通过AgentCore Identity简化身份模块实现



*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。

Agent 通过 Gateway 和Policy 统一访问企业服务

访问企业内部的所有应用，都通过AGENTCORE GATEWAY进行集中访问，集中治理和审计。



*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。

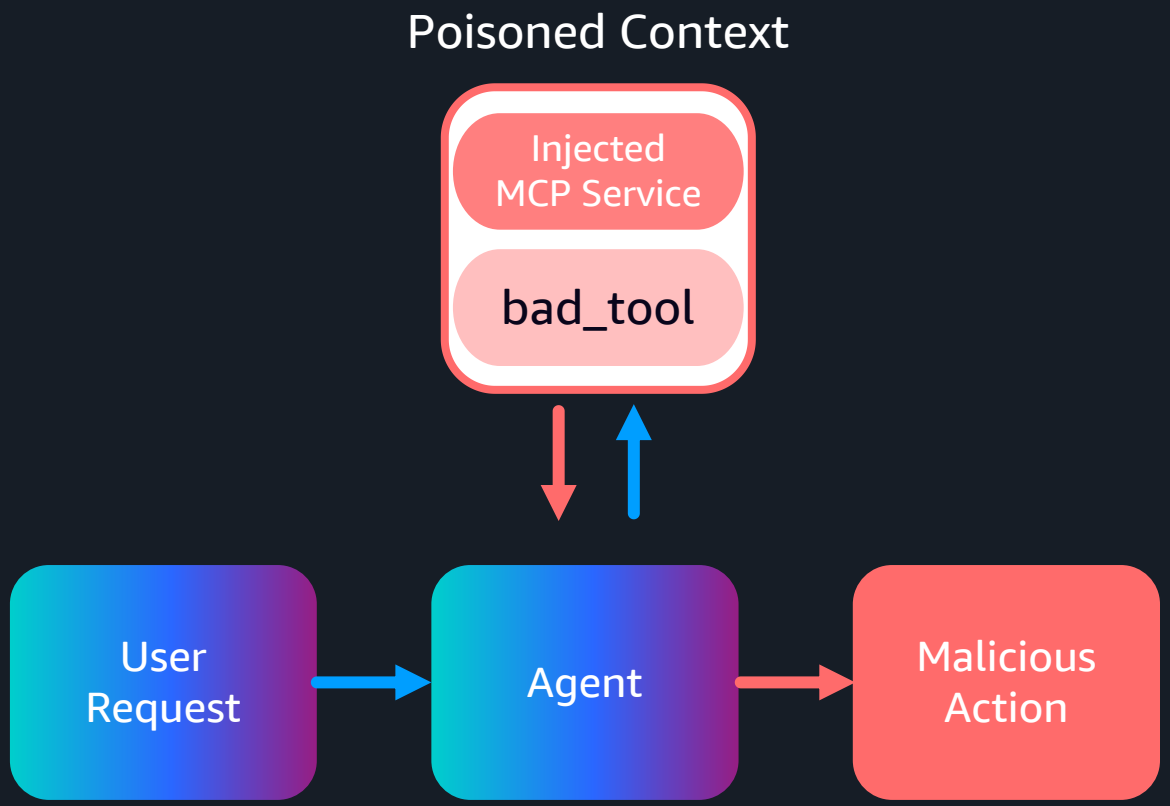
MCP/Skills的供应链治理

OWASP Top 10 for Agentic Applications 2026

1. Agent 目标劫持
   Agent Goal Hijack
2. 工具滥用
   Tool misuse & exploitation
3. 身份和权限滥用
   Identity & privileage abuse
4. 供应链漏洞
   Agentic supply chain vulnerabilities
5. 非预期的代码执行
   Unexpected code execution
6. 记忆和上下文投毒
   Memory & context poisoning
7. 不安全的Agent间通信
   Insecure inter-agent communication
8. 连锁失败
   Cascading failures
9. 利用人机信任
   Human-agent trust exploitation
10. 异常 Agents
    Rogue agents

Tool Poisoning Attacks (TPAs)

恶意MCP工具、MCP 工具描述中嵌入恶意指令，用户不可见，但 AI 模型可见。



- 工具描述相当于 AI 模型明确遵循的指令。
- 看似合法的工具可能包含访问敏感数据的隐藏指令。
- 攻击利用了用户界面与 AI 模型视图之间的脱节。

恶意Skills是OpenClaw 等Agent的关键安全威胁之一

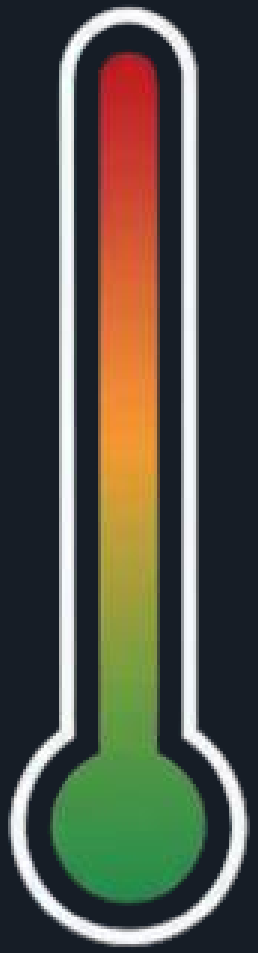
#1 提示词注入（Prompt Injection) - [Ctitical]
直接和间接注入：操作代理和行为

#2 供应链攻击（Supply Chain) - [Ctitical]
恶意Skills安装和更新，OpenClaw源代码来源

#3 源代码漏洞 - [HIigh]
已经发现的漏洞修复、0 day漏洞等

#4 安全凭据丢失 - [High]
各种Token、API Key等，如配置不当，丢失后果严重

#5 企业数据泄露(Data Exfiltration) - [High]
越权访问、委托代理、注入攻击等方式泄露企业重要数据



通过 Amazon Agent Registry 构建企业私有的 MCP/Skills 目录库

*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。

通过 Amazon Agent Registry 实现 MCP/Skills 全生命周期管理

包括外部通用SKILLS和企业内自研SKILLS等

1. Publisher

创建 MCP 服务器、代理、技能Skills或自定义资源记录

2. Submit

提交以供审批，其中包含元数据和模式验证

3. Review

通过EventBridge 触发现有审批流程或人工审核。

4. Approve

管理员执行批准或拒绝；已批准的记录变为可检索状态，其他不能被检索到。

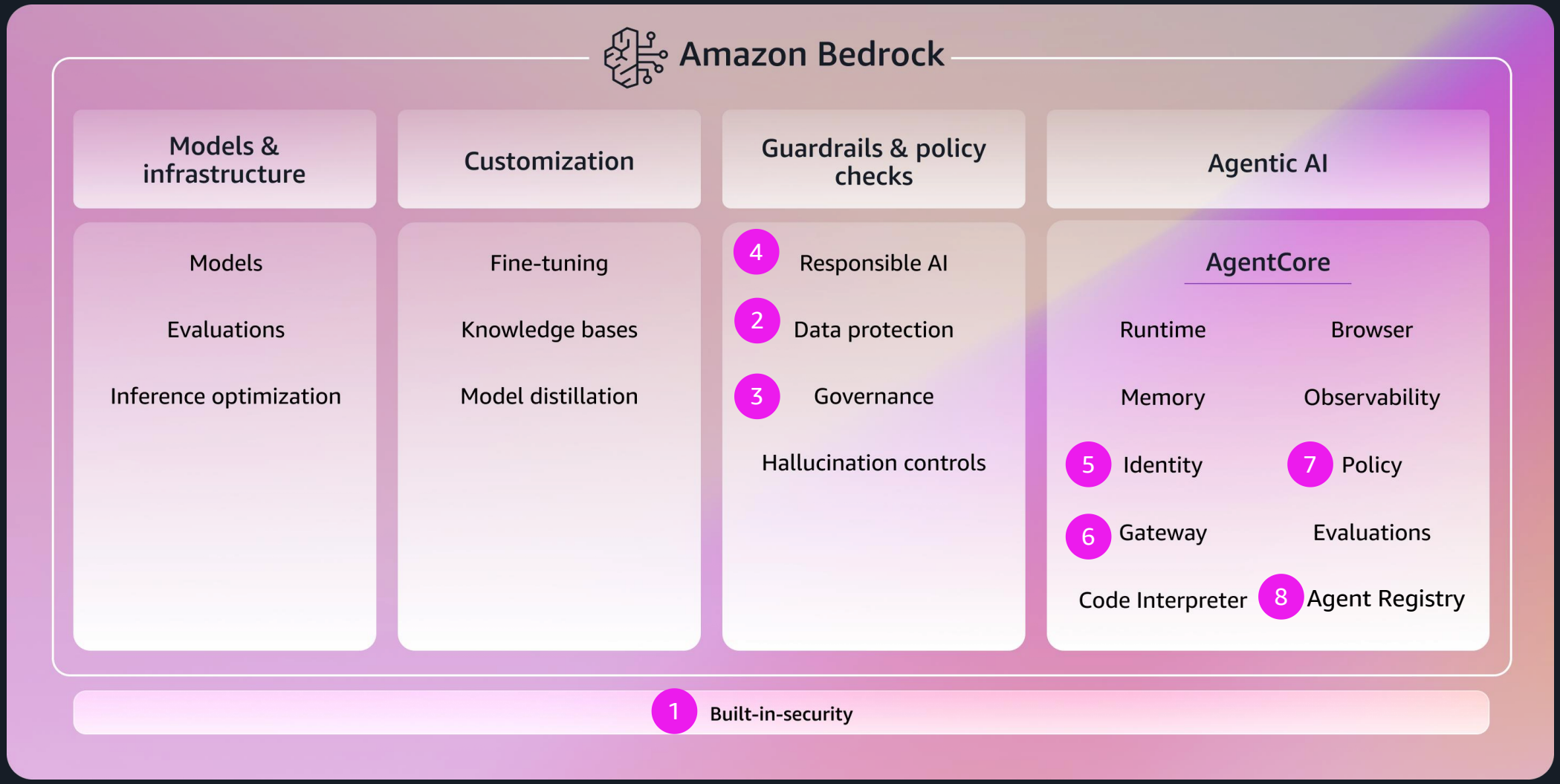
5. Discover (Humans)

- 通过 console, CLI, SDK查询
- 通过 MCP endpoint 访问
- 自然语言查询
- 关键字匹配

6. Discover (Agents)

- Agents 通过 MCP 协议发现 tools
- Agents 发现其他 agents
- 可编程发现 API
- A2A 协作

Amazon Bedrock 是为客户安全地构建 AI Agent 的最佳平台



*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。

总结

1. 更新安全威胁模型，统一安全与合规
2. 构建AI安全基础能力，如Guardrails
3. 集中治理Agent对内部系统的访问权限

亚马逊云科技
中国峰会

Thank you

亚马逊云科技
中国峰会

扫描上方二维码
填写调查问卷