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

制造业企业

AI 平台 从 0 到 1

不是从最顶的开始 — 而是从最基础的开始

庄颖勤

亚马逊云科技 Solutions Architect

赵舒霞

韶音科技 AI 建设负责人

BREAKOUT SESSION

30 min

今天的 30 分钟：四段路线

Agenda — 我们会讲什么 · 现在讲到哪 · 谁来讲

I	**框架与选型逻辑** 五层栈模型 · 用对资源（SA + FDE）	SA 庄颖勤 · 约 5 min
II	**韶音落地现状** 安全纵深管控 · 统一模型网关	客户 赵舒霞
III	**实战的坑与可观测** 三个真实的坑 · 从不可见到全可见	客户 赵舒霞
IV	**下一阶段与原则** Agent 规划 · 三条原则 · 行动清单	SA 庄颖勤 · 收尾

— 接下来 30 分钟，

由我们的客户分享
实际落地实战

客户
实战

赵舒霞

韶音科技 Shokz · AI 建设负责人

- 🌍 连续三年全球开放式耳机品牌出货量第一
- 🏢 60+ 国家销售，总部深圳
- 🚀 AI 成熟度：分散项目落地 → 业务规模化落地期

亚马逊云科技上海峰会独家首发 · 实战全披露

* 根据 Omdia 最新发布的数据显示：2025 年全年，韶音全球开放式耳机出货量达到 930 万台，占整体全球开放式耳机出货量的 19%，位列全球第一。
根据 Omdia 最新发布数据显示：自 2023 年起至 2025 年止，Shokz 韶音品牌开放式耳机出货量连续 3 年保持全球第一。

韶音科技 AI 平台：五层技术栈定位

全球骨传导耳机领先品牌 · AI 成熟度：分散落地 → 业务规模化期

L5	**Agents** 📋 明确需求	AI 开发流水线 | AI 设计 | 用户研究 | 算法研究
L4	**Agentic 平台** 🟡 建设中	AgentCore 评估 | MCP | 治理 | 安全
L3	**数据与知识** ✅ 已完成	S3 + Glue + Redshift POC | RAG / 向量库 | RDS
L2	**模型层** ✅ 已完成	Shokz Gateway | Claude / GPT / DeepSeek / Qwen
L1	**基础设施** ✅ 已完成	Amazon Web Services Cloud | 网络安全架构 | VPC | On-premise

关键发现

⚠ AI 使用分散

各业务独立引入，个人账号，团队账号混用，缺统一治理

⚠ 数据孤岛

多业务系统与渠道数据待打通

✅ 网关已就位

统一模型访问 + 成本追踪

✅ 安全基线完成

VPC + 网络架构 + 合规框架

亚马逊云科技
中国峰会

为什么要建 L2 AI GateWay 平台

真实的情况是.....

没有统一网关与治理，AI 工具在企业里就这样“裸奔”

成本	**Token 爆炸**	直接分发 API Key，次月账单远超预期
安全	**数据泄密**	新品 / 研发信息会不会被喂给外部 AI?
效率	**模型错配**	用最贵的模型，做最简单的事
安全	**Token 异常**	API Key 泄露或被盗用，无从追溯
运维	**配置复杂**	一个项目至少 3 个 Key 起配，难以统一管理

L1 基础设施：纵深安全管控

设计理念：Shokz Gateway（基于 LiteLLM）是唯一 AI 出口 + VPC Endpoint Policy 兜底

第 6 层	**CloudTrail 审计**	全量操作日志
第 5 层	**GuardDuty**	异常行为检测
第 4 层	**IAM Policy**	最小权限原则
第 3 层	**VPC Endpoint Policy**	限制可访问服务（兜底）
第 2 层	**SG + NACL**	网络层过滤
第 1 层	**域名白名单 (DNS)**	只允许已批准域名

Shadow AI 如何被堵住

场景 1：员工直连未审批的外部 AI 服务

→ DNS 层拦截，域名不在白名单

场景 2：绕过网关调 Bedrock

→ VPC EP 限定可访问服务 + IAM 限定仅网关 Role 可调

场景 3：新模型服务偷偷接入

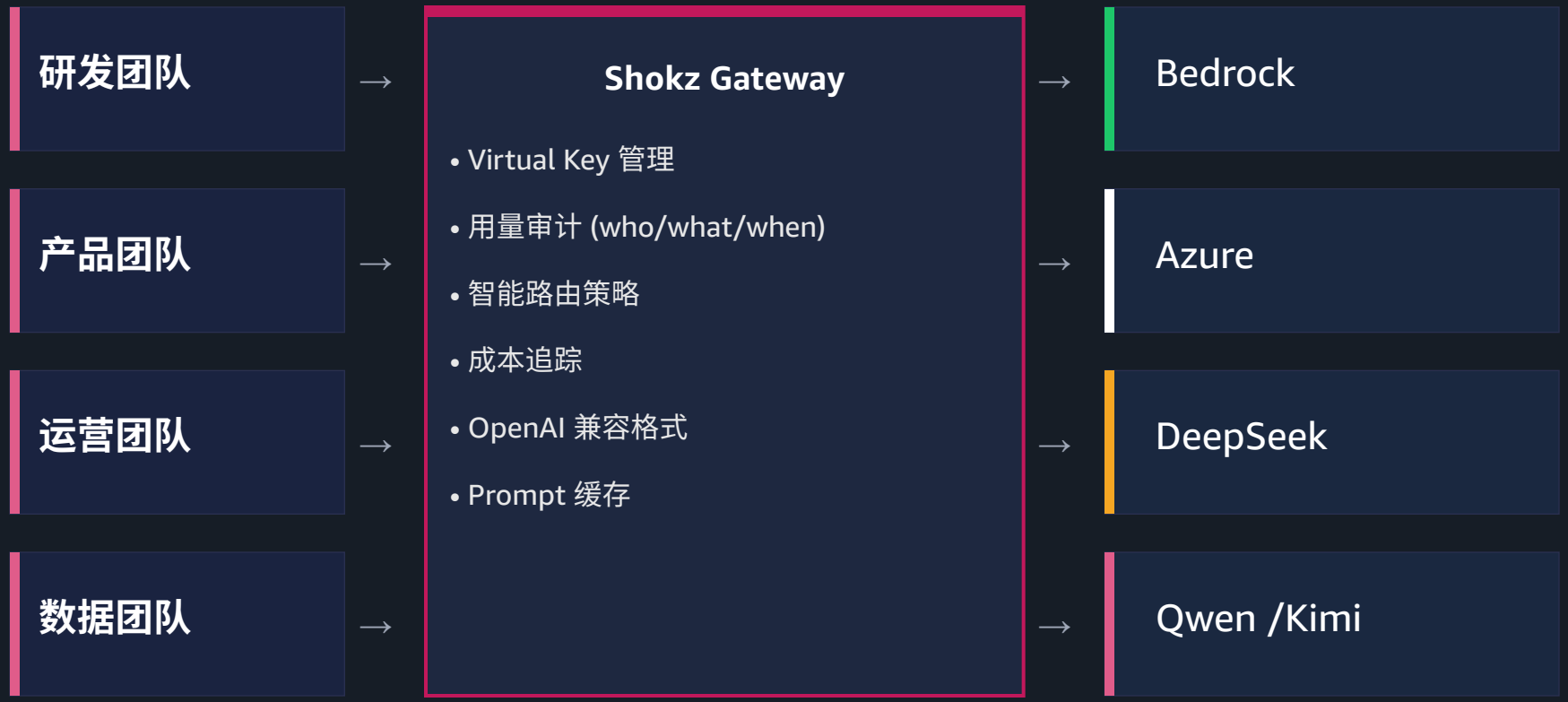
→ 网络层 + IAM 双重拦截

✅ **结果：受管网络内未发现绕过网关的 AI 调用**

新模型申请流程：申请 → SA 评审 → 2 工作日内开通

L2 模型层：Shokz Gateway 统一网关架构

Shokz Gateway + Amazon Bedrock：网关做路由审计，Amazon Bedrock 做模型供给



💡**选型结论：**自研网关（周期长） vs Bedrock-only（多模型受限） vs **LiteLLM + Bedrock（最迅速上线 + 多模型）**✅
对开发者：改一行 URL 即可接入 | 对管理者：经网关的调用，who/what/when/how much 全量可审计

两个真实的坑 — 韶音怎么趟过去的

从 Day 0 到能跑，实打实的踩坑与解法

i.

跨境延迟

- 网关在海外，大陆用户体感差
- 解法：就近路由 + Prompt 缓存
- 体感延迟显著改善

制造业 AI 落地的第一个体验 Gap：不是模型慢，而是链路长

大陆用户访问海外 AI 服务时，跨境链路会把首 Token、整体响应与抖动一起放大

为什么慢？ 网络传输 + 模型推理 两段时间叠加，用户只感知“慢”

中国用户 → 企业出口 → **跨境链路** → 境外 AI 网关 → 模型推理

用户体感：等待首字、请求排队、偶发卡顿

就近路由

Nearest Path / Routing

- 优化企业出口与访问路径
- 减少跨境跳数与链路抖动
- 让首 Token 更早出现

Prompt Cache

Repeat Question Acceleration

- 制度 / SOP / FAQ 高复用
- 相同上下文直接命中缓存
- 同时降低 Token 成本

体感延迟、访问不稳定显著改善

首 Token 更快
用户不再盯着空白页面等响应

重复问题秒回
高频知识问答直接吃缓存

成本一起下降
少算一次，就少付一次 Token

💡 路由解决“路太远”，Cache 解决“重复算”；两者叠加后，用户感受到的是整体速度变快。

两个真实的坑 — 韶音怎么趟过去的

从 Day 0 到能跑，实打实的踩坑与解法

ii.

团队精力分配与交期

业务不会停，安全不能等

- 解法：不造轮子+合作伙伴全力合作
- 两周后团队可自主运维

制造业 IT 团队的精力 Gap 不是障碍 — 关键是用对资源

多线并发 IT 团队，怎么撑 AI 平台？SA 架构师 + FDE 驻场 双轨模型

SA 架构师
Solutions Architect

- 架构设计 · 最佳实践
- 技术选型 · 风险评估
- 商业价值对齐

FDE 驻场
Field Delivery Engineer

- 现场执行 · 快速交付
- 运维支撑 · 日常答疑
- 技能转移

1 人 IT 团队 能力地图

网络/基础架构	安全合规	数据工程	AI/ML基础	业务理解

SA + FDE 补齐 AI 架构和执行短板，内部团队专注业务知识 + 数据资产

从不可见 到 网关流量全可见

网关让经网关的 AI 调用 100% 变成可测量的数字资产

97 %

经网关的 AI 调用均被审计追踪

实施前

- 不知道谁在用哪个模型
- 账单来了才发现超支
- Shadow AI 无从排查

实施后：全量可视仪表盘

模型使用分布
多模型并存，分布全量可见（按多维度下钻）

团队用量排名
各团队用量清晰可比，识别重点消耗方

成本控制
通过缓存 + 路由优化，持续压降 API 成本

响应时间
端到端延迟与错误率纳入监控，可观测、可告警

安全事件
Shadow AI 收敛至单一入口，非法调用实时拦截并留痕

*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。

下一阶段：从管住模型到管好 Agent

The roadmap — 韶音 AI 平台 2026 H2 规划

已完成 ✅

L1-L2 基础层

- Shokz Gateway（基于 LiteLLM）
- 统一模型访问
- 纵深安全管控
- 全量审计可视

进行中 🟡

L3-L4 平台层

- L3: Glue + Redshift 数据层
- L4: AgentCore 评估中
- MCP 网关测试
- RAG 知识库已建设

落地中 📋

L5 Agent 应用层

- AI 设计 agent
- AI Coding 全流程
- 算法研究
- 用户研究

💡 AgentCore 是关键决策节点：是否采用 亚马逊云科技托管 Agent 框架，还是继续自建?
评估维度：运维复杂度 vs 灵活性 vs 与 Bedrock 生态集成深度

统一网关的真实部署拓扑

从逻辑到落地 — 韶音生产环境参考架构（已脱敏）

说明：本图为脱敏参考架构，省略真实账号 / VPC / 资源标识；模型与区域可用性以亚马逊云科技官方文档为准。

*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。

可观测性的真实实现：Serverless 审计链路

每次模型调用 → 可查询的审计资产（已脱敏 · 全 Serverless · 按用量计费）

① 数据采集与转换

ECS · Shokz Gateway
每次调用写审计记录

↓

审计表（DynamoDB）
写入调用明细

↓

Streams → Lambda
变更触发 / 转 NDJSON

↓

审计存储（S3）
Hive 分区 year/month/day

② 数据治理与查询

Glue Catalog
Partition Projection（免维护）

↓

Athena Workgroup
SQL 分析审计日志

↓

查询 API（API Gateway）
REST 查询接口

↓

Lambda（query）
执行查询逻辑

③ API 服务与看板

Cognito（JWT 鉴权）
保护 API 访问

↓

React SPA
CloudFront + S3 托管

↓

📊 **审计看板**
用量 · 成本 · 模型统计

↓

👤 **管理员 / 审计员**
按角色访问

*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 0 到 1 的三条不二原则

What will be remembered — 带回去的三件事

01

先建网关，再建 Agent

没有网关，Agent 是裸跑。网关是 AI 平台的第一道基础设施，
不是可选项，是必选项。

02

安全不是事后补

纵深安全从 Day 0 开始建。等到出事再补，
成本高得多，而且往往补不完整。

03

用好资源，不要硬干

SA + FDE 是加速器，不是依赖。
IT 团队的上限不是技术，不是人力，是如何把有限时间花在刀刃上。

回去后：本周与本月的 6 件事

Take-home actions — 今天听完马上能做的事

This Week

部署 LiteLLM 网关（POC 环境）

2h 搭起来，接入 1 个团队先跑

梳理公司现有 AI 使用清单

发一封邮件，收集各部门在用什么 AI

网络层基础检查

确认 VPC、SG、DNS 白名单机制是否到位

This Month

网关全量上线

所有团队统一走网关，老接口迁移

建立用量报告机制

每周一张成本 +用量报告发给 IT 负责人

规划 L3 数据层 POC

选 1 个数据源做 Glue + RAG 小 Demo

Thank & Q/A

“AI 平台从「0 到 1」—

不是从最顶的开始，而是从最基础的开始。”

庄颖勤 亚马逊云科技 Solutions Architect

赵舒霞 韶音科技 · AI 建设负责人

亚马逊云科技
中国峰会

Thank you

亚马逊云科技
中国峰会

Thank you