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

小红书大模型基础设施实践
Relax 与 MaaS 驱动的模型能力闭环

MaaS in the loop: Data Generation · Judge · Eval · Serving

杨睿

小红书 RedAI Infra

Multi-modal Agentic RL + Model Capability Infrastructure

朱文军

亚马逊云科技

解决方案架构师

Relax × MaaS

training-time and serving-time capability loop

Agenda

- 背景：复杂 Agentic RL 需要模型能力闭环
- Relax：训练编排、Rollout 与后端协同
- MaaS in the loop：数据生成、Judge 与评测
- 演进：从训练闭环到 AI Native 基础设施

Correction

MaaS 不是部署终点
而是训练全周期模型能力层

TALK ABSTRACT

MaaS 不是下游部署，而是 Relax 全流程依赖的模型能力底座

本次分享围绕 Relax 的 Agentic RL 基础设施，以及 MaaS 如何贯穿数据生成、Judge、评测和线上服务。

Relax

复杂交互任务训练
Rollout / Reward
训练后端与推理引擎协同

MaaS in-loop

训练前数据生成
训练中 Judge Model
Checkpoint 评测回归

Serving

Token 级生产调度
动态 Batch / KV Cache
成本与 SLO 治理

Amazon

弹性计算与权重分发
网络 / 存储 / 观测
AI Native 基础设施演进

主线：Relax 消费 MaaS 的模型能力完成训练闭环，MaaS 再承接训练结果与线上反馈，二者是互相驱动的基础设施关系。

亚马逊云科技
中国峰会

CHAPTER 01

背景与问题

Complex Agentic RL only matters when model capabilities can become online services.

BACKGROUND

三大核心瓶颈推动框架演进

从 SFT 到 RL 规模化，工程门槛随模型、模态和生产要求同步抬升。

01 训推耦合

Actor 与 Rollout 共享 GPU，推理和训练互相阻塞；资源利用率和吞吐都受限。

02 单模态局限

主流框架偏文本 RL，图 / 视频 / 音频端到端 RL 缺少统一工程路径。

03 可靠性缺失

大规模训练需要在线扩缩、驱逐处理、权重一致和服务级恢复。

Relax 的目标：解耦训推、统一全模态、提供生产级弹性容错

DESIGN PILLARS

三大设计支柱

同一套框架同时服务算法验证、大规模全异步训练和生产运维。

全模态统一

Qwen3 / Qwen3-VL / Qwen3-Omni
一份 YAML 切换文本、视觉、音频、视频

完全异步

Ray Serve 服务化
Actor / Rollout / Reference 独立 GPU 集群

生产级

REST API 扩缩容
HealthManager + DCS + Metrics 全链路托底

16xH800
evaluation cluster

GRPO+
RL algorithms

4/8/16
precision choices

Zero-Drop
elastic operations

TECH STACK

Megatron-LM × TransferQueue × SGLang

训练、数据、推理三大平面真正解耦；任意一层可独立升级。

Megatron-LM

TP / PP / CP / EP 完整并行
MoE 超大模型支持
Megatron Bridge：HF ↔ MCore

TransferQueue

StreamingDataLoader 流式消费
max_staleness 控制新鲜度
字段级 ready / backpressure

SGLang

高吞吐推理引擎
多模态 payload 接入
暂停 / 恢复 / 权重同步 API

工程边界清晰：训练后端、数据平面、推理引擎都可以单独替换

亚马逊云科技
中国峰会

CHAPTER 02

Relax：训练基础设施

Data generation, rollout, reward, training backend, and inference engine work as one system.

ARCHITECTURE

三平面正交解耦

任何一层独立升级，都不需要改动其他角色的运行方式。

Control Plane	Controller 发出 generate / advantage / step；按算法注册角色
Compute Plane	Training -> Megatron-LM；Inference -> SGLang；Reward 可插拔
Data Plane	TransferQueue 是唯一跨角色数据通道；字段级独立读写

Independent Upgradability

RAY SERVE

Relax 六层服务化架构

角色即服务，每层职责独立，故障域和扩缩容边界清晰。

Entrypoints	train.py · signal handling · Ray cluster	01
Orchestration	Controller · Service · Registry	02
Components	Actor · Rollout · Critic · GenRM	03
Engine	Rollout Engine · Reward · Router · Filters	04
Backends	Megatron-LM training · SGLang inference	05
Distributed	Ray Actor Groups · DCS Checkpoint Service	06

EXECUTION MODE

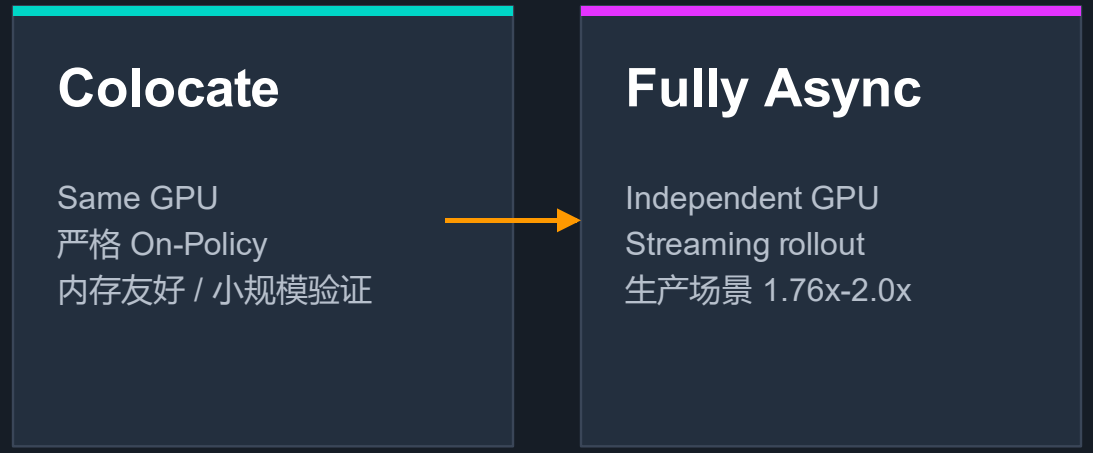
Colocate 与 Fully Async 自由切换

同设备严格同步用于验证；独立集群全异步用于大模型和慢 rollout 场景。

1.76x-2.0x

端到端吞吐提升

TransferQueue 解耦训推，Actor / Rollout 独立 GPU 集群，staleness 可调以平衡数据新鲜度与吞吐。



0-N
staleness control

0%
resource contention

2.0x
Qwen3-Omni-30B

亚马逊云科技
中国峰会

CHAPTER 02

Relax 核心机制

Throughput, stability, and extensibility for large-scale multi-modal Agentic RL.

TRANSFERQUEUE

训推解耦的核心数据平面

所有阶段都在 TQ 上读字段、写字段，以 micro-batch 粒度流式衔接。

01 Rollout

IN: prompts
OUT: responses

02 ActorFwd + Ref

IN: prompts + responses
OUT: actor/ref log_probs

03 Reward + Adv

IN: log_probs + responses
OUT: rewards + advantages

04 Train Actor

IN: advantages
OUT: new weights -> DCS

字段级 ready / Streaming 流水 / 背压容量公式：total = bs × (max_staleness + 1) × n_samples

DCS

分布式 Checkpoint 服务

权重同步控制面把 rollout 暂停、KV 清理、NCCL 同步和恢复串成可运维流程。

1	**Pause Rollout** HTTP /pause_generation
2	**Flush KV Cache** HTTP /flush_cache
3	**NCCL Broadcast** all_gather -> broadcast
4	**Resume Rollout** HTTP /continue_generation

< 5%

sync overhead
with overlap

<30s

pause/resume
latency

*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。

OPERATIONS

弹性扩缩与故障恢复能力

把 rollout 推理池从固定资源升级为可扩、可摘、可恢复的训练生产面。

扩容

ready 的 replica 立即进入 bring-up；超时实例标记失败，已成功实例保留。

缩容

先摘流，再释放，再清理拓扑；不低于初始池，不打断权重更新。

恢复

engine 级健康监控 + 服务级全局重启；checkpoint path 自动修正。

OMNI + AGENTIC

全模态 Agentic RL 的统一接入

数据、rollout、processor、Megatron 训练后端都围绕多轮 token 级任务和 agent 请求组织。

多模态训练

placeholder -> content lint
HF processor resize
ViT CP support
T2I + CP REF

MaaS Model Calls

teacher model 数据扩写
judge model 质量判断
checkpoint 中间评测

Example: DeepEyes

VLM tool call + 多轮搜索工具
训练时由模型服务提供可控判断
线上再进入统一 serving 链路

关键点：Agentic RL 的数据、Judge 和 Eval 调用都需要 MaaS 提供稳定、可治理的模型能力。

亚马逊云科技
中国峰会

CHAPTER 03

MaaS in the loop

承前启后：MaaS 不是 Relax 之后的部署平台，而是训练前、训练中、评测与服务共同依赖的模型能力层。

MaaS 的定位：Relax 训练链路里的模型能力层

把模型能力抽象成可调用、可调度、可观测、可治理的服务单元，供训练和线上同时消费。

Data Gen	Rollout	Judge	Eval	Serve
teacher / self-play 多模态样本生成 任务扩写与难例挖掘	环境交互 工具调用 轨迹采样	LLM-as-judge reward / rubric 偏好与安全判断	checkpoint 评测 回归门禁 能力面板	在线 API 动态 batch 成本治理

Relax 不是只把模型交给 MaaS 部署；Relax 在训练前、中、后持续调用 MaaS，MaaS 也从训练反馈中迭代服务策略。

MAAS 02

训练前：用 MaaS 做数据生成与任务扩展

高质量 Agentic RL 不是只等人工数据，而是通过模型服务批量生成、筛选和扩展任务分布。

Seed Tasks	业务种子问题 / 多模态上下文 / tool-use 场景 / 安全与风控边界
MaaS Calls	teacher models / self-play agents / prompt rewriting / trajectory generation
Data Output	候选样本 / 多轮轨迹 / hard cases / 初始评测集 / 数据质量 trace

MaaS 提供的是可规模化的数据生产能力，Relax 再把这些样本变成可训练、可回放、可评测的 RL 数据流。

训练中：Judge Model 与 Reward 计算依赖 MaaS

复杂任务的 reward 不只来自规则函数，还会调用 MaaS 上的 judge / critic / verifier 模型。

ROLLOUT
Relax 生成多轮轨迹
工具调用与环境反馈

→

JUDGE
MaaS 调用 judge model
rubric / pairwise / safety

→

REWARD
聚合规则分与模型分
advantage / filtering

→

TRAIN
Actor 更新
checkpoint 产出

多版本 Judge

不同任务可路由到不同 judge / verifier，支持灰度和回归对比。

训练稳定性

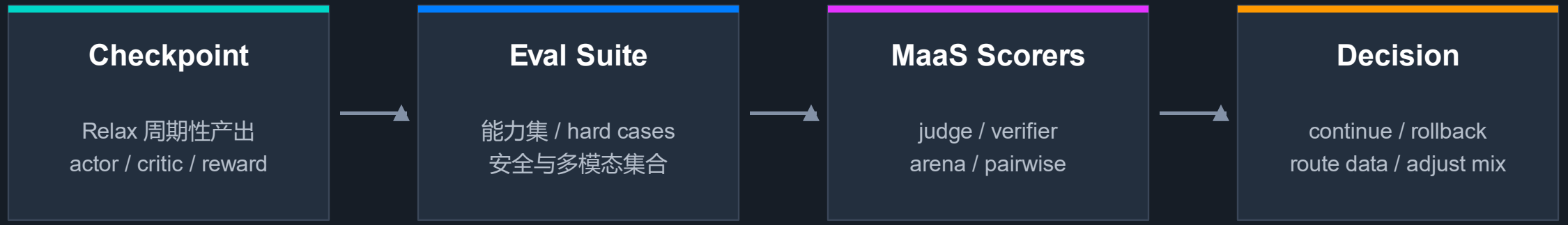
Judge 调用需要限流、超时、fallback，避免训练被服务波动拖垮。

成本可控

Reward 调用进入 token budget 与配额治理，训练吞吐和成本一起优化。

训练间：Checkpoint 评测与回归门禁依赖 MaaS

中间评测不是离线附属流程，而是决定继续训练、回滚、调参和采样策略的在线信号。



MaaS 让评测模型、线上候选模型和历史 checkpoint 可以被同一个服务面调度；评测结果再回流到 Relax 的数据采样和训练决策。

MaaS × Amazon：训练与服务共享的云基础设施底座

Amazon EC2 P6 训推一体：为什么 Relax 选择 EC2 P6

ODCR 统一资源池，训练与推理弹性切换，消除传统方案的资源割裂与搬运开销。

GPU 利用率

传统：训练、rollout、评估服务存在明显潮汐
空闲浪费 30-50%

EC2 P6：统一资源池
利用率 >85%

1.8 TB

权重搬运

传统：跨集群复制
延迟 >60s

EC2 P6：NVLink 互联
同步 <10s

6.4Tbps

Rollout 吞吐

传统：受限网络带宽

EC2 P6：NVSwitch
+ 1.8TB HBM
本地高速 rollout

2-3x

弹性扩缩

传统：跨集群调度复杂

EC2 P6：EKS 统一编排
托管控制面，自定义CA灵活扩缩容

<10s

*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。

MAAS ON AMAZON

MaaS 平台 × Amazon： EC2 P6 训推一体的基础设施选型

计算、存储、网络三层选型，全部围绕 Amazon EC2 P6 ODCR 训推一体架构设计。

计算层 Compute

- Amazon EC2 P6-B300
  训推统一资源池
- ODCR 长期算力锁定
  确保容量可用
- Self-managed NodeGroup分区
  集群生命周期管理

存储层 Storage

- S3 Express One Zone
  Checkpoint <100ms
- FSx for Lustre
  训练数据高速共享
- EBS io2 Block Express
  权重热加载

网络与调度

- EFA 6.4 Tbps
  NCCL 集合通信
- PrivateLink
  MaaS API 内网调用
- EKS + Karpenter
  推理 Pod 扩缩
- **Topology-aware scheduling**确保同一训练 Job 的 Pod 落在同 Rack / 同 Switch 下，减少跨 Rack 延迟

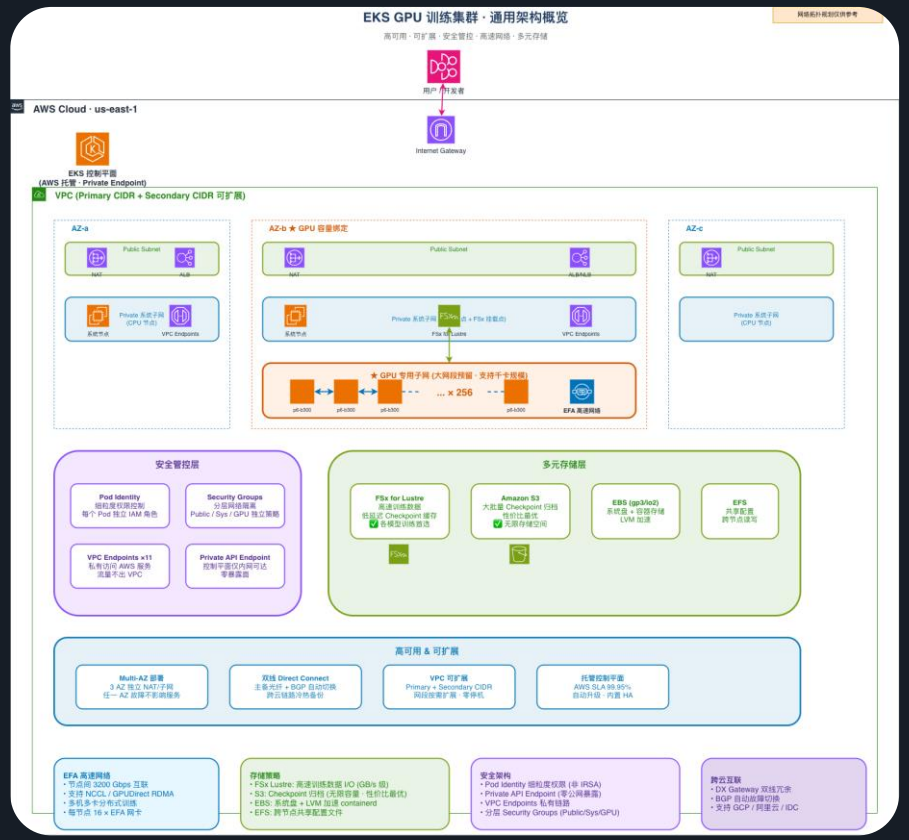
100%
算力保障 (ODCR)

>85%
GPU 利用率

99.9%
服务可用性

*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。

EC2 P6 训推一体的高可用架构设计



基础设施强韧 · 高可用设计

跨 3 AZ 部署，独立 NAT + 子网，EKS 控制平面 Amazon 托管（SLA 99.95%）；双线 Direct Connect 主备冗余，BGP 秒级故障切换，跨云链路无单点。

EFA 高速互联 · 千卡级分布式训练

每节点 16× EFA 网卡，节点间 6400 Gbps 全互联；原生支持 NCCL + GPUDirect RDMA，200+ 台 Amazon EC2 P6 Instances 组成大规模训练集群，通信延迟极低。

多层存储组合 · 兼顾性能与性价比

FSx for Lustre 提供 GB/s 级训练数据吞吐与 Checkpoint 高速缓存；S3 承接海量 Checkpoint 归档，无限容量、成本最优；EBS/EFS 覆盖系统盘与共享配置场景。

纵深安全管控 · 零信任架构

Private API Endpoint 零公网暴露；Pod Identity 细粒度最小权限；11 个 VPC Endpoints 保证流量不出 VPC；三层 Security Groups 网络隔离。

VPC 弹性扩展

Primary + Secondary CIDR 双段设计，按需零停机扩容，应对未来业务扩展。

MAAS GOVERNANCE

MaaS 治理：Token 经济与服务分级

全部基于 Amazon EC2 P6 ODCR，按优先级分区调度，训练与服务资源最优分配。

Tier 0 — 线上 Serving (P99 <200ms)
ODCR 固定分区 · 高优先级独占 GPU

Tier 1 — 训练 Judge (P99 <2s)
ODCR 共享分区 · 训练间隙复用推理资源

Tier 2 — Eval & Data Gen (尽力)
ODCR 弹性分区 · 训练空闲时段批量执行

Token 计量	按调用方分账 · CloudWatch 看板
限流与背压	Judge 突增自动降级 · 保护 Serving
模型灰度	Eval→Judge→线上 逐级验证
成本归因	Cost Explorer 标签化 · 精细分摊

3x
Token 吞吐/$ 提升

<1%
训练中断率

分钟级
模型热切换

100%
ODCR 无抢占

*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。

END-TO-END ON AMAZON

全链路：Relax + MaaS + EC2 P6 在 Amazon 上的端到端

EC2 P6 ODCR 训推一体 + Amazon 全栈，训练与服务运行在同一资源池。

数据生成 (MaaS)

EC2 P6 ODCR 复用
Teacher / Self-play
百万级样本/天

训练 (Relax)

EC2 P6–B300
训推一体
利用率 >85%

评测 (MaaS)

EC2 P6 弹性分区
Checkpoint 评测
回归 <30min

服务 (MaaS)

EC2 P6 ODCR 固定分区
P99 <200ms
99.9% 可用

Amazon 统一基础设施：EFA · S3 Express · FSx · EKS · CloudWatch

ODCR 统一资源池 — 四阶段共享同一 EC2 P6 算力，按优先级动态分区

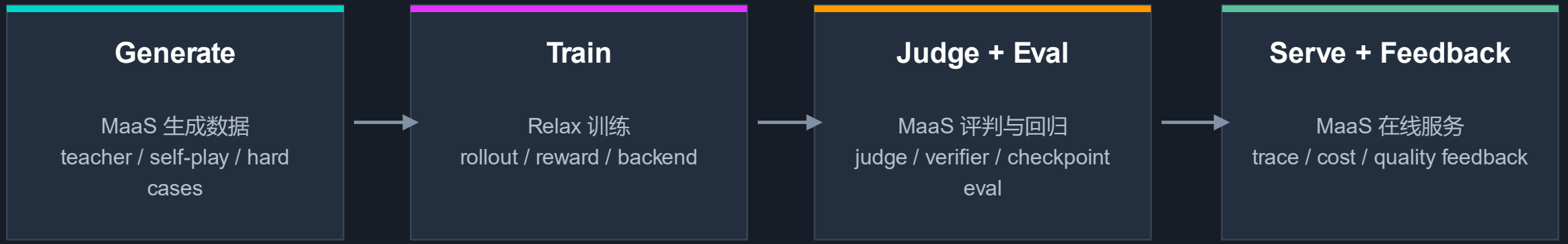
阶段	Amazon 服务	关键指标
Data Gen	EC2 P6 ODCR 复用	百万级样本/天
Training	EC2 P6 + FSx	利用率 >85%
Eval	EC2 P6 弹性分区 + CloudWatch	回归 <30min
Serving	EC2 P6 ODCR 固定分区 + EKS	P99 <200ms

*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。

CHAPTER 04

从训练闭环到 AI Native 基础设施演进

Relax 与 MaaS 的关系不是单向部署链路，而是训练生产、模型服务和反馈治理共同组成的闭环。



演进路径：MaaS 提供模型能力平面，Relax 提供训练闭环，两者共同推动 AI Native 基础设施从“任务系统”走向“能力系统”。

亚马逊云科技
中国峰会

Thank you

Q&A

Relax: github.com/redai-infra/Relax
主题：小红书大模型基础设施实践
Relax 与 MaaS 驱动的模型能力闭环

欢迎交流
欢迎共建

亚马逊云科技
中国峰会

Thank you

Q&A

Relax: github.com/redai-infra/Relax
主题：小红书大模型基础设施实践
Relax 与 MaaS 驱动的模型能力闭环

欢迎交流
欢迎共建