
AWS
re:Invent
DECEMBER 1 – 5, 2025 | LAS VEGAS, NV

AIM383

Build more effective agents through model customization

Davide Gallitelli

Senior WW Specialist SA GenAI/ML

(he/him)

Sam Palani

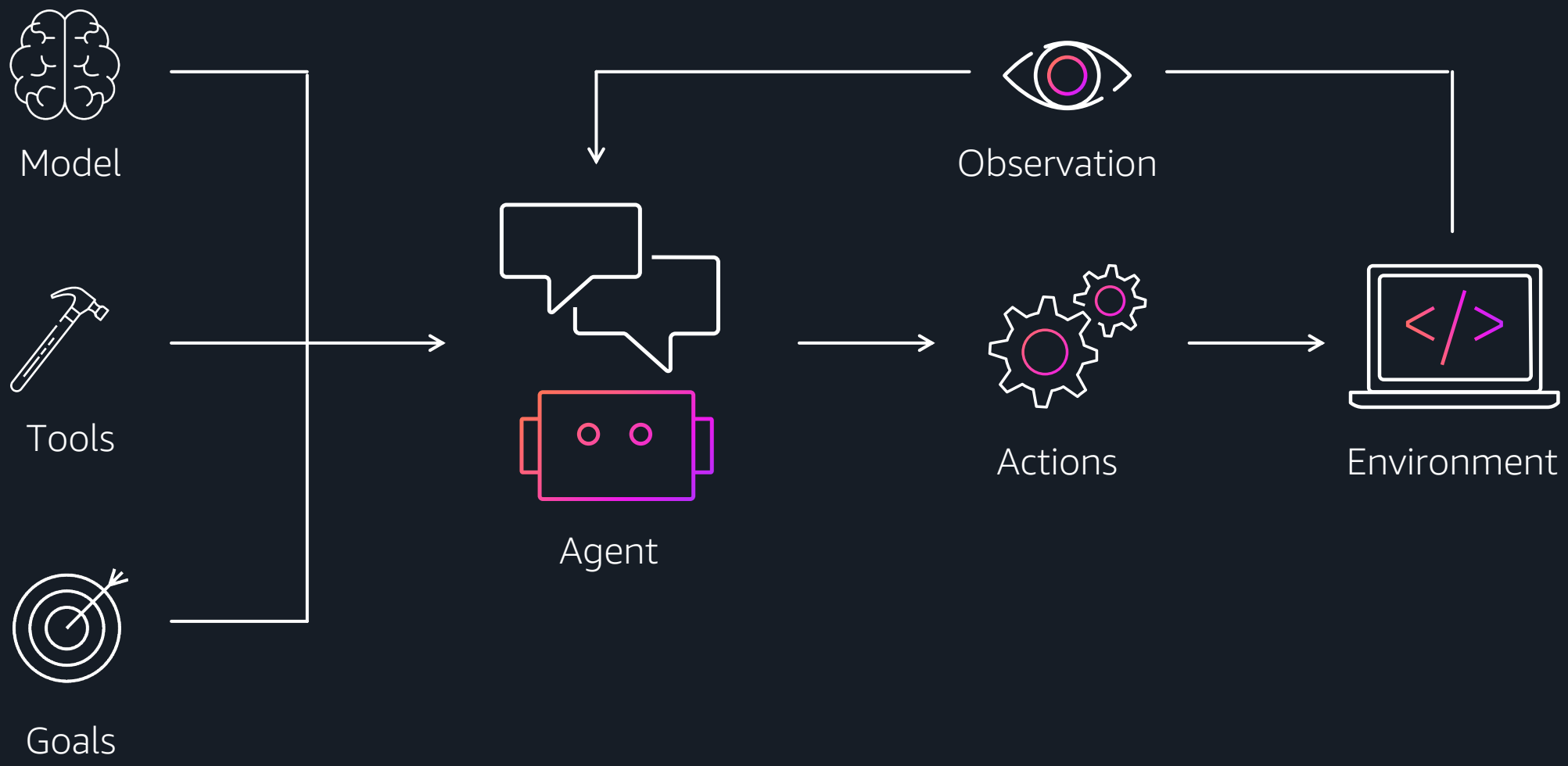
Head Foundation Models Data Science & SA GTM

(he/him)

Agenda

- AI Agents and Customization
- Amazon Nova Foundation Models
- Nova Customization on Amazon Bedrock
- Customizing Nova with SageMaker AI
- Demo
- Customer Story (?)

What is agentic AI?



Model reasoning capabilities

Scalability and cost-effectiveness

Secure data infrastructure

Development tools

Enterprises are doubling down on agents

33%

of enterprise software apps will include Agentic AI by 2028, up from less than 1% in 2024.

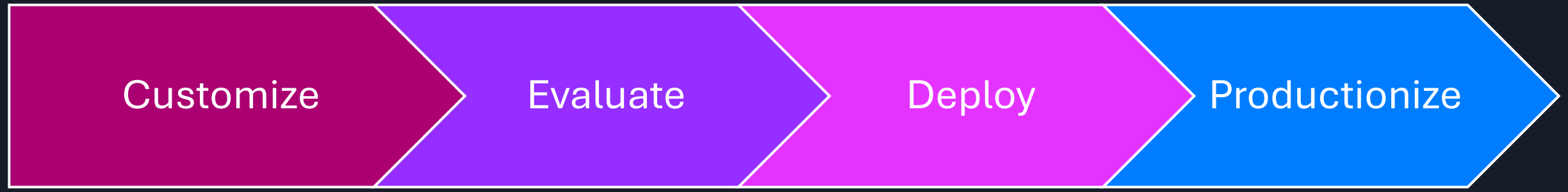
Gartner, “Top strategic Technology Trends for 2025,” October 2024

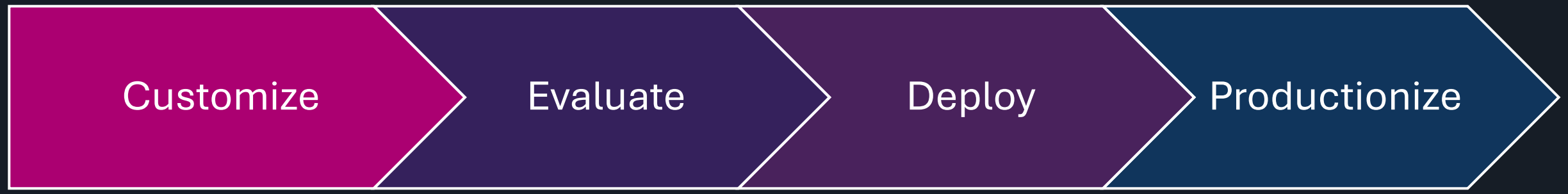
15%

of day-to-day work decisions will be made autonomously through Agentic AI by 2028.

Gartner, “Top Strategic Technology Trends: Agentic AI—the Evolution of Experience” February 2025

The path to your perfect AI Agent





Why Model Customization for Agents

Brittle by Default

General models impress in demos but fail on your domain and edge cases—customization makes them dependable.

Domain Mismatch

Your acronyms, taxonomies, and formats aren’t “internet average”—customization aligns to your semantics

Tool-Schema Friction

Teach schema-true JSON so tool calls pass validation the first time, slashing retries and glue code.

Long-Horizon Drift

Customize micro-policies to shorten plans and prevent small errors from compounding across steps.

Latency & Cost Control

Distill complex behaviors into lighter variants to cut p50/p95 latency and dollars per task.

Customization: Progression of Options

Optimizing the Output

- Prompt engineering
- Retrieval augmented generation (RAG)
- Context Engineering

Modifying the Model

Post-Training

- Fine-tuning (SFT)
- Model distillation
- Direct preference optimization (DPO)
- Reinforcement learning (RL)

Pre-training

- Continued pre-training (CPT)

Level of customization

Level of effort

Model Customization Techniques: a deep dive

Fine-Tuning (PEFT and Full)	Adapt models for a specific task with a (small) number of labelled dataset
DPO (PEFT and Full)	A simple way to align an LLM with what people prefer without training a separate reward model or running RL
Proximal Policy Optimization (PPO)	Improve your models using real-world user interactions and associated feedback to mimic human behavior
Distillation	Use teacher model to distill smaller student models using proprietary data for cost-efficiency and lower latency
Continuous Pre-Training	Continue training the base model checkpoints on a large corpus of unlabeled data – Domain adaptation

Amazon Nova Foundation Models

State-of-the-art foundation models that deliver frontier intelligence and industry-leading price performance

Amazon Nova **Multimodal Embeddings**

UNDERSTANDING MODELS

Amazon Nova **Micro**

Amazon Nova **Lite**

Amazon Nova **Pro**

Amazon Nova **Premier**

CREATIVE CONTENT GENERATION MODELS

Amazon Nova **Canvas**

Amazon Nova **Reel**

(Reel 1.1)

SPEECH-TO-SPEECH MODEL

BROWSER USE

Amazon Nova **Sonic**

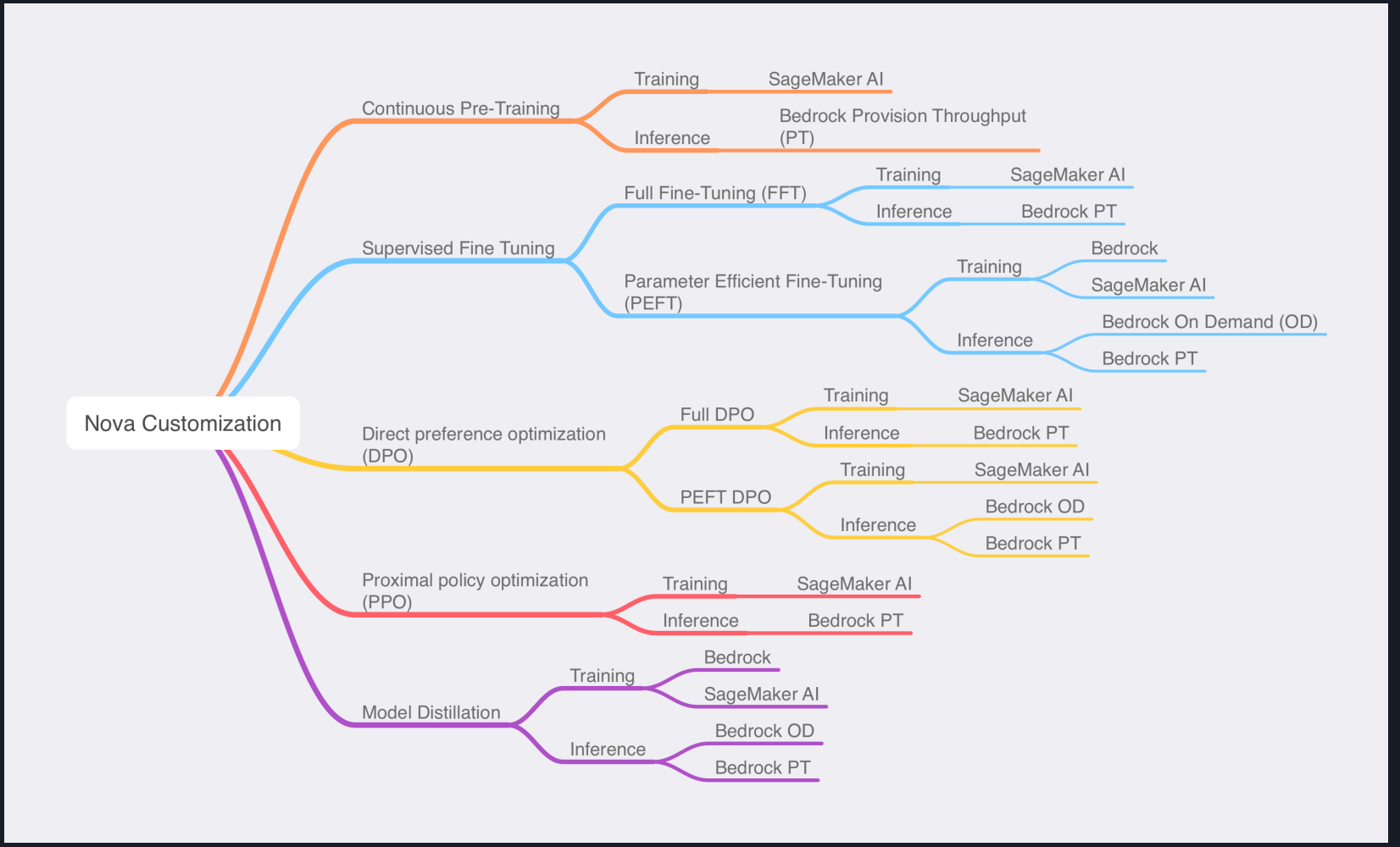
Amazon Nova **Act**

(Preview)

NOVA WEBSITE

aws

Customization Options for Amazon Nova



Nova Customization on Amazon Bedrock

Fine-Tuning PEFT

Supervised PEFT based fine-tuning for Agentic tasks such as tool calling and persona switching

Model Distillation

Distill the behaviors of the high performing model into a more efficient student model - productionizing a high performing prototype

Deployment

Available both On-Demand Inference and Provisioned Throughput.

SageMaker AI offers two training options

Purpose-built infrastructure for FM training

Fully managed training jobs

Fully managed resilient infrastructure for large-scale and cost-effective training

Focus on model building rather than IT

Provide access to flexible on-demand GPU cluster with a pay as you go option

Amazon SageMaker HyperPod

Resilient and **self orchestration** infrastructure for maximum resource control

Customize and manage cluster orchestration (Slurm or EKS)

Schedule workloads to maximize cluster utilization across teams

Optimized recipes help get started

Business value

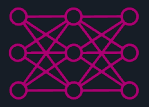
Reduce FM training setup time and get started in minutes vs. days or weeks

Pre-tested training stack configurations for fine-tuning and pre-training publicly available FMs

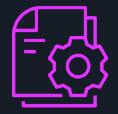
Easily switch between GPU and Trainium instances

Open-source and customizable, with a growing list of 90+ recipes

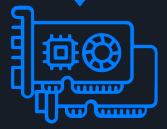
How it works



Specify training and validation data directories

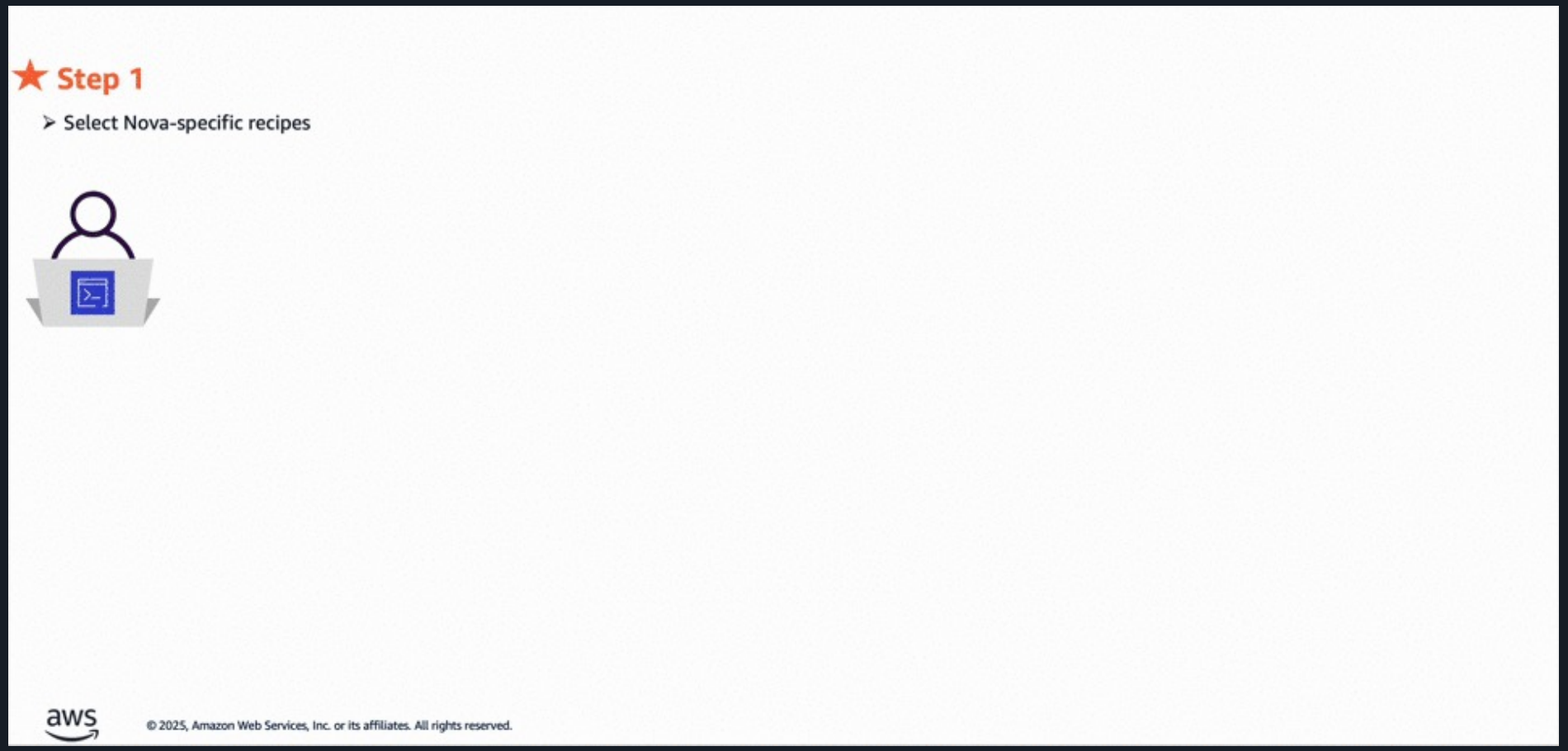


Run the recipe on SageMaker HyperPod or fully managed Training Jobs

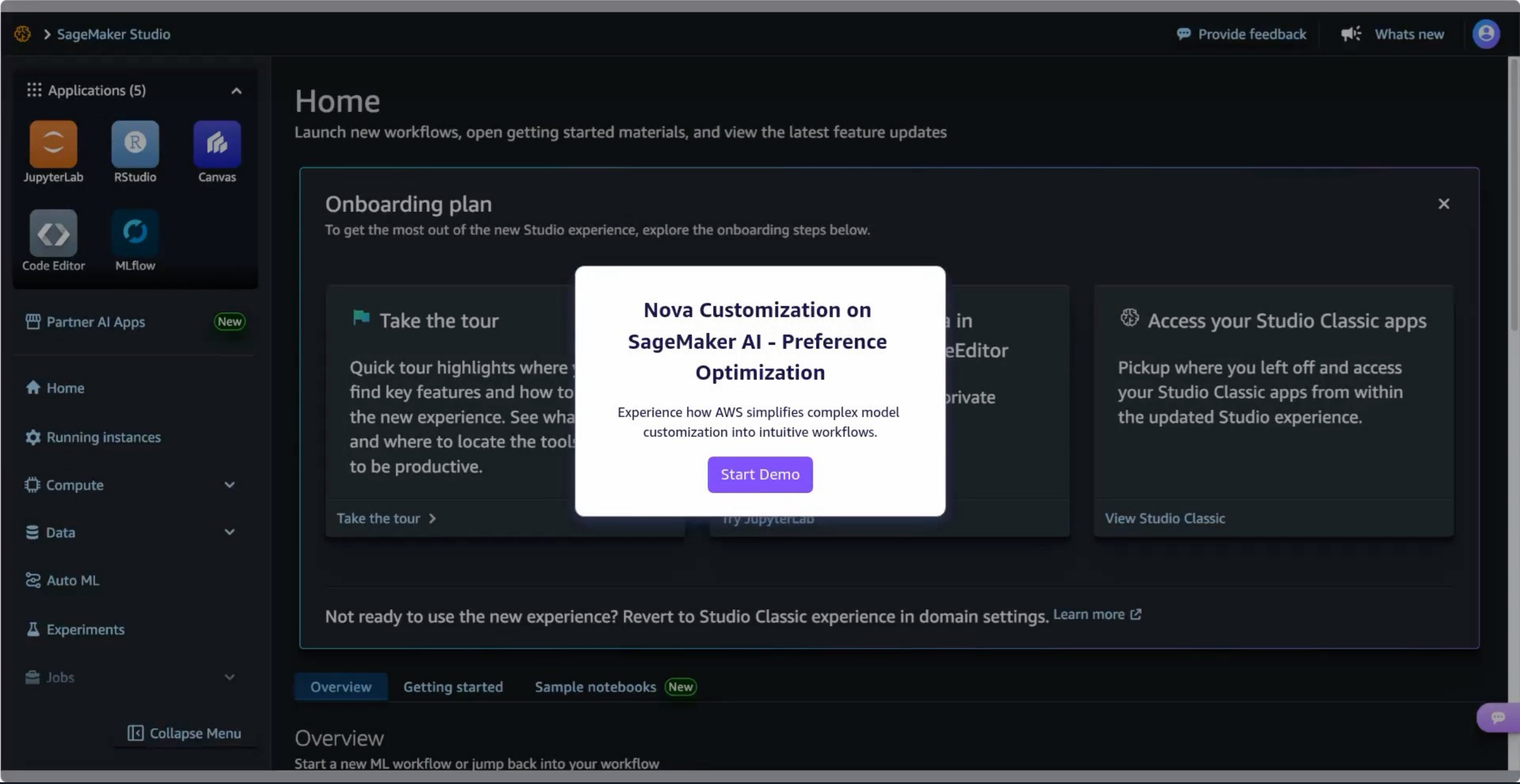


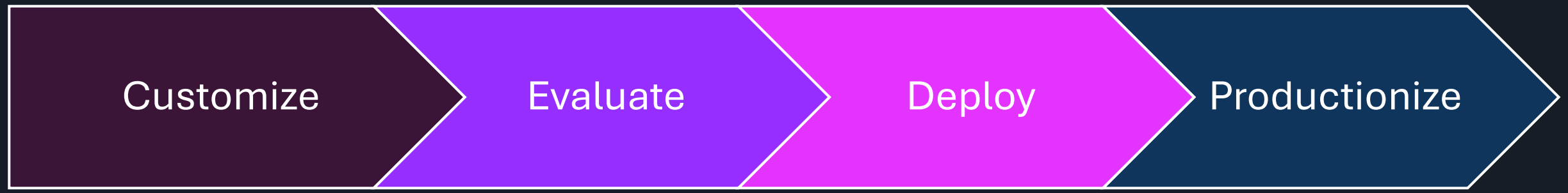
Select a recipe for popular publicly-available FMs (DeepSeek, Llama, etc.) or Amazon Nova

SageMaker Recipes simplify customization



Learn more:





LLM Evaluation

Assess their performance across different tasks and applications

1

General Metrics

Perplexity, accuracy, cross-entropy loss

2

Text Quality Metrics

BLEU, ROUGE, METEOR

3

Task-Specific Metrics

Question-Answering Accuracy, Named Entity Recognition Metrics, Sentiment Analysis Metrics

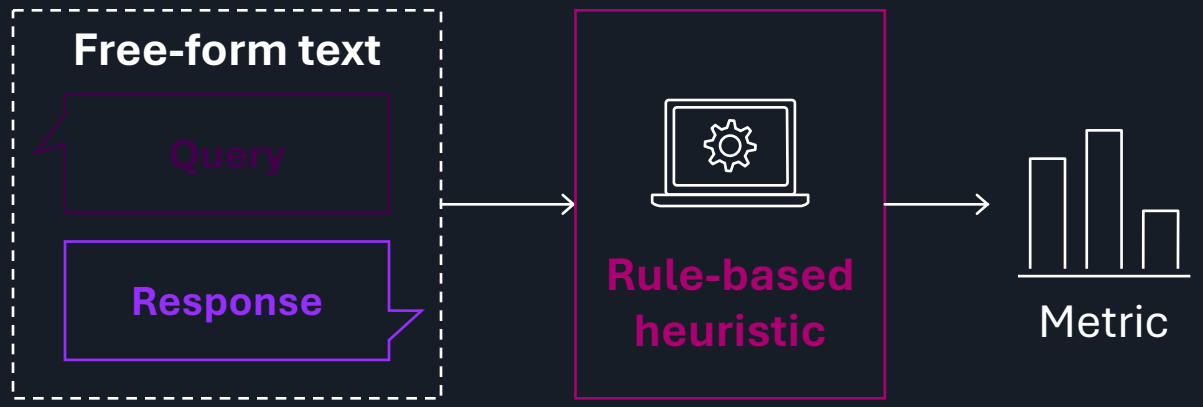
4

Ethical and Robustness Metrics

Hallucination Detection, Bias and Toxicity Metrics, Diversity

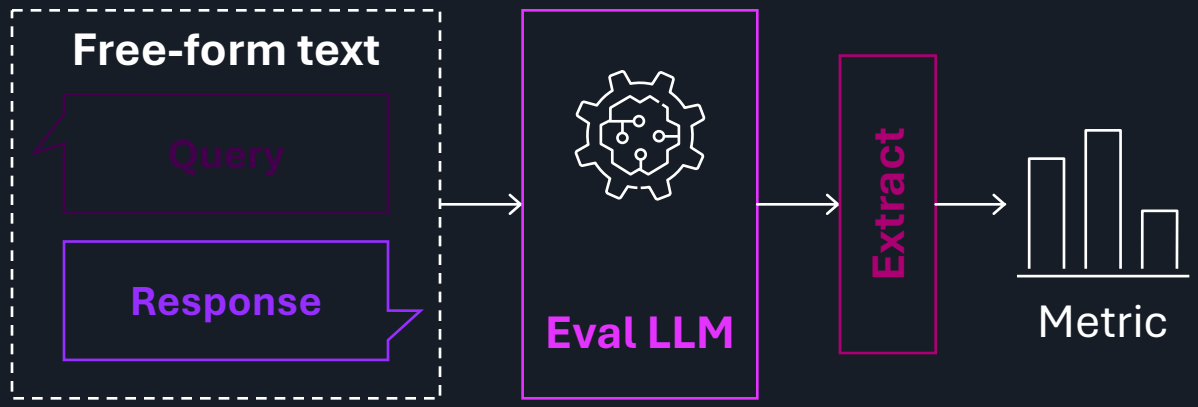
Evaluation with LLM-as-a-judge

Rule-based heuristics



Fast, scalable, cheap to run
Leverage standard metrics (F1, ROUGE…) or helper models (sentiment, toxicity…)
Will the metrics align well with human preferences?

LLM-based critique



Flexible, customizable checks
Checking an answer usually easier than creating it
Is it biased by the evaluator?
Affordable to run?

LLM-as-a-judge metrics

01 Correctness

02 Completeness

03 Faithfulness

04 Helpfulness

05 Coherence

06 Relevance

07 Following instructions

08 Professional style and tone

09 Readability

10 Harmfulness

11 Stereotyping

12 Answer refusal

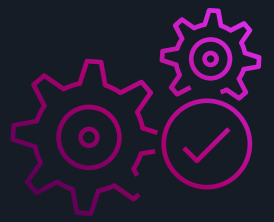
Choose your deployment option

Bedrock Inference

Custom Model Import (CMI)

Pay-per-use format

Available both On-Demand (PEFT, DPO, Distillation) and Provisioned Throughput (all techniques)



Amazon SageMaker AI

Real-Time Inference Endpoint or Inference on Hyperpod

Choose your deployment infrastructure
Choose your serving stack
Always available, pay-per-minute

Currently not possible for Amazon Nova models

Applications (5)
JupyterLab
RStudio
Canvas
Code Editor
MLflow
Partner AI Apps New
Home
Running instances
Compute
Data
Auto ML
Experiments
Jobs
Pipelines
Models
Collapse Menu

Nova models support Training and Evaluation with recipes but are not currently compatible with the SageMaker Python SDK. Learn about Nova

Nova Micro

by Amazon

Train | Deploy | Optimize | Evaluate

About

This is a generative AI model from Amazon. Nova Micro is a low-latency, text-only model that delivers responses at a very low cost. It is multilingual model that supports text inputs and outputs. Nova Micro can be trained as a student of Nova Pro. To learn more about the model, please read the Amazon Nova documentation .

Tags: Nova, Text Generation

Provider: Nova

Task: Text Generation

Prerequisites

Access to Nova recipes

Choose your approach to customize:

1. **SageMaker training jobs** - Train Nova model on the SageMaker training job platform. You don't need to create a cluster.

2. **HyperPod clusters** - Train Nova model on HyperPod. You need to create a HyperPod EKS cluster with Restricted Instance Group (RIG).

Model customization

Base model recipe

Get the appropriate base recipe from the recipes repository based on your customization goals.

Starter notebook

Use a sample Jupyter notebook to start a training job.

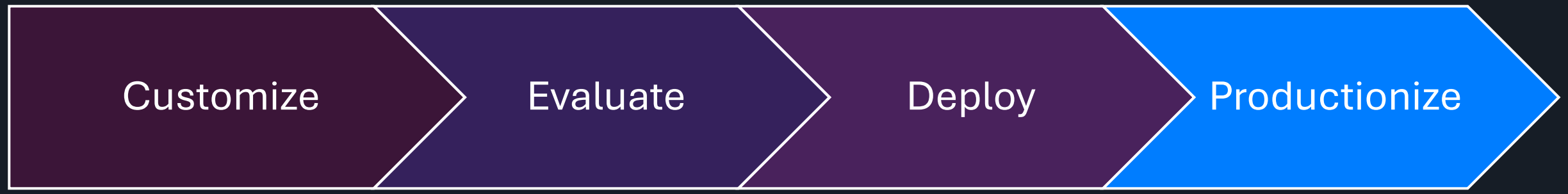
Bedrock inference

Deploy trained model to Bedrock for production usage.

Resources

GitHub repository

Customizing Amazon Nova models with SageMaker AI



EMBRACING OPEN SOURCE

Frameworks for building agents

STRANDS AGENTS SDK

LANGGRAPH, LANGCHAIN

OpenAI

OPEN AI AGENTS SDK

CREW.AI

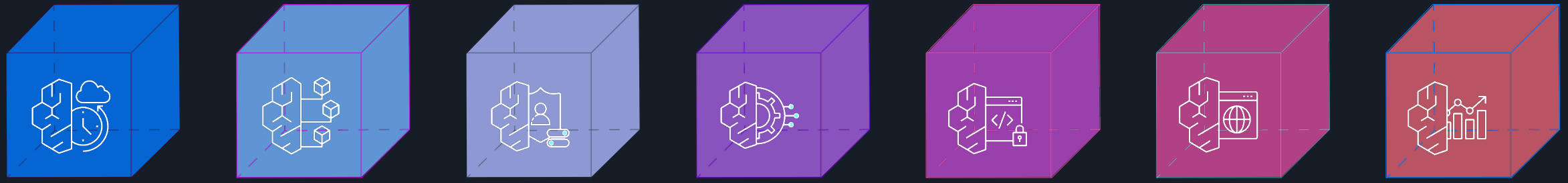
GOOGLE ADK

LLAMAINDEX

+ many more

Amazon Bedrock AgentCore

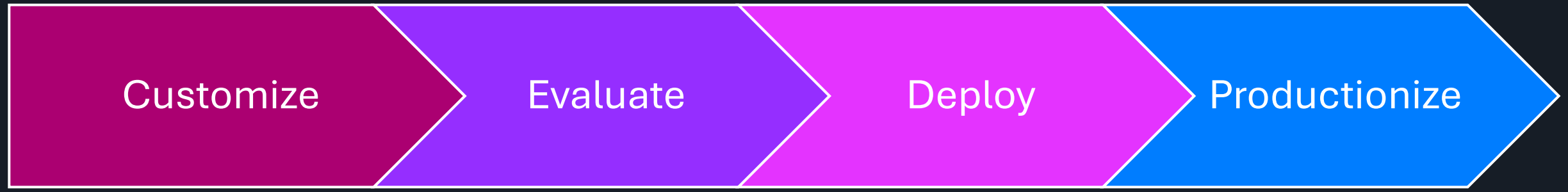
Everything you need for getting agents into production



Runtime | **Memory** | **Identity** | **Gateway** | **Code Interpreter** | **Browser Tool** | **Observability**

Wrapping up

The path to your perfect AI Agent



When to Customize

Performance

- Higher accuracy, consistency
- Domain-Specific Needs
- Brand Alignment
- Control and Compliance

Business

- Cost Efficiency
- Competitive Advantage
- Operational Scale
- Integration Needs

Technical

- Quality training data available
- Team with technical expertise
- Identified specific performance/business targets

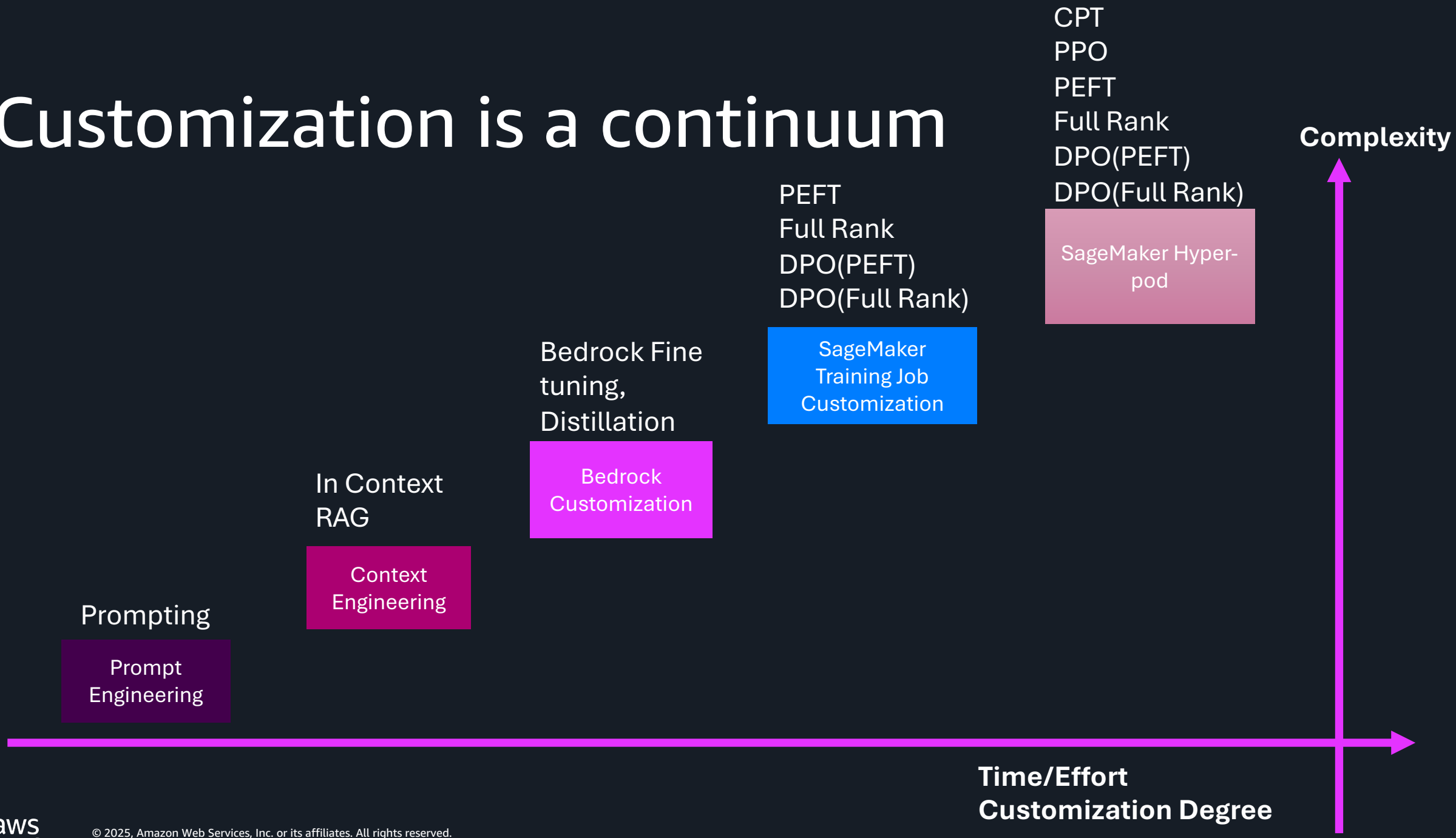
Supported Fine-tuning Techniques

Methods	Bedrock	SageMaker AI
Parameter-efficient fine-tuning	☑	☑
Full fine-tuning		☑
Parameter-efficient Direct Preference Optimization (DPO)		☑
Full model DPO		☑
Proximal Policy Optimization (PPO)		☑
Continuous Pre-training		☑
Distillation	☑	☑

Only text-only modalities supported on Bedrock

- *Available in N. Virginia Region*
- *SageMaker based customization is available for Micro, Lite, Pro models*

Customization is a continuum



Continue your learning journey on AWS Skill Builder

Dive deeper with 1,000+ free, expert-led training on AWS's official online learning center

Scan to access more trainings on:

Amazon Nova

Thank you!

Please complete the session survey in the mobile app