
AWS
re:Invent
DECEMBER 2 – 6, 2024 | LAS VEGAS, NV

AIM350-NEW

Supercharge your generative AI applications with Amazon Nova models

Sruthi Karuturi
She/Her
Sr. Manager, Software Development
Amazon, Artificial General Intelligence

Ryan Hoium
He/His
Sr. Manager, Solutions Architecture
Amazon, Artificial General Intelligence

Amazon Nova roadmap

Model Type	Model	Input/Output	Context Length	Customization
Text-to-text	Amazon Nova Micro	Input – Text Output - Text	128K	Fine-Tuning, Distillation (w/ Pro as Teacher)
Multimodal Understanding	Amazon Nova Lite	Input – Text, Image, Video, Output - Text	300K	Fine-Tuning, Distillation (w/ Pro as Teacher)
	Amazon Nova Pro	Input – Text, Image, Video, Output - Text	300K	Fine-Tuning
	Amazon Nova Premier	Input – Text, Image, Video, Output - Text	Coming soon!	Coming soon!
Multimodal Generation	Amazon Nova Reel	Input – Text, Image, Video Output – Video	NA	Coming soon!
	Amazon Nova Canvas	Input – Text, Image Output - Image	NA	Coming soon!

Amazon Nova understanding models overview

PRO, LITE, AND MICRO

Best-in-class models for text generation with image and video understanding, offering the best price-performance and accuracy on multimodal RAG and agents. Easily customizable with proprietary data in a secure Amazon Bedrock environment.

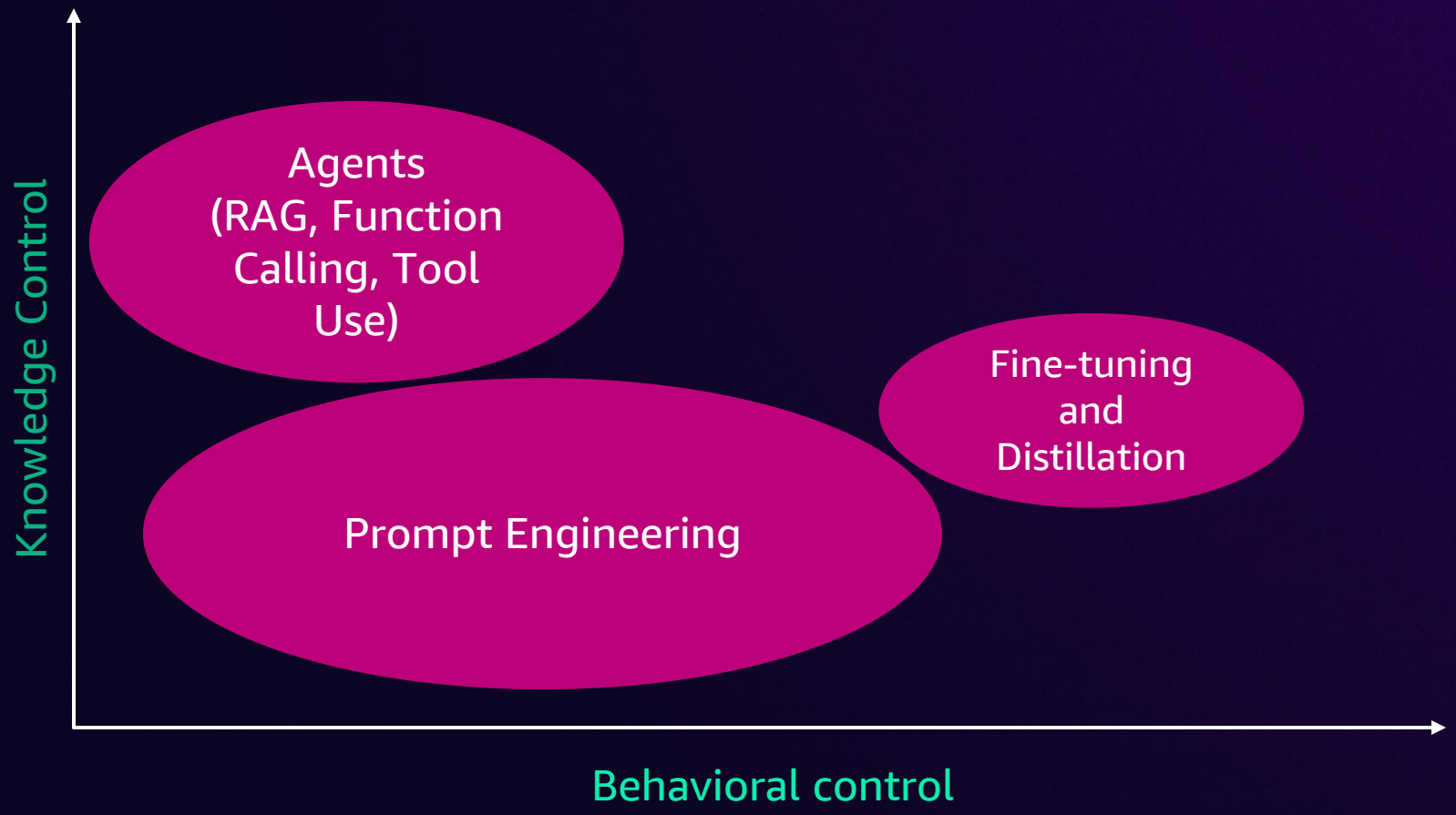
Key Attributes

- ✓ Input modalities: Text, Image, Video
- ✓ Output modalities: Text
- ✓ Max tokens: 300K tokens *(equivalent to 225K words, or 100 documents, or 30 mins of video)*

- ✓ **Image, documents, and video understanding and reasoning** to perform tasks such as captioning, Q&A, summarization in addition to wide range of text-related tasks, such as code generation, and conversational chat
- ✓ **Create text and multimodal agents** that can perform and automate multistep tasks for your customers on images, documents, or videos
- ✓ **Customize** on your proprietary data through custom fine tuning text, image, and video modalities
- ✓ **Distill** the intelligence of Pro, at the cost and speed of Lite or Micro using Model Distillation on Bedrock

Amazon Nova improves your applications

Behavior: How the model responds
Knowledge: What the model knows or is using for context



Build AI agents with Amazon Nova

Agents orchestrate interactions between foundation models (FMs), data sources, software applications, and user conversations

Common terms: Tool calling, function calling, plug ins

Tool types

APIs
Python interpreter
SQL clients
RAG systems

Functionality

Single tool call
Parallel tool calls
Sequential tool calls
Planning

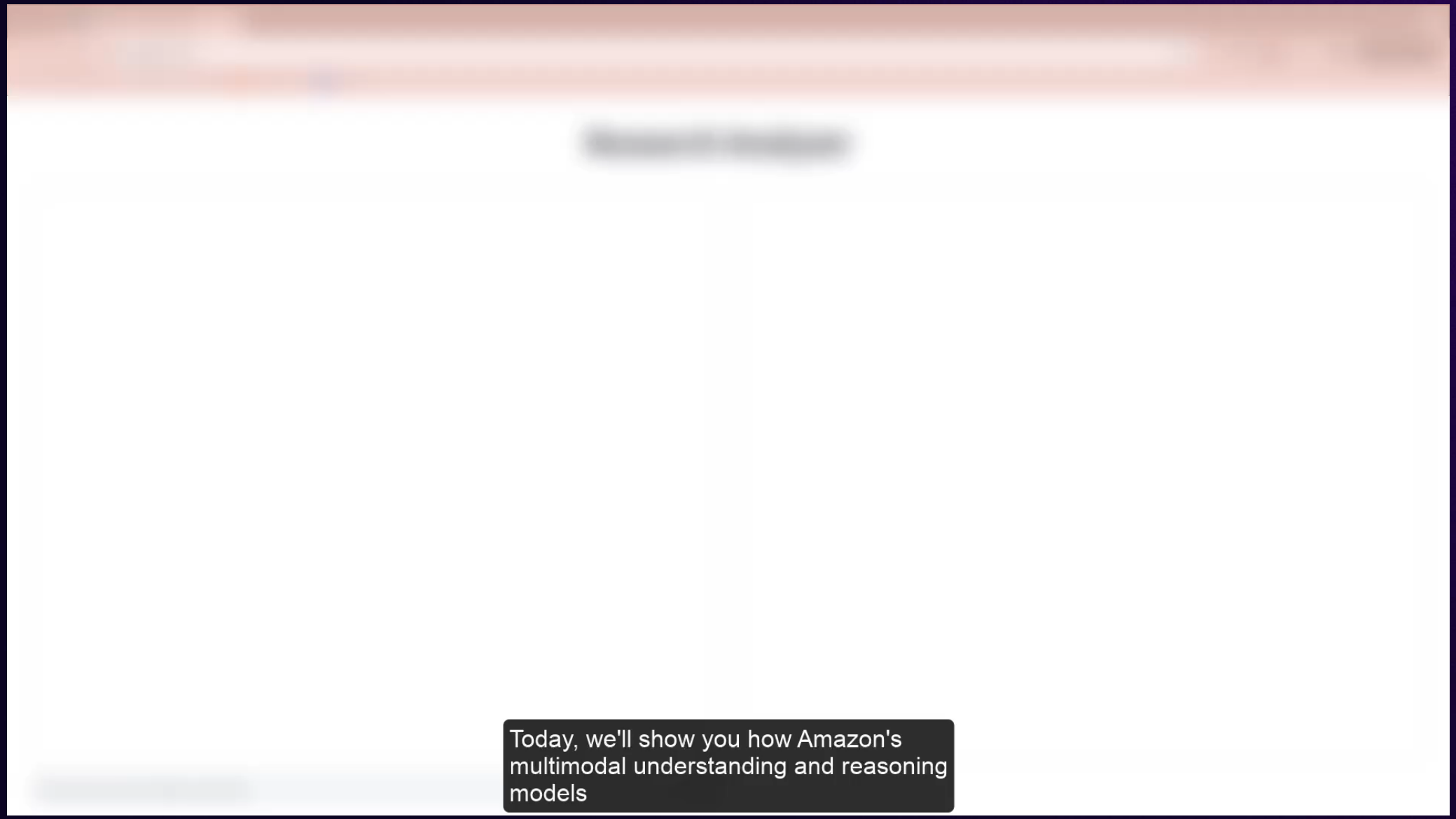
Frameworks

Amazon Bedrock Agents
Langchain
Custom framework

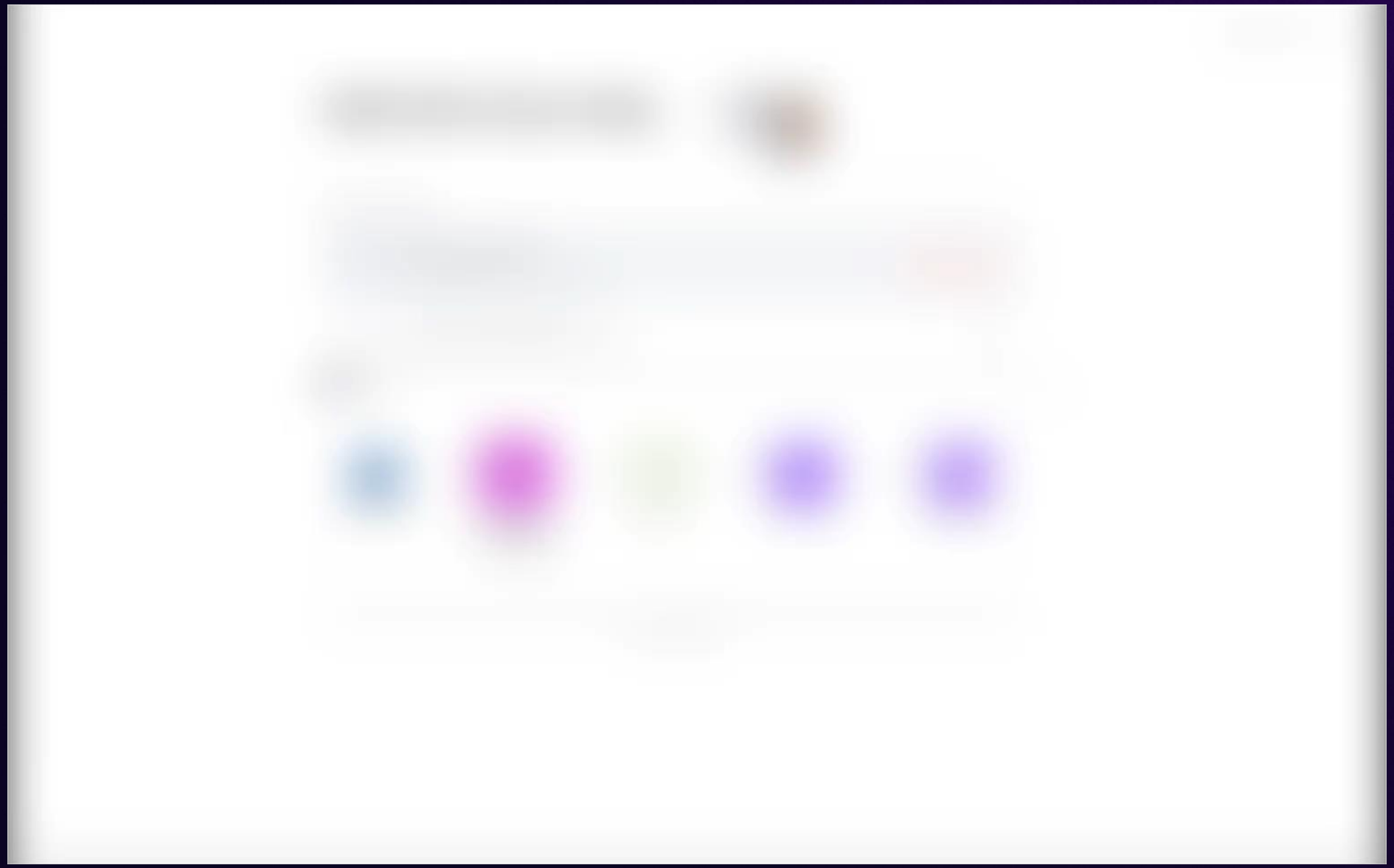
Applications

Chatbots
Process Automation
Multi-agent networks

Amazon Nova Pro + Lite demo | Research analyzer



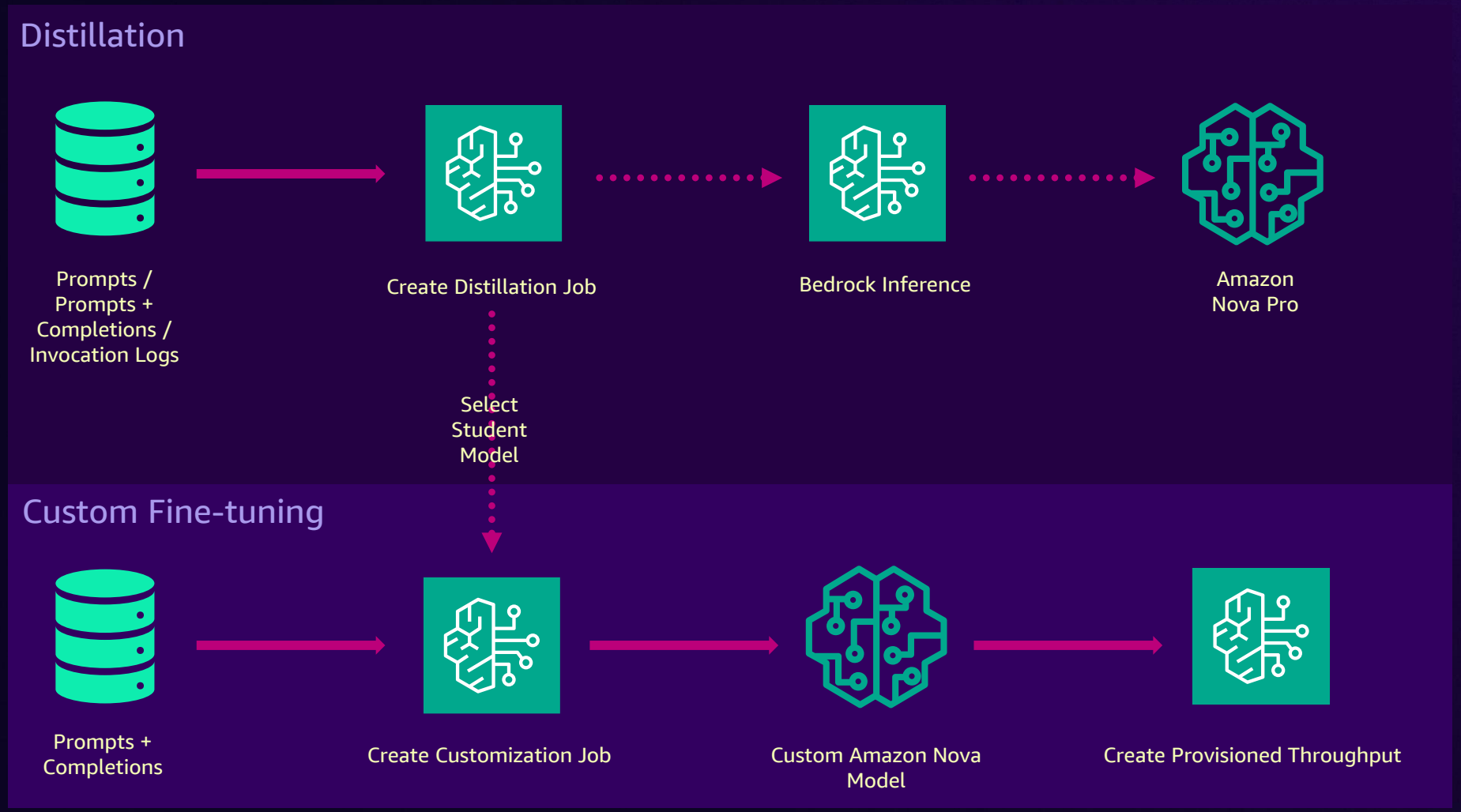
Amazon Nova Pro demo I AWS Support agentic applications



In this demo, we show how Amazon Nova
models can power computer agents

In this demo, we show how Nova can power computer agents

Amazon Nova customization high-level system design



Amazon Nova customization capabilities

Model Family	Input Modality			Output Modality	Distillation From Pro (Text Input Only)	Distillation From Premier
	Text	Image / Document	Video			
Amazon Nova Micro	✓	✗	✗	Text	✓	Coming soon!
Amazon Nova Lite	✓	✓	✓	Text	✓	Coming soon!
Amazon Nova Pro	✓	✓	✓	Text	✗	Coming soon!
Amazon Nova Canvas	Coming soon!	✗	✗	Image (Coming soon!)	✗	✗
Amazon Nova Reel	Coming soon!	✗	✗	Video (Coming soon!)	✗	✗

Amazon Nova Customization

PRO, LITE, AND MICRO

Fine-tuning (Micro/Lite/Pro)

- You want a consistent zero-shot response from a customized model that can't be achieved with prompting
- You have the data to train the model and improve its responses on your tasks
- Context Length: 32K
- Image Understanding – Improve context specific responses on image inputs
- Video Understanding – Teach new context (e.g. improved captioning, context specific commentary on sports content)

For text input modality

+ Distillation (from Pro to Micro/Lite)

- You get better responses from Pro but need the speed and cost efficiency of Lite/Micro
- You have small set of labeled data (e.g., ~100)
- You have prompts but not labeled data
- You have too much label noise (e.g., different ground truth styles from different annotators)
- You want generalization power of Pro on top of simple finetuning

Amazon Nova customization use cases

Diverse set of fine-tuning tasks (public and internal)

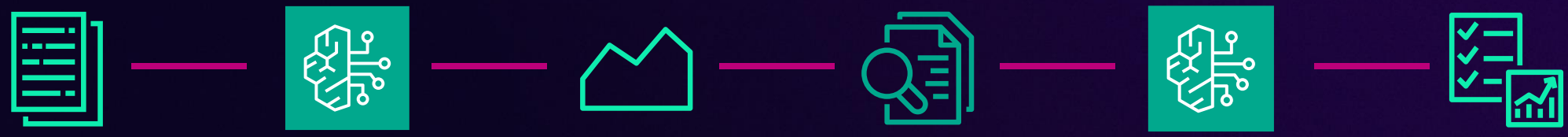
Goal:

- *Maximize performance on target tasks*
- *Minimize catastrophic forgetting (e.g. RAI, foundational capabilities)*

Example Task	Modality	Metric	Data Size	Relative Improvement over Base Model
Structured Response (e.g., JSON)	Text	Exact Match	100 – 20K samples	~30%
Classification (binary and multi-label)	Text	F1		~50%
Question Answering (e.g., medmcqa)	Text	Exact Match		>50%
Query Understanding/Routing	Text	F1		>30%
Table Q&A (e.g., tat-qa)	Text	Accuracy		~50%
Sentiment Analysis (e.g., fpb)	Text	F1		~50%
Tool Calling (API Invocation)	Text	Accuracy		~20%
Tool Calling (API Argument Accuracy)	Text	Accuracy		~10%
Chart/Graph Q&A Anomaly Detection	Image	Accuracy		~15%
Visual Table Understanding	Image	JSON value Accuracy		>50%
Image Understanding (e.g., VQA-Radiology)	Image	Accuracy		>50%
Video Caption Generation	Video	Rouge		~25%

Customize Amazon Nova models on Amazon Bedrock

Customers typically go through the following process (possibly in iterations) before finalizing and deploying custom models into their enterprise production



Format Data

{ "messages": [{ "role": "user", "content": [{ "text": "What is AWS?" }] }, { "role": "assistant", "content": [{ "text": "It is Amazon Web Services." }] }] }

Trigger Customization

Via AWS console or API with specified parameters

Monitor

Track or stop jobs in AWS console

Analyze Training Results

Analyze training and validation loss in S3

Inference with Provision Throughput

Purchase throughput, choose custom model, invoke inference

Evaluate Custom Model Performance

Measure performance using proprietary metrics and datasets

Questions?

Thank you!

Please complete the session survey in the mobile app

Sruthi Karuturi
Email: sruthisk@amazon.com
LinkedIn: @sruthi-karuturi

Ryan Hoium
Email: rhoium@amazon.com
LinkedIn: @ryan-hoium