AWS re:Invent

DECEMBER 2 – 6, 2024 | LAS VEGAS, NV

AIM398-NEW

Practical generative AI using Amazon Nova

Jay Lee
(he/him/his)
Sr Manager Product
Artificial General Intelligence

Shubham Katiyar
(he/him/his)
Director Engineering
Artificial General Intelligence

Patrick Nguyen
(he/him/his)
Director Engineering
AWS Support

Gerard Medioni
(he/him/his)
VP, Distinguished Scientist
Prime Video & Amazon MGM Studios

Jamie St. Onge
(she/her/hers)
Sr Manager Product
Amazon Q

Agenda

01 Application of gen AI

02 Amazon Nova Models

03 Amazon Nova in action

- ✓ AWS Support
- ✓ Prime Video
- ✓ Amazon Q Developer
- ✓ Amazon Advertising

04 Next steps and resources

Applications of gen AI

Intelligent Document Processing

Workflow Streamlining, Automation & Assistants

Assistants for Coding or Customer Service

Incident & Case Management

Creative Content Generation & Creative Asset Reuse

Hyper-Personalized Customer Experience

Gen AI Applications Across Amazon

Amazon Ads Video Generator

Prime Video Season Recap

AWS Agentic Support

Amazon Q Developer

Amazon Legal Contract Analysis

Amazon Finance Accounting Reconciliation

Introducing Amazon Nova

Amazon Nova

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

UNDERSTANDING MODELS

Amazon Nova **Micro**

Amazon Nova **Lite**

Amazon Nova **Pro**

Amazon Nova **Premier**
COMING SOON

CREATIVE CONTENT GENERATION MODELS

Amazon Nova **Canvas**

Amazon Nova **Reel**

Amazon Nova Foundation Models

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

Frontier Intelligence

Speed

Price performance

Agentic Workflows & RAG

Customization

Responsible

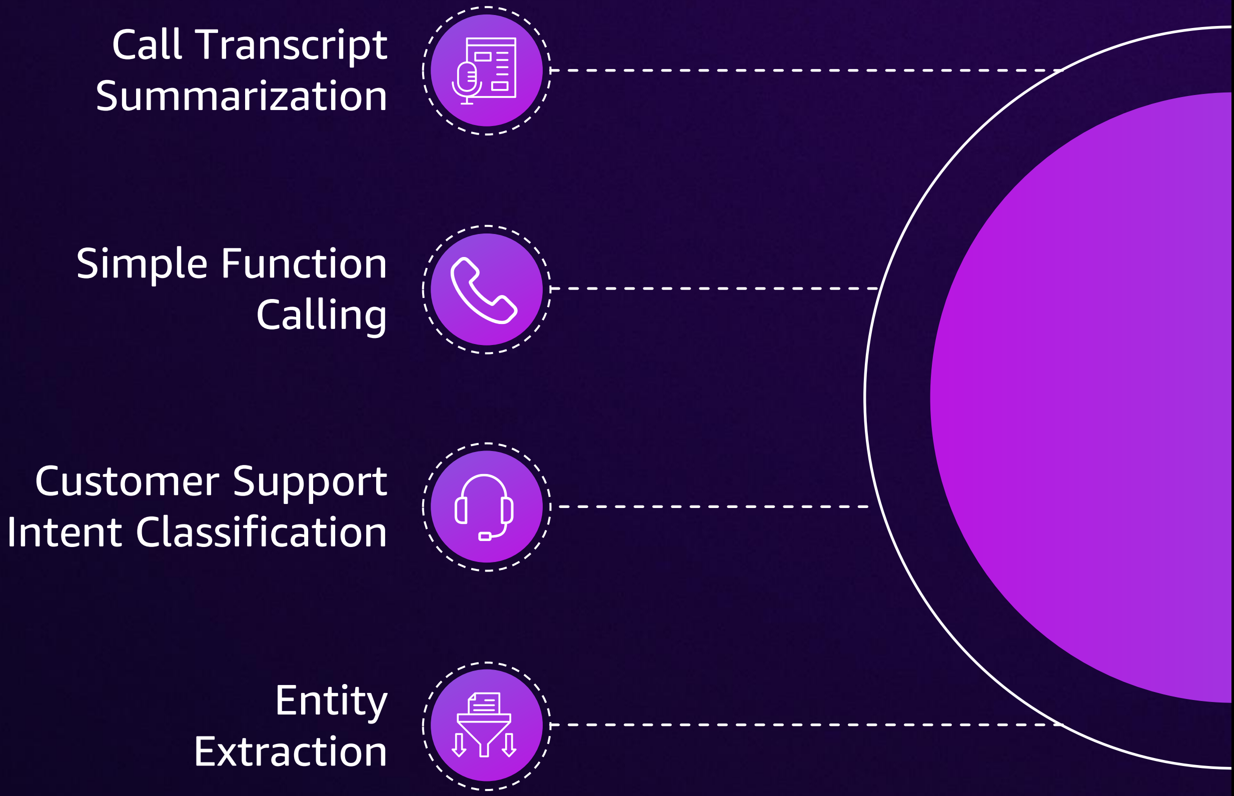
Amazon Nova Micro

Text-only model that delivers the lowest latency responses at very low cost

BEST COMBINATION OF SPEED, ACCURACY, AND COST

Key attributes

- ✓ Input modalities: Text
- ✓ Context Length: 128K tokens
- ✓ Languages: up to 200
- ✓ Customization: Fine-Tuning & distillation (student)
- ✓ Latency: 210 tokens per second
- ✓ Price/M Tokens: $0.035 input; $0.14 output



Call Transcript Summarization

Simple Function Calling

Customer Support Intent Classification

Entity Extraction

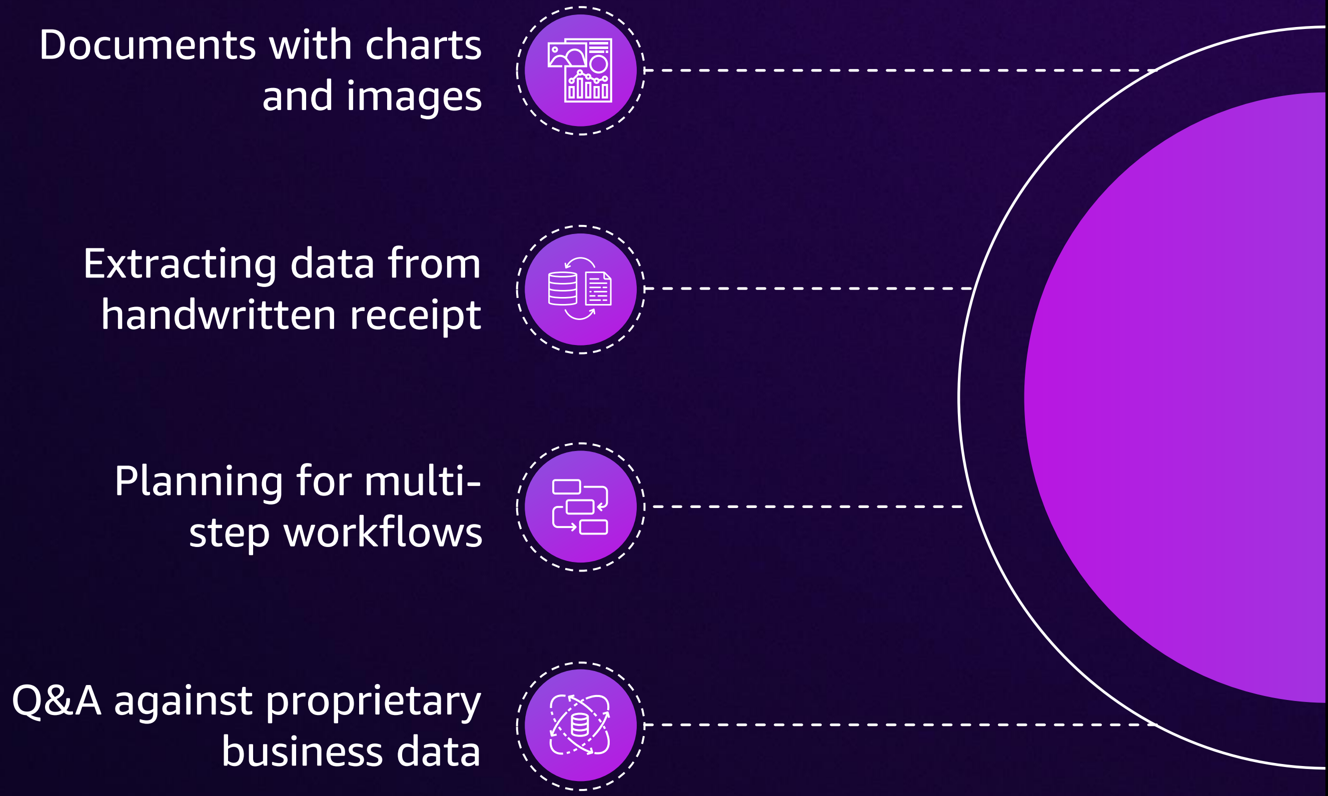
Amazon Nova Lite

Lowest-cost multimodal model that is lightning-fast for lightweight tasks

BEST COMBINATION OF SPEED, ACCURACY, AND COST

Key attributes

- ✓ Input modalities: Text, image, video
- ✓ Context Length: 300K tokens
- ✓ Languages: up to 200
- ✓ Customization: Fine-Tuning (text, images, video) & distillation (student)
- ✓ Latency: 157 tokens per second
- ✓ Price/M Tokens: $0.06 input; $0.24 output



Documents with charts and images

Extracting data from handwritten receipt

Planning for multi-step workflows

Q&A against proprietary business data

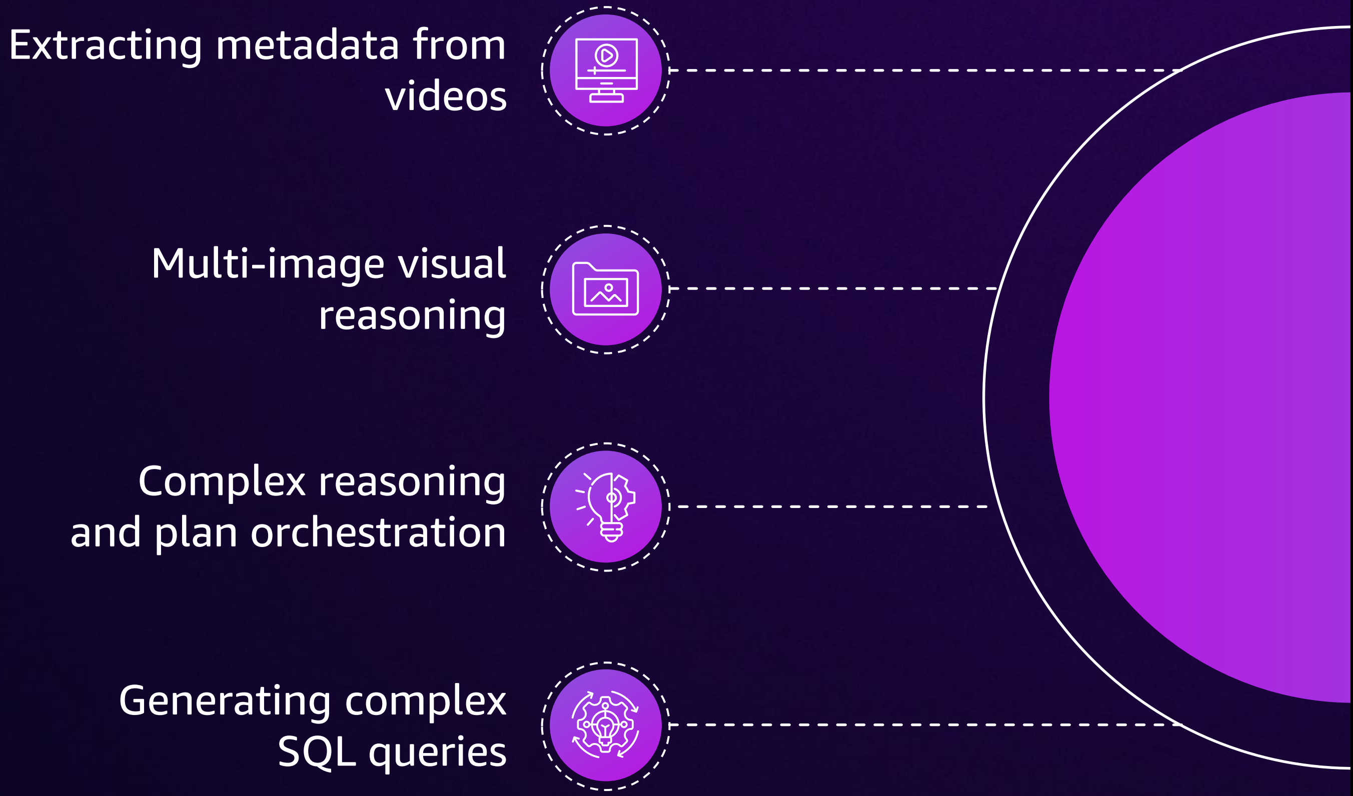
Amazon Nova Pro

Highly capable multimodal model with the best combination of accuracy, speed, and cost for a wide range of tasks

BEST COMBINATION OF SPEED, ACCURACY, AND COST

Key attributes

- ✓ Input modalities: Text, image, video
- ✓ Context Length: 300K tokens
- ✓ Languages: up to 200
- ✓ Customization: Fine-Tuning (text, images, video) & distillation (teacher)
- ✓ Latency: 100 tokens per second
- ✓ Price/M Tokens: $0.8 input; $3.2 output



Extracting metadata from videos

Multi-image visual reasoning

Complex reasoning and plan orchestration

Generating complex SQL queries

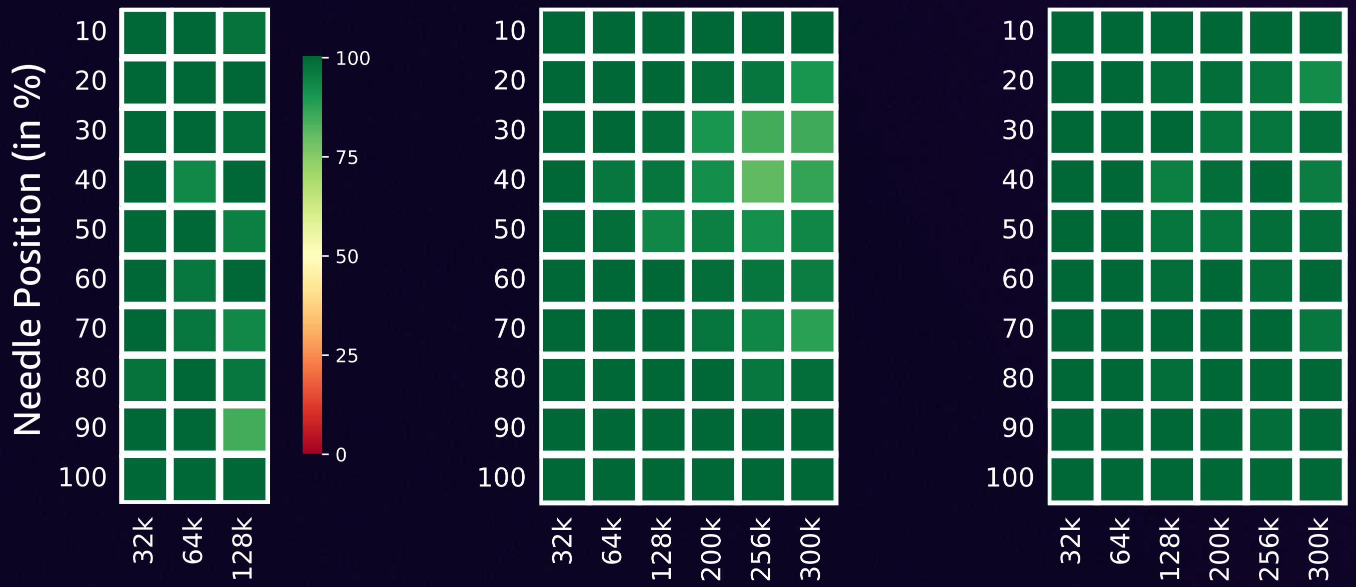
Amazon Nova's Context

HIGH ACCURACY ACROSS DIFFERENT CONTEXT LENGTHS

Amazon Nova Micro 128k

Amazon Nova Lite 300k

Amazon Nova Pro 300k



Amazon Nova Long Context

- ✓ Designed for high accuracy on long documents
- ✓ Ideal for complex understanding and retrieval tasks
- ✓ Strong multi-modal understanding

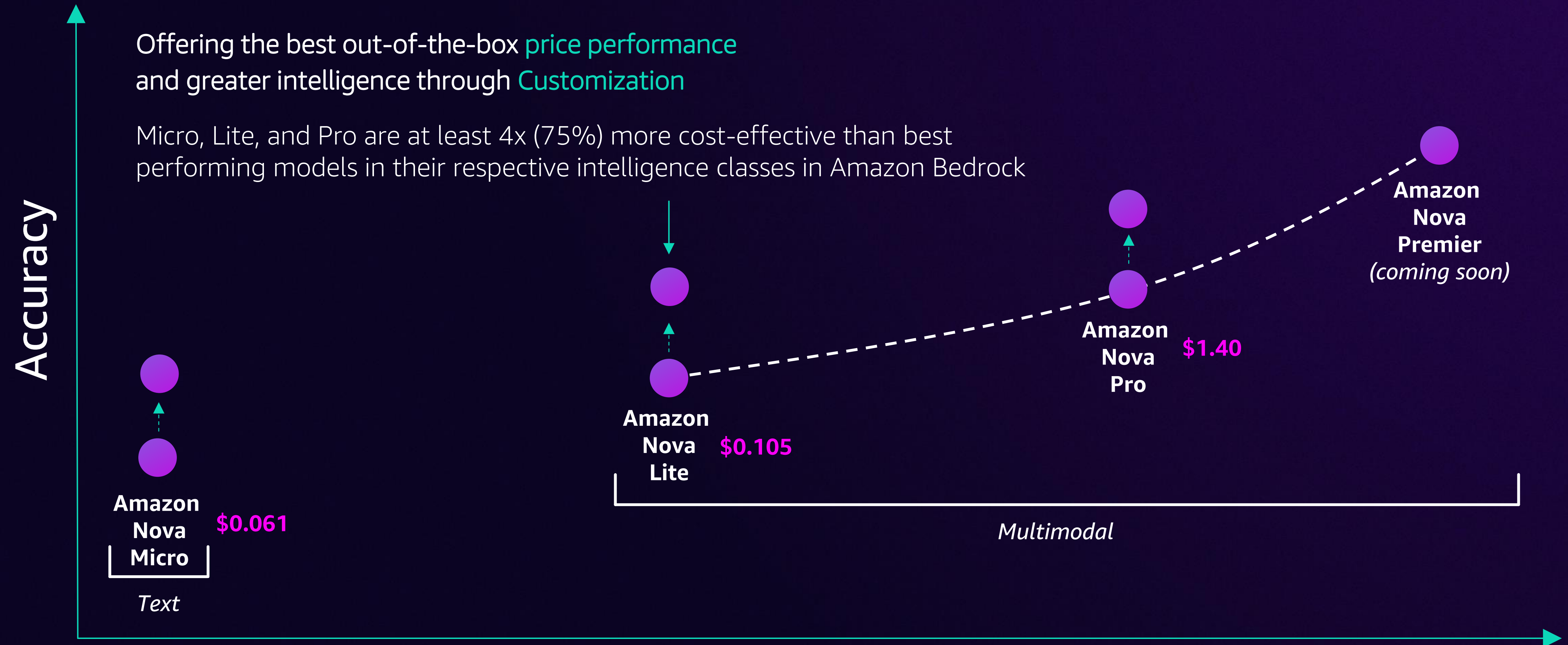
Needle-In-A-Haystack Accuracy at 128K

Micro	Lite	Pro
96.5%	98.4%	98.8%

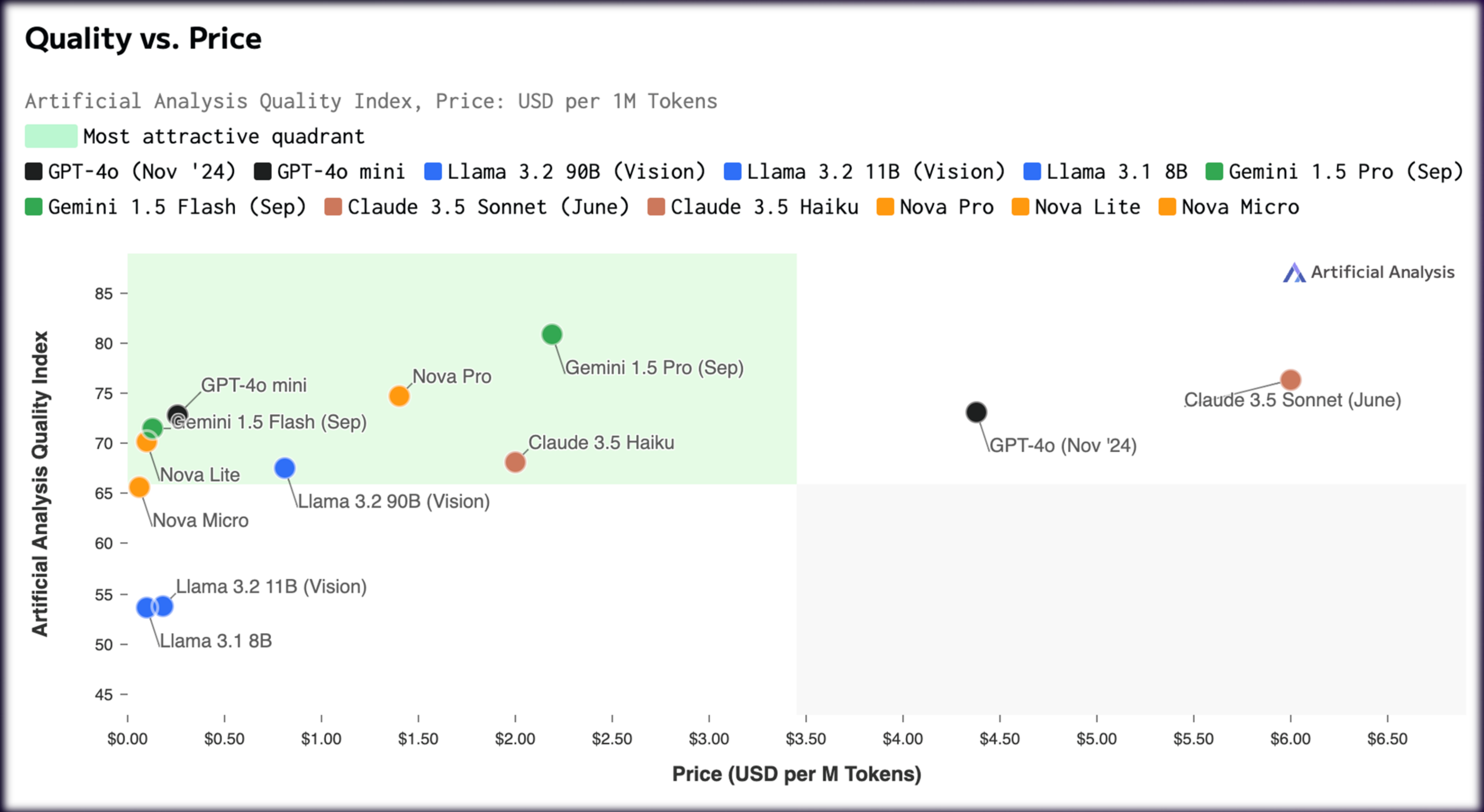
Amazon Nova: Leading Price Performance

Offering the best out-of-the-box price performance and greater intelligence through Customization

Micro, Lite, and Pro are at least 4x (75%) more cost-effective than best performing models in their respective intelligence classes in Amazon Bedrock



Amazon Nova: Industry-Leading Price Performance



Quality vs. Price

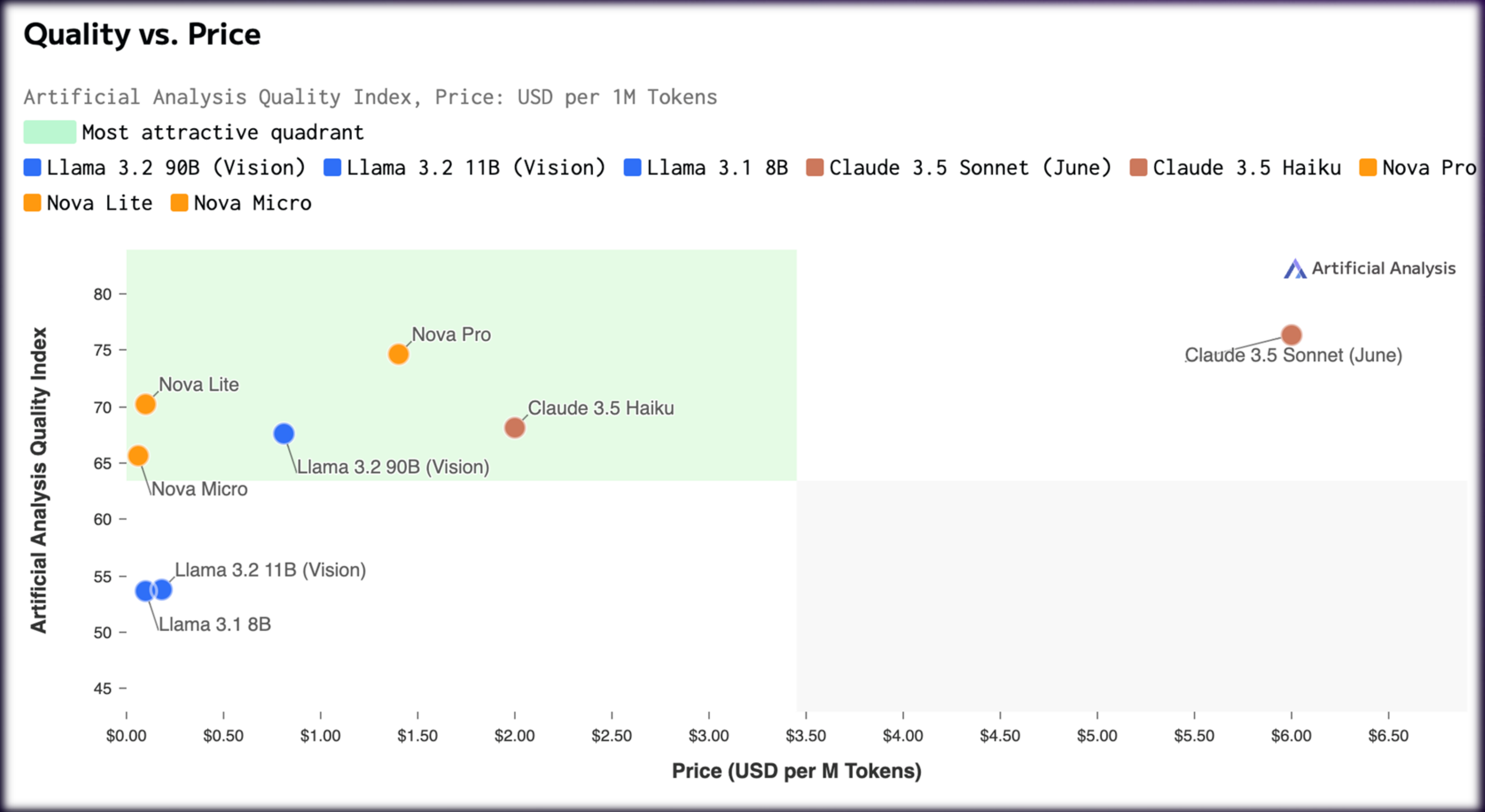
Artificial Analysis Quality Index, Price: USD per 1M Tokens

🟩 Most attractive quadrant

⬛ GPT-4o (Nov '24) ⬛ GPT-4o mini 🟦 Llama 3.2 90B (Vision) 🟦 Llama 3.2 11B (Vision) 🟦 Llama 3.1 8B 🟩 Gemini 1.5 Pro (Sep) 🟩 Gemini 1.5 Flash (Sep) 🟫 Claude 3.5 Sonnet (June) 🟫 Claude 3.5 Haiku 🟧 Nova Pro 🟧 Nova Lite 🟧 Nova Micro

Amazon Nova: Industry-Leading Price Performance

Micro, Lite, and Pro are at least 4x (75%) more cost-effective than best performing models in their respective intelligence classes in Amazon Bedrock



Quality vs. Price

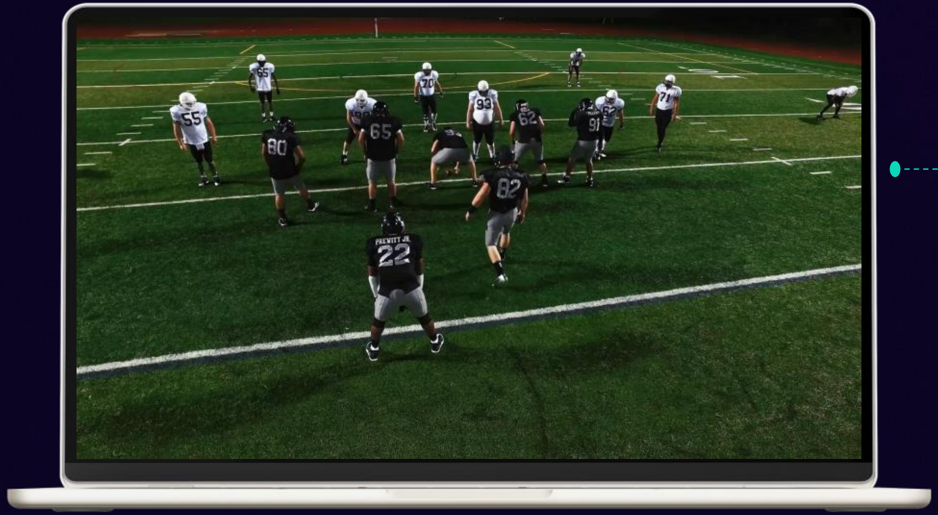
Artificial Analysis Quality Index, Price: USD per 1M Tokens

Native Video Understanding

Input Prompt: Describe the video

Prompt

Input Video:



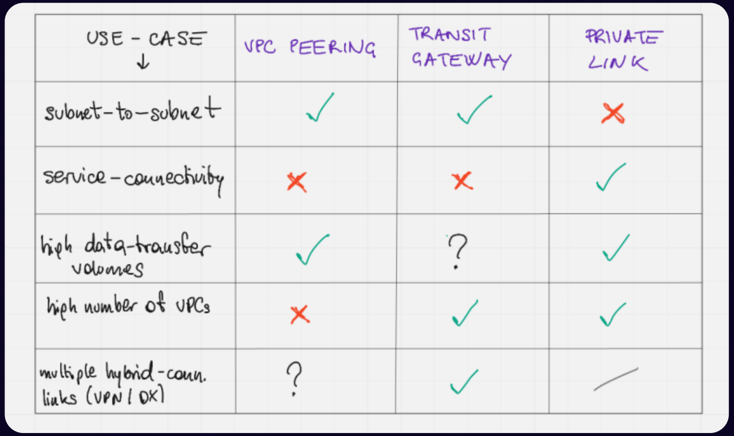
Video

Output

The video depicts a football game in progress, featuring two teams on a well-marked green field. Initially, the teams are positioned in a classic formation, with players from both sides facing each other, ready for the play. The players are equipped with helmets and jerseys, indicating a formal match. As the video progresses, the play commences with the offensive team executing a pass, leading to a tackle by the defensive team. The action is intense, showcasing the physicality and strategy inherent in football. The camera captures the play from an aerial perspective, providing a clear view of the field and the players' movements.

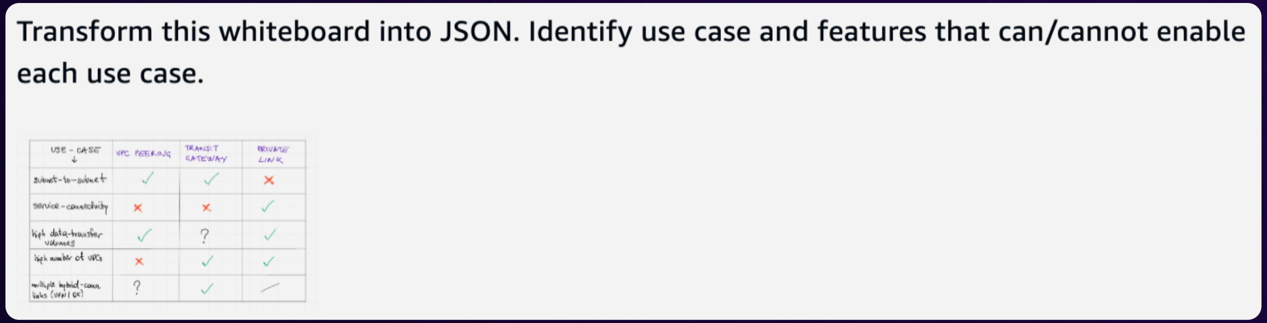
Handwriting understanding

Input Prompt: Transform this whiteboard into JSON. Identify use case and features that can/cannot enable each use case.



USE - CASE ↓	VPC PEERING	TRANSIT GATEWAY	PRIVATE LINK
subnet-to-subnet	✓	✓	✗
service-connectivity	✗	✗	✓
high data-transfer volumes	✓	?	✓
high number of VPCs	✗	✓	✓
multiple hybrid-conn. links (VPN / DX)	?	✓	—

Image



Transform this whiteboard into JSON. Identify use case and features that can/cannot enable each use case.

Amazon Nova

State-of-the-art creative content generation models

Amazon Nova **Canvas**

State-of-the-art image generation model

Amazon Nova **Reel**

State-of-the-art video generation model

Amazon Nova Canvas

IMAGE GENERATION MODEL

Enables creation and editing of studio-quality images for advertising, marketing, ecommerce, and other domains (e.g., a natural exfoliating body scrub with a vibrant colorful splash of mint and cucumber juice in a bathroom with a beautiful natural view)

Image dimensions

Up to 2K x 2K

Core features

Image generation and image editing

Fine-tuning

Coming Soon!

Advanced features

Color palette using hex codes, background removal, image conditioning, **RAI (watermarking, output indemnity)**

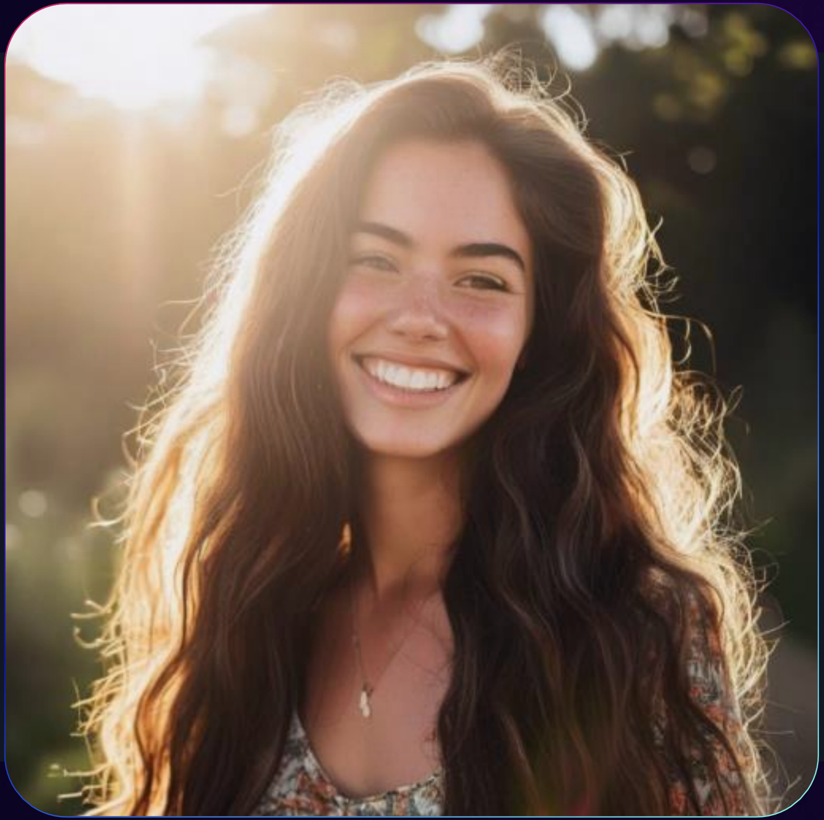
Recommended Use Case

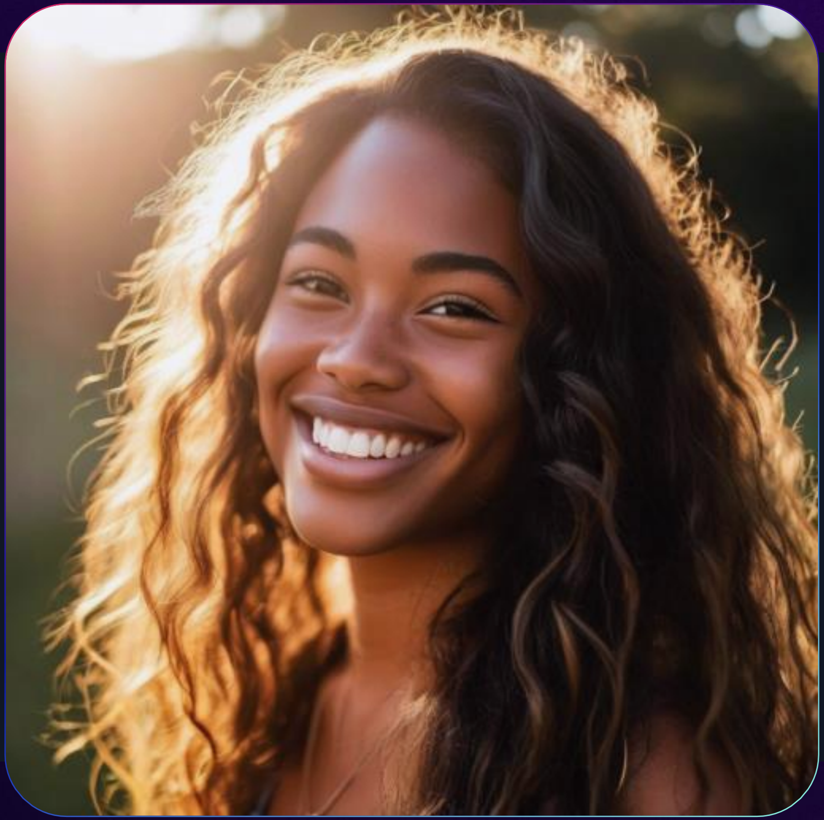
Creation of images for advertising, marketing, social media, publishing, ecommerce

Text to image

GENERATE IMAGES FROM TEXT PROMPTS

A portrait of a smiling young woman with long, flowing hair, standing in natural sunlight



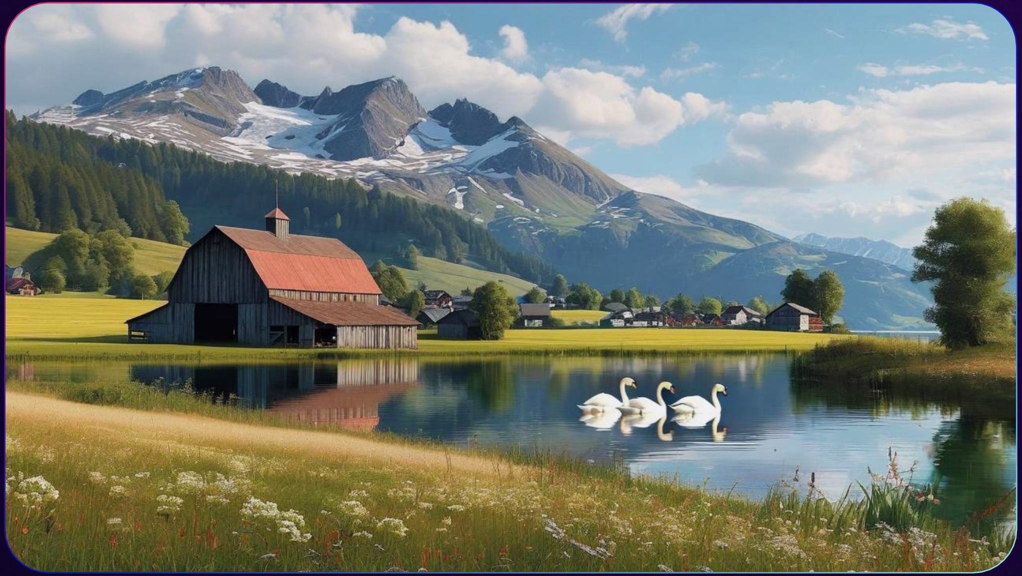


Inpainting

EDIT AN IMAGE

Input image

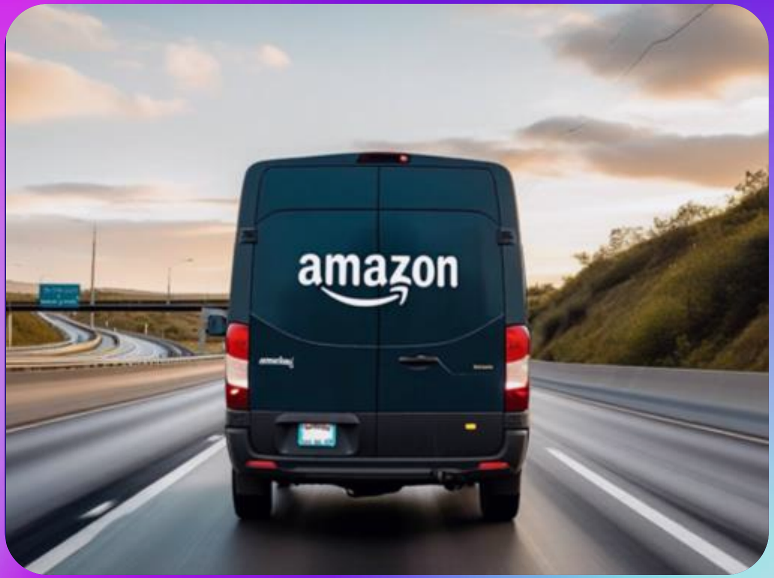
Inpainting the image with a group of swans



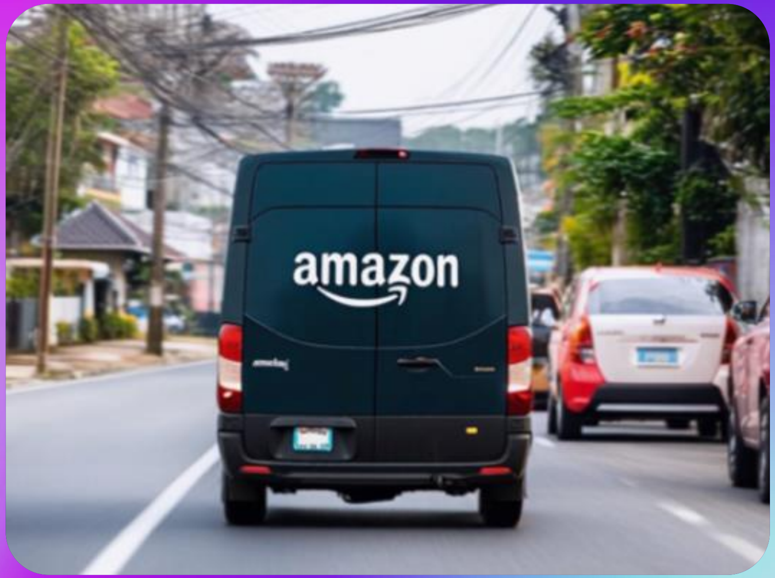
Outpainting

GENERATE DIFFERENT BACKGROUNDS

Input Image



Generated Images With Different Background



Color palette

USING HEX CODES TO CONTROL OUTPUT

Input Image

a jar of salad dressing in a rustic kitchen surrounded by fresh vegetables with studio lighting

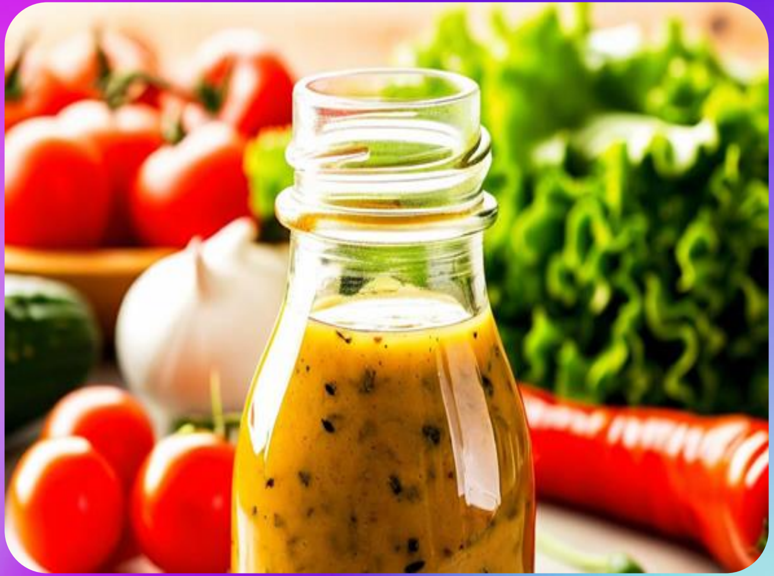
+

Input hex codes

['#ff8080', '#ffb280', '#ffe680', '#e5ff80']

→

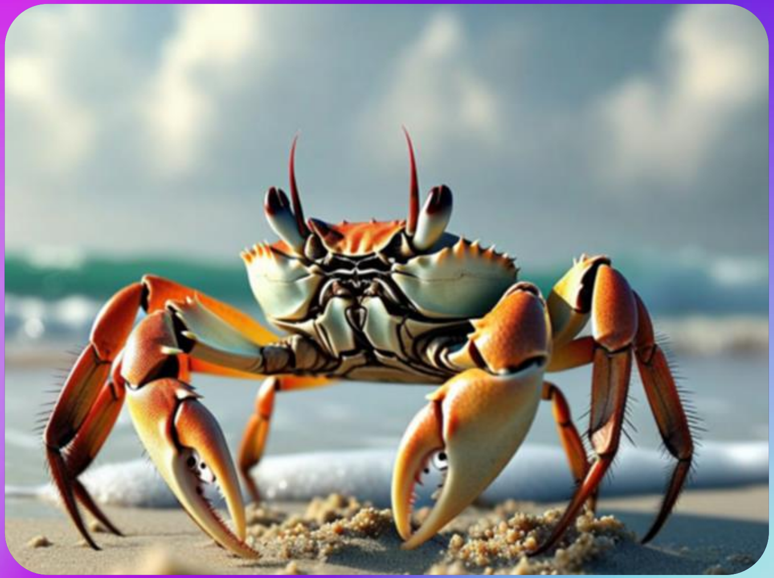
Color-controlled output



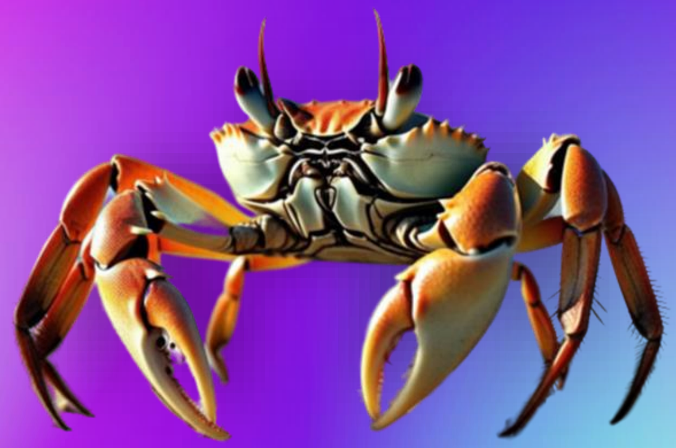
Background Removal

REMOVE BACKGROUND FROM IMAGES

Input Image



Removed background



Amazon Nova Reel

VIDEO GENERATION MODEL

Enables creation of short videos through simple prompting via input images or a natural language prompts (e.g., perfume bottle with jasmine and citrus themes and a calming, natural setting).

Video duration

6 seconds

Upcoming features

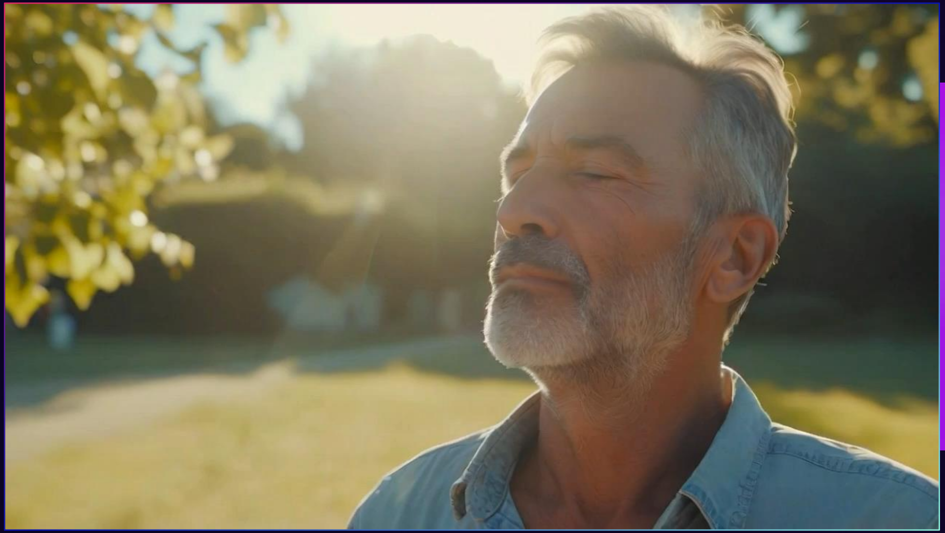
Video editing, long videos with storyboarding

Recommended use case

Creation of short video clips for marketing, advertising, entertainment, and social media applications

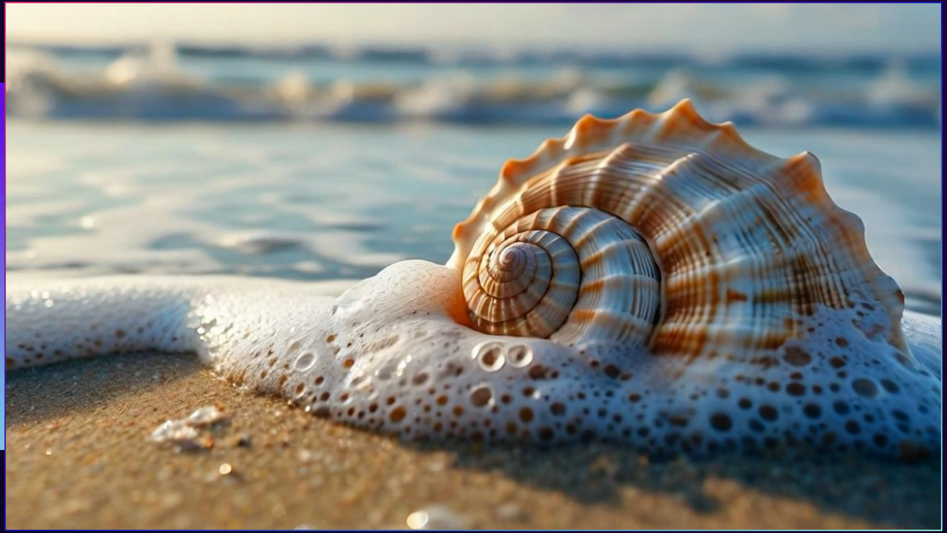
Features

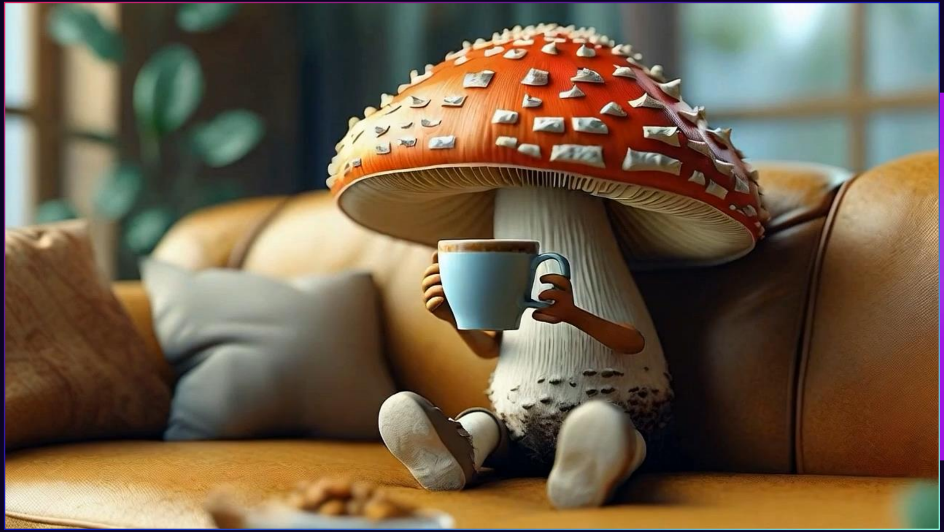
Text-to-video generation, image-to-video generation



"Slow cam of a man middle age; 4k; Cinematic; in a sunny day; peaceful; highest quality; dolly in"

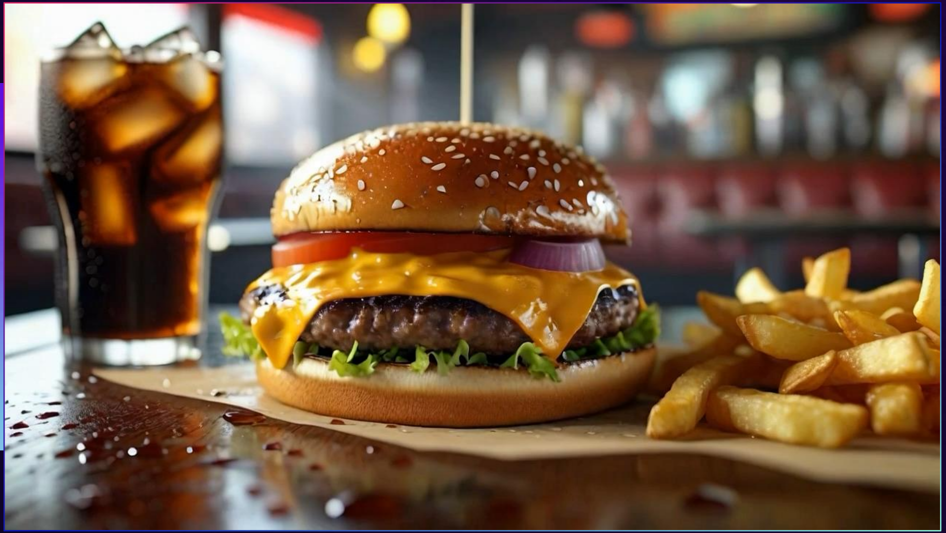
"Closeup of a large seashell in the sand. Gentle waves flow around the shell. Camera zoom in."





"A mushroom drinking a cup of coffee while sitting on a couch, photorealistic."

"Cinematic dolly shot of a juicy cheeseburger with melting cheese, fries, and a condensation-covered cola on a worn diner table. Natural lighting, visible steam and droplets. 4k, photorealistic, shallow depth of field



Amazon Nova in Action

AWS Support | Agent Application

Overview

aws

Support Engineering

Amazon Nova Micro, Lite, and Pro multimodal understanding, customer intent classification, and customer issue resolution

Challenge

Complex technical issues can take a few hours to resolve

Solution

Amazon Nova multimodal capabilities help Support staff diagnose issues and recommend best practices

Impact

Resolution Time reduced to under ~**5 mins**

Why Amazon Nova?

Agentic Workflows & RAG

RAG/in-context learning, and tools increase automation

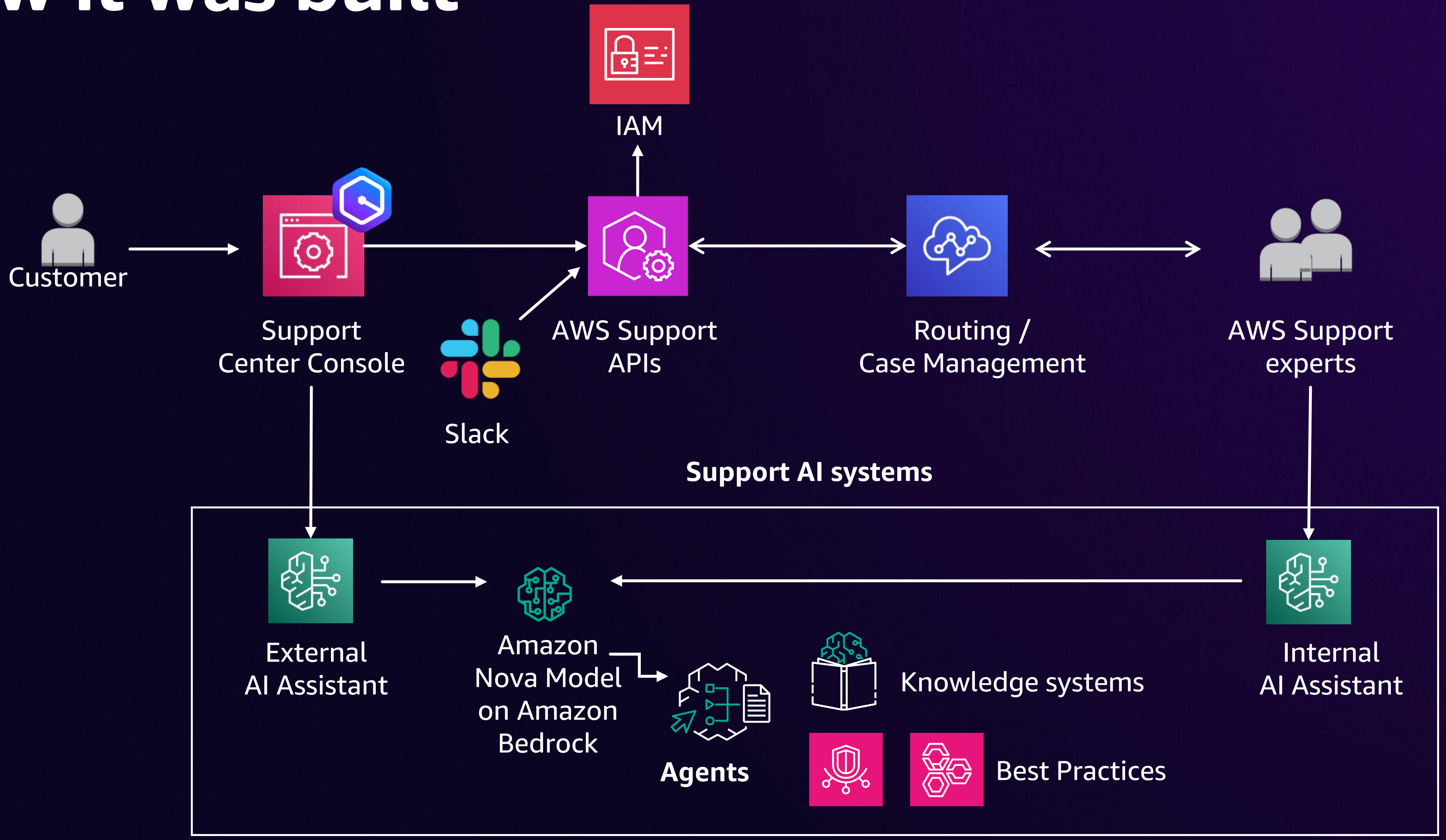
Customization

Customization enhances model's domain knowledge

Frontier Intelligence

Understands diverse inputs: logs, diagrams, more

How it was built



Recap

aws

Support Engineering

Amazon Nova assists AWS Support front-line staff to reduce resolution time from hour to minutes for complex problems.

✓ Agentic Workflows & RAG

Amazon Nova RAGs/in-context learning, and tools deliver domain knowledge.

✓ Customization

Amazon Nova customization enhances Model's technical understanding

✓ Frontier Intelligence

Amazon Nova understands diverse inputs: logs, diagrams, more.

Prime Video | Season Recap

Overview

Amazon Nova Pro was used to build a cinematic quality recap from hours of video for *Bosch: Legacy*

Challenge

Show recaps take weeks to generate

Solution

Amazon Nova can analyze hours of video content, deeply understands the story and plot and compiles it into a recap video

Impact

Cinematic quality season recap in a matter of hours

Why Amazon Nova?

Frontier Intelligence

Trained on videos. Generates denser descriptions with fewer hallucinations.

Long Context

Process hours of video content, resulting in better understanding and reasoning

How it was built

Bosch Legacy S2 (10 episodes, ~450 minutes)

Gen AI →

3-minute Narrated Recap

Synopsis Generation → Voice Synthesis → Visual Montage

Synopsis Generation

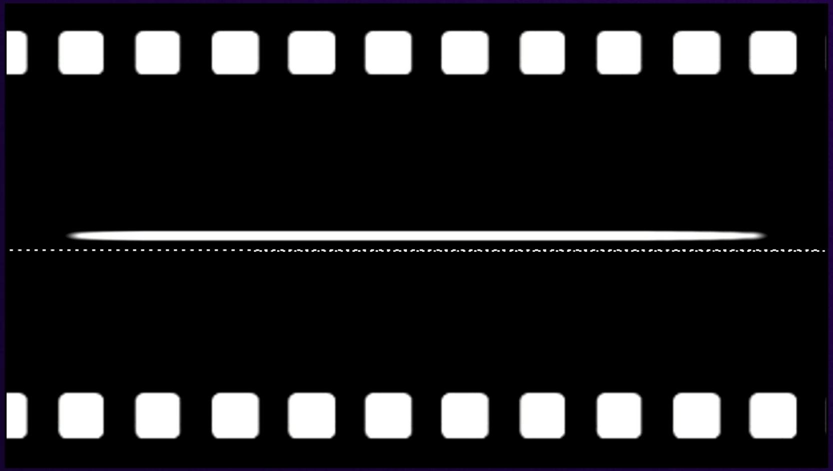
Synopsis →

In the gripping second season of "Bosch: Legacy", Harry Bosch finds himself entangled in a complex web of murder, corruption and personal struggles. Harry's daughter Maddie is abducted by Kurt Dockweiler

Audio Synthesis

In the gripping second season of "Bosch: Legacy", Harry Bosch finds himself entangled in a complex web of murder, corruption and personal struggles…

Narration



Visual Montage

From narration to final audio visual assembly

In the gripping second season of "Bosch: Legacy",

Harry Bosch finds himself entangled in a **complex web** of **murder**, **corruption** and **personal struggles**

Recap

Amazon Nova Pro can be used to build recaps for streaming video, as you can see with this example from Amazon MGM Studios Original Bosch: Legacy

✓ **Frontier Intelligence**

Trained on entire videos.

✓ **Long Context**

Understand and process hours of video content

Amazon Q Developer

Overview

Amazon Nova Pro powers natural language to tool-use

Challenge

Managing AWS resources requires complex data retrieval across multiple pages

Solution

Use Amazon Q in natural language to introspect across account resources

Impact

Simplification of using AWS service with increased developer's productivity

Why Amazon Nova?

Frontier Intelligence

Provides 94.3% accuracy (+3.4% gain) in generating natural language to tool use

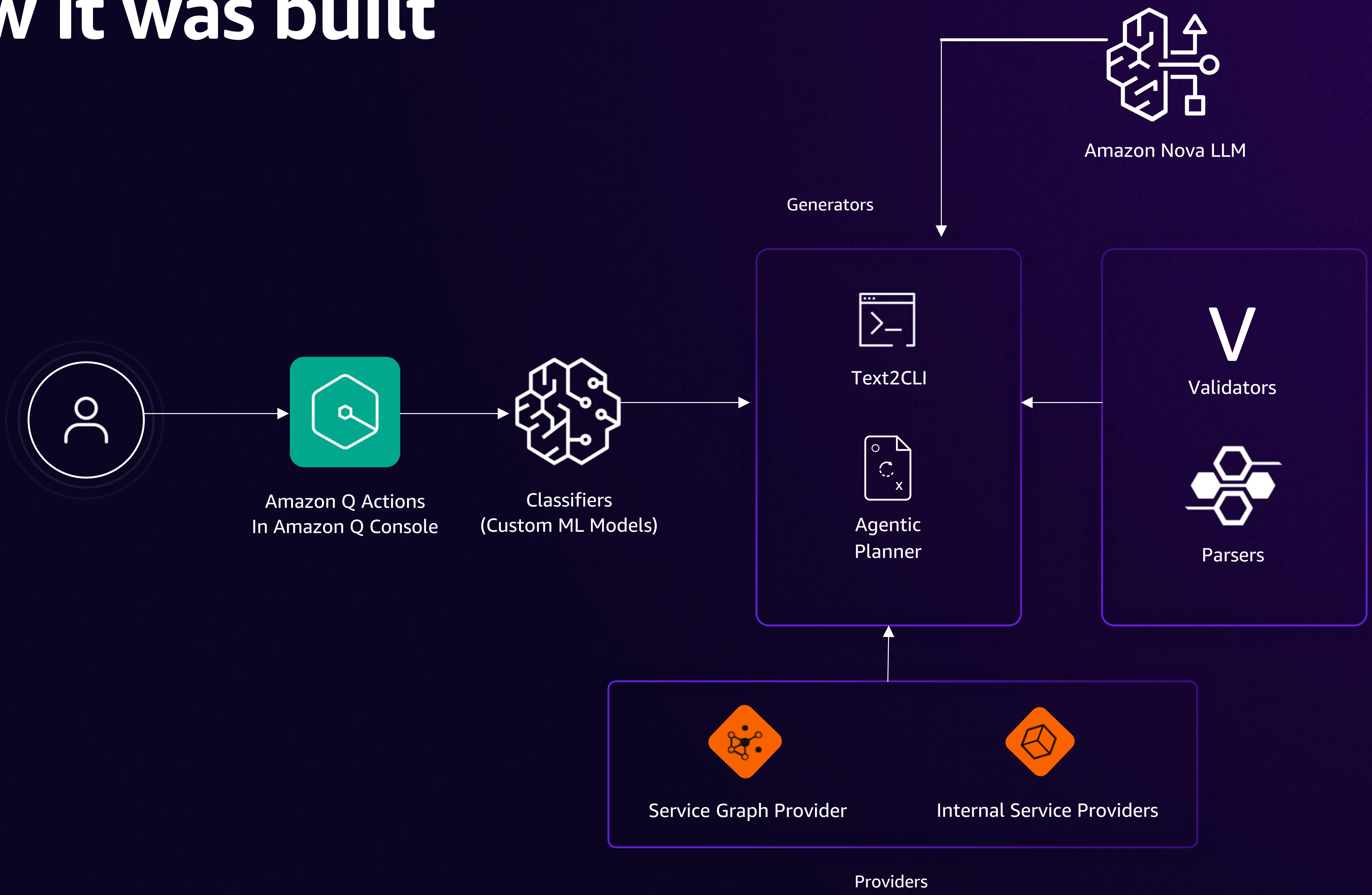
Speed

Twice as fast as its competitors in same weight class

Price Performance

Provided higher accuracy, higher speed at a lower price point

How it was built



Recap

Amazon Nova Pro assists Amazon Q Developer to achieve higher accuracy and speed in a cost effective manner for natural language to tool use

✓ Frontier Intelligence

Amazon Nova provides higher accuracy for converting natural language to tool use

✓ Speed

Amazon Nova is twice as fast as its competitors

✓ Price Performance

Amazon Nova provides better performance at a lower price point

Amazon Ads Video Generator

Overview

amazonadvertising

Amazon Nova Reel helps Amazon Ads reduce cost of Video Ads production

Challenge

Video ads drive better engagement but are costly

Solution

Amazon Ads leverages Amazon Nova Reel to make AI-assisted video ads

Impact

Amazon Nova Reel makes higher engagement ads affordable for more advertisers

Why Amazon Nova?

Responsible AI controls

Strong RAI controls like Invisible watermarks and indemnity

Advanced Controls

Supports advanced features for generation and editing

Customization

Fine-tuning for alignment with brand aesthetics (coming soon!)

Recap

amazonadvertising

Amazon Nova Reel helps Amazon Ads reduce cost of Video Ads production.

✓ **Responsible AI Controls**

Amazon Nova Content Creation models outperform comparable models

✓ **Advanced Controls**

Supports advanced features for generation and editing

✓ **Customization**

Fine-tuning for alignment with brand aesthetics (coming soon!)

Next Steps and Resources

Amazon Nova
Foundation Models

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

Understanding Models

- ✓ Amazon Nova Micro
- ✓ Amazon Nova Lite
- ✓ Amazon Nova Pro
- ✓ Amazon Nova Premier (coming soon)

Content Creation Models

- ✓ Amazon Nova Canvas
- ✓ Amazon Nova Reel

What's Next for Amazon Nova?

NEW

Amazon Nova
Speech-to-Speech

COMING SOON

NEW

Amazon Nova
Any-to-Any

COMING SOON

Get started today with Amazon Nova

✓ **Run evaluations** comparing Amazon Nova in your Amazon Bedrock implementations

✓ Test Amazon Nova on **Amazon Bedrock Playground**

✓ Check out **Prompt Guidance** and **Quick-Start Guides** available via the link on the right

Thank you!

Please complete the session survey in the mobile app

Jay Lee

jaywlee@amazon.com

Shubham Katiyar

shubham@amazon.com

Gerard Medioni

medioni@amazon.com

Jamie St. Onge

grecjami@amazon.com

Patrick Nguyen

nguyenpt@amazon.com