

The background features a dark blue gradient with large, overlapping, semi-transparent shapes in shades of purple and magenta. Two thin, light blue lines cross the scene diagonally. The text is positioned on the left side of the image.

AWS re:Invent

DECEMBER 2 - 6, 2024 | LAS VEGAS, NV

AIM395 - NEW

Amazon Nova understanding models

Firat Elbey

(he/him/his)
Principal Product Manager
Amazon AGI

Imre Kiss

(he/him/his)
Director, Applied Science
Amazon AGI

Fabian Triefenbach

(he/him/his)
Sr. Manager, Applied Science
Amazon AGI

Peter Goldstein

(he/him/his)
Chief Product & AI Strategist
Hearst Corporation



Amazon Nova Understanding Models

01 Introduction

02 Key Attributes and Benchmarks

03 Core Capabilities

04 Customer Success Story

05 Key takeaways

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

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

Higher Intelligence →

Amazon Nova Micro

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

GENERALLY AVAILABLE

Amazon Nova Lite

Very low cost multimodal model that is lightning fast for processing image, video, and text inputs

GENERALLY AVAILABLE

Amazon Nova Pro

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

GENERALLY AVAILABLE

Amazon Nova Premier

Most capable multimodal model for complex reasoning tasks and for use as the best teacher for distilling custom models

COMING SOON

← Lower Latency & Cost



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

Key attributes

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

TEXT INTELLIGENCE				
TASK	NOVA MICRO	GEMINI 1.5 FLASH 8B	LLAMA 3.1 8B	
1 Language Understanding (MMLU - CoT)	77.6%	68.1%	73.0%	
2 General Reasoning (DROP - CoT)	79.3%	<u>68.1%</u>		
3 Common Sense Reasoning (ARC-c)	90.2%	<u>88.7%</u>	83.4%	
4 Deep Reasoning (GPQA - Main)	40.0%	<u>33.5%</u>	30.4%	
5 Mathematics (GSM8K)	92.3%	<u>84.5%</u>	84.5%	
6 Mathematics (MATH - CoT)	69.3%	58.7%	51.9%	
7 Python Code Generation (HumanEval)	81.1%	<u>81.1%</u>	72.6%	
8 Multi-step reasoning (Big Bench Hard - CoT)	79.5%	69.5%		
9 Instruction Following (IFEval)	87.2%	<u>86.1%</u>	<u>85.0%</u>	
10 Translation: EN to 14 Languages (Flores200)	40.2	<u>38.2</u>	<u>32.7</u>	
11 Translation: 14 Languages to EN (Flores200)	42.6	<u>41.4</u>	<u>36.5</u>	
AGENTIC WORKFLOWS				
12 RAG - Text (CRAG)	43.1%	<u>37.7%</u>	<u>42.2%</u>	
13 Agent API Orchestration (BFCLv3 - 11/17)	56.2%		49.9%	

*The numbers for other models are a mix of self-reported and measured evaluations (underlined). For a more detailed overview of evaluations, please refer to the **Amazon Nova Tech Report***



Amazon Nova Lite

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

Key attributes

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



TEXT INTELLIGENCE

TASK	NOVA LITE	CLAUDE 3.5 HAIKU	GPT-4o MINI	GEMINI 1.5 FLASH	LLAMA 3.2 11B
1 Language Understanding (MMLU - CoT)	80.5%	80.3%	82.0%	78.9%	73.0%
2 General Reasoning (DROP - CoT)	80.2%	83.1%	79.7%	78.4%	
3 Common Sense Reasoning (ARC-c)	92.4%	<u>90.9%</u>	<u>92.3%</u>	<u>94.3%</u>	83.4%
4 Deep Reasoning (GPQA - Main)	42.0%	<u>37.5%</u>	<u>41.7%</u>	<u>45.1%</u>	32.8%
5 Mathematics (GSM8K)	94.5%	<u>93.8%</u>	<u>86.4%</u>	86.2%	84.5%
6 Mathematics (MATH - CoT)	73.3%	69.4%	70.2%	77.9%	51.9%
7 Python Code Generation (HumanEval)	85.4%	88.1%	87.2%	<u>81.1%</u>	72.6%
8 Multi-step reasoning (Big Bench Hard - CoT)	82.4%	86.6%	<u>81.0%</u>	85.5%	
9 Instruction Following (IFEval)	89.7%	85.9%	<u>87.4%</u>	<u>91.6%</u>	<u>85.0%</u>
10 Translation: EN to 14 Languages (Flores200)	41.5	<u>40.0</u>	<u>41.1</u>	<u>40.0</u>	<u>33.0</u>
11 Translation: 14 Languages to EN (Flores200)	43.1	<u>40.2</u>	<u>41.9</u>	<u>42.9</u>	<u>36.3</u>

VISUAL INTELLIGENCE

12 Visual Reasoning (MMMU - CoT)	56.2%		59.4%	62.3%	50.7%
13 Image Understanding (ChartQA)	86.8%		<u>79.2%</u>	85.4%	83.4%
14 Document Understanding (DocVQA)	92.4%			89.9%	88.4%
15 Video Captioning (VATEX)	77.8%			57.1%	
16 Visual Question Answering (TextVQA)	80.2%		<u>70.3%</u>	78.7%	<u>71.3%</u>
17 User Interface Grounding (GroundUI-1K)	80.2%		<u>7.2%</u>	<u>59.9%</u>	<u>3.7%</u>

AGENTIC WORKFLOWS

18 RAG - Text (CRAG)	43.8%	<u>31.9%</u>	<u>49.9%</u>	<u>42.4%</u>	<u>42.2%</u>
19 Agent API Orchestration (BFCLv3 - 11/17)	66.6%		60.7%	56.2%	49.9%
20 Web Browser Grounding (VisualWebBench)	77.7%		<u>71.3%</u>	<u>76.1%</u>	<u>65.1%</u>
21 Multimodal Web Agents (Mind2Web)	60.7%		<u>58.6%</u>	<u>46.2%</u>	<u>22.1%</u>

Amazon Nova Pro

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

Key attributes

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



TEXT INTELLIGENCE

TASK	NOVA PRO	CLAUDE 3.5 SONNET V2	GPT-4o	GEMINI 1.5 PRO	LLAMA 3.2 90B
1 Language Understanding (MMLU - CoT)	85.9%	89.3%	88.7%	85.9%	86.0%
2 General Reasoning (DROP - CoT)	85.4%	88.3%	83.4%	74.9%	
3 Common Sense Reasoning (ARC-c)	94.8%	<u>96.3%</u>	<u>96.2%</u>	<u>95.4%</u>	94.8%
4 Deep Reasoning (GPQA - Main)	46.9%	<u>58.0%</u>	<u>48.4%</u>	<u>55.1%</u>	46.7%
5 Mathematics (GSM8K)	94.8%	<u>96.5%</u>	<u>92.6%</u>	90.8%	95.1%
6 Mathematics (MATH - CoT)	76.6%	78.3%	76.6%	86.5%	68.0%
7 Python Code Generation (HumanEval)	89.0%	93.7%	90.2%	<u>87.8%</u>	80.5%
8 Multi-step reasoning (Big Bench Hard - CoT)	86.9%	93.2%	<u>83.0%</u>	89.2%	
9 Instruction Following (IFEval)	92.1%	90.2%	<u>89.8%</u>	<u>91.7%</u>	<u>90.9%</u>
10 Translation: EN to 14 Languages (Flores200)	43.4	<u>42.5</u>	<u>43.1</u>	<u>43.0</u>	<u>39.7</u>
11 Translation: 14 Languages to EN (Flores200)	44.4	<u>43.5</u>	<u>43.9</u>	<u>45.6</u>	<u>43.7</u>

VISUAL INTELLIGENCE

12 Visual Reasoning (MMMU - CoT)	61.7%	70.4%	69.1%	65.9%	60.3%
13 Image Understanding (ChartQA)	89.2%	90.8%	85.7%	87.2%	85.5%
14 Document Understanding (DocVQA)	93.5%	94.2%	92.8%	93.1%	90.1%
15 Video Captioning (VATEX)	77.8%			64.6%	
16 Visual Question Answering (TextVQA)	81.5%	<u>61.7%</u>	<u>77.2%</u>	78.7%	<u>80.7%</u>
17 User Interface Grounding (GroundUI-1K)	81.4%	16.3%	13.4%	35.2%	<u>8.3%</u>

AGENTIC WORKFLOWS

18 RAG - Text (CRAG)	50.3%	<u>52.6%</u>	<u>52.0%</u>	<u>48.9%</u>	<u>45.2%</u>
19 Agent API Orchestration (BFCLv3 - 11/17)	68.4%	61.3% (V1)	68.9%	59.8%	54.3%
20 Web Browser Grounding (VisualWebBench)	79.7%	<u>76.7%</u>	<u>77.5%</u>	<u>76.4%</u>	<u>73.2%</u>
21 Multimodal Web Agents (Mind2Web)	63.7%	<u>61.6%</u>	<u>55.0%</u>	<u>58.4%</u>	<u>21.6%</u>

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

Quality vs. Price

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

Most attractive quadrant

■ Llama 3.2 90B (Vision) ■ Llama 3.2 11B (Vision) ■ Llama 3.1 8B ■ Claude 3.5 Sonnet (June) ■ Claude 3.5 Haiku ■ Nova Pro
■ Nova Lite ■ Nova Micro



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

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

Amazon Nova Micro

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

GENERALLY AVAILABLE

Amazon Nova Lite

Very low cost multimodal model that is lightning fast for processing image, video, and text inputs

GENERALLY AVAILABLE

Amazon Nova Pro

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

GENERALLY AVAILABLE

Amazon Nova Premier

Most capable multimodal model for complex reasoning tasks and for use as the best teacher for distilling custom models

COMING SOON



Native Video Understanding

Input Prompt: Describe the video

Input Video:



Prompt

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.

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)	?	✓	/

Handwriting understanding

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)	?	✓	—



```
{  
  "name": "high_number_of_VPCs",  
  "features": {  
    "VPC_PEERING": "disabled",  
    "TRANSIT_GATEWAY": "enabled",  
    "PRIVATE_LINK": "enabled"  
  }  
}
```

Diagram understanding

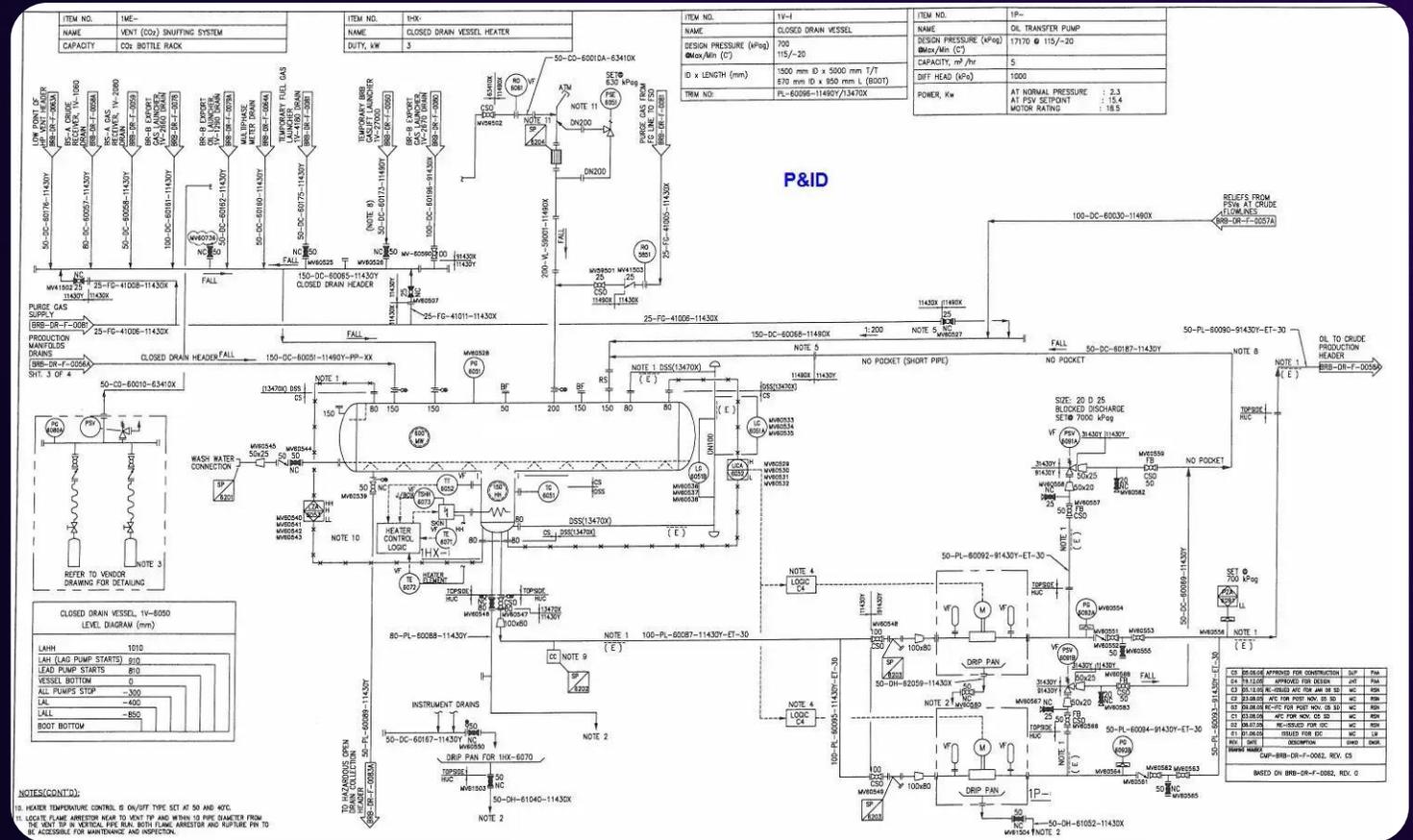
Input Prompt 1

You are an engineer, review the diagram and tell me what the process flow is about in detail and where it would be used

Input Prompt 2

How would you improve this process?

Input Image



Powered by Amazon Nova

Where Vision meets Agents



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

Deep Dive

Building Agents with Nova



Agentic Workflows: Industry-leading price-performance



Agentic Workflows

Table 1: Results on the Berkeley Function Calling Leaderboard (BFCL) v3 as of the 11/17/24 update. We include latest versions of the models available on the leaderboard at that time.

Category	Metric	Nova Pro	Claude 3.5 Sonnet (Jun)	Gemini 1.5 Pro 002	GPT- 40 (Aug)	Nova Lite	Gemini 1.5 Flash 002	GPT- 4o-mini (Jul)	Nova Micro	Llama 3.1 8B
Overall Cost latency	Accuracy (1)	68.36 +/- 1.33	61.29	59.80	68.94	66.58 +/- 1.35	55.33	60.72	56.22 +/- 1.42	49.86
	\$ per 1k Calls (L)	3.8	8.61	6.54	8.22	0.32	0.42	0.51	0.16	N/A
	Seconds (1)	0.95	3.88	2.99	1.51	0.56	1.14	1.64	0.52	N/A
Non-Live	AST (1)	90.08	70.04	88.0	85.9	87.54	79.69	84.25	87.15	83.62
	Execution (1)	89.82	66.27	91.41	85.64	86.41	80.64	84.12	89.71	87.29
Live	AST (1)	84.71	67.89	81.06	75.41	77.82	74.52	76.41	76.18	59.84
Multi-Turn	Overall (1)	45.12	40.00	16.25	45.25	50.25	12.50	28.25	15.50	10.50
Hallucination	Relevance (1)	95.12	68.29	75.61	63.41	97.56	78.05	80.49	87.80	78.05
	Irrelevance (1)	65.10	74.58	75.10	82.93	49.14	75.65	71.77	57.55	41.62

Agentic Workflows: Industry-leading price-performance



Agentic Workflows

Table 2 shows the results of our models on multimodal agent workflows along with other publicly-reported results. Both Amazon Nova models - Lite and Pro - demonstrate strong visual reasoning and agentic capabilities and achieve high scores on all three benchmarks.

	Amazon Nova Pro	Amazon Nova Lite	Claude 3.5 Sonnet (Oct)	Claude 3 Sonnet	GPT-4 (Apr)	GPT-4o (MMM)	GPT-4o mini (Jul)	Gemini 1.0 Pro (001)	Gemini 1.5 Pro (002)	Gemini 1.5 Flash (002)
VisualWebBench	79.7	77.7	76.7	65.8	64.6	64.6	71.3	48.0	76.4	76.1
MM-Mind2Web	63.8	60.7	61.6	-	36.8	-	58.6	17.9	58.4	46.2
GroundingUI-1K	81.4	80.2	17.3	-	-	13.4	-	-	35.2	-



Nova Computer Agent

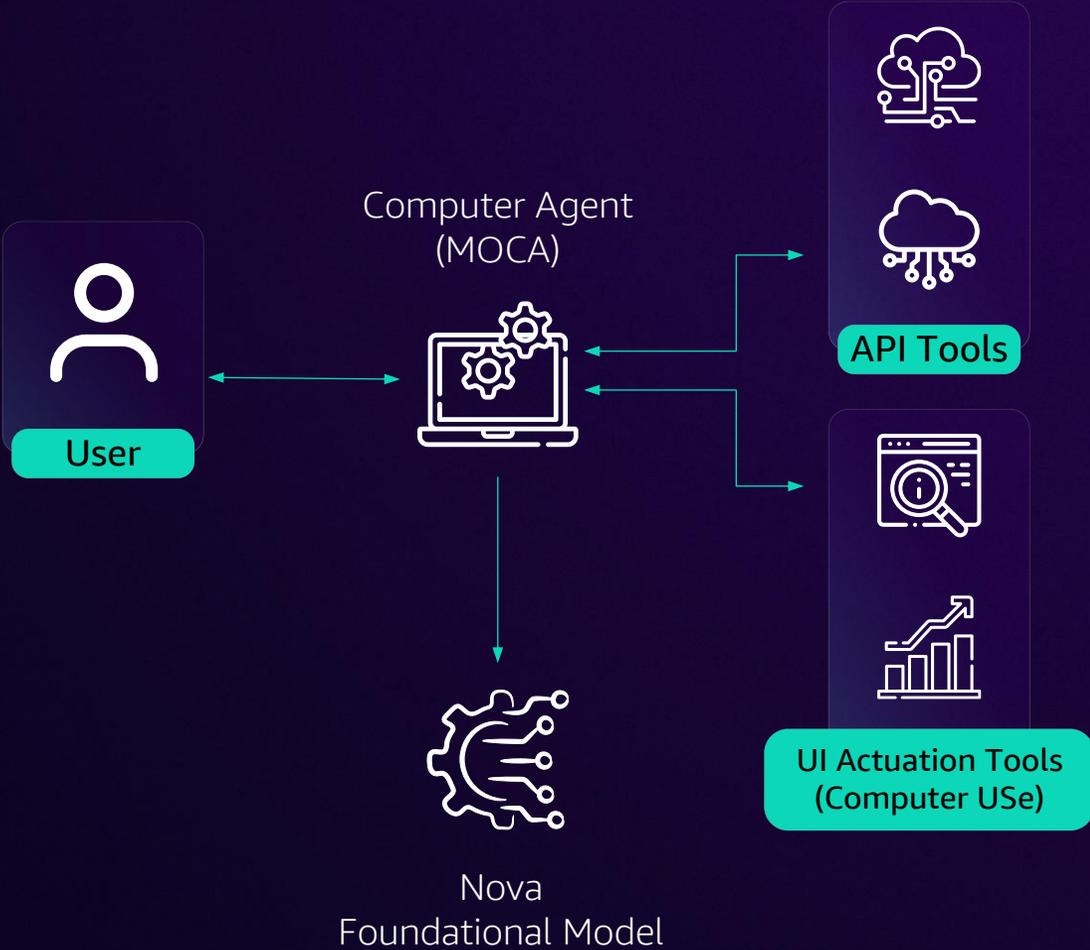
My on-Call Agent



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

Agentic Architecture: My on-Call Agent

 Agentic Workflows



Agentic Workflow: My on-Call Agent



Agentic Workflows



Prompt "You are an on-call engineer. Your task is to identify the root cause of the issue"

Tool Library Prompt "You are an on-call engineer. Your task is to identify the root cause of the issue"

API Tools: API Calls
UI Actuation Tools
Computer Agent
Drivers Browser

Prompt "Given the data collected so far determine the root cause and mitigation or request more data using the tool library"

Customizability with Nova

Fine-tune Nova models with your proprietary data



Easy customization via fine-tuning on custom data

Accuracy boost:

- ✓ Better zero-shot accuracy
- ✓ Better price-performance trade-off

Multimodal: Improve context-specific responses on image and video inputs



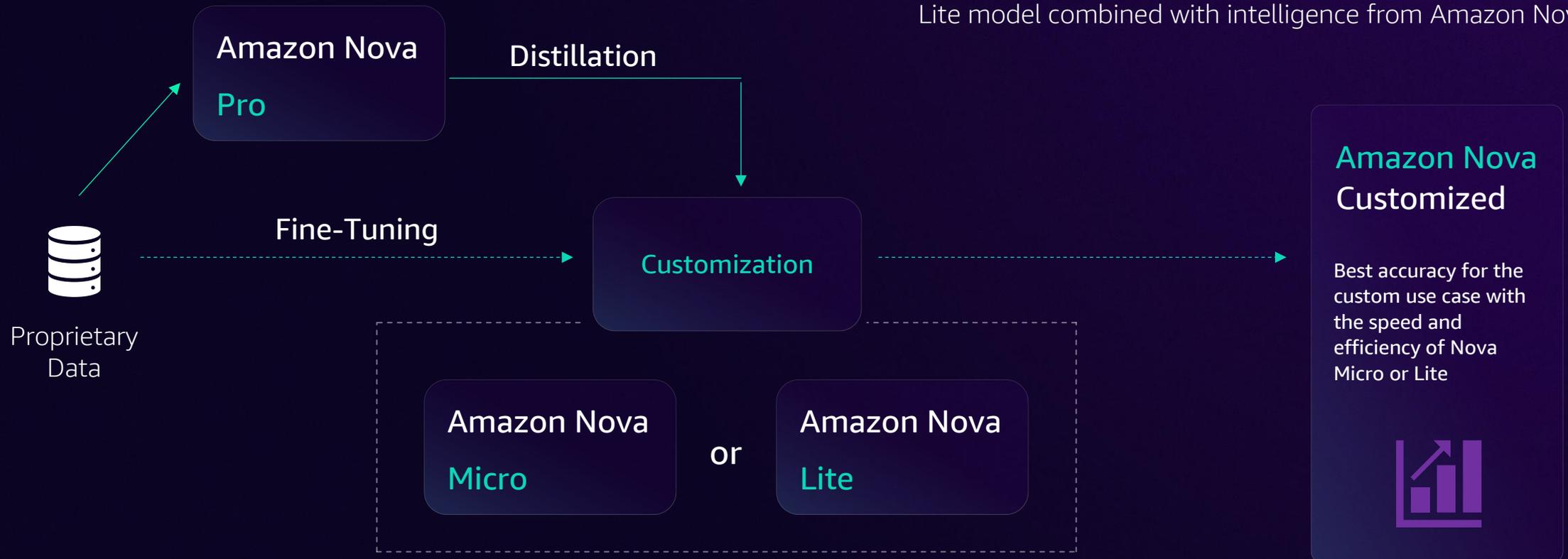
Customizability with Nova

Distill from Nova Pro with your proprietary data



Distillation: Generate synthetic text data using Amazon Nova Pro as teacher

Better Price-Performance: Speed of Amazon Nova Micro & Lite model combined with intelligence from Amazon Nova Pro



Customizability with Nova

Use Case: AI assistant that takes action!



Customers have been fine-tuning early versions of Amazon Nova Micro and Lite

Task: Function calling use cases with 17 APIs

Better Accuracy via Fine-Tuning:

- ✓ Fine-tuned Amazon Nova Micro is matching out-of-the-box performance of GPT4o-Mini & Nova Lite
- ✓ Fine-tuned Amazon Nova Lite performing best

Best Combination of Accuracy, Speed and Cost:

- ✓ up to 30% faster than GPT4o-Mini
- ✓ Delivering Industry leading price-performance



Start today and supercharge your gen AI applications with custom Amazon Nova models!



Think Fast, Think Nova

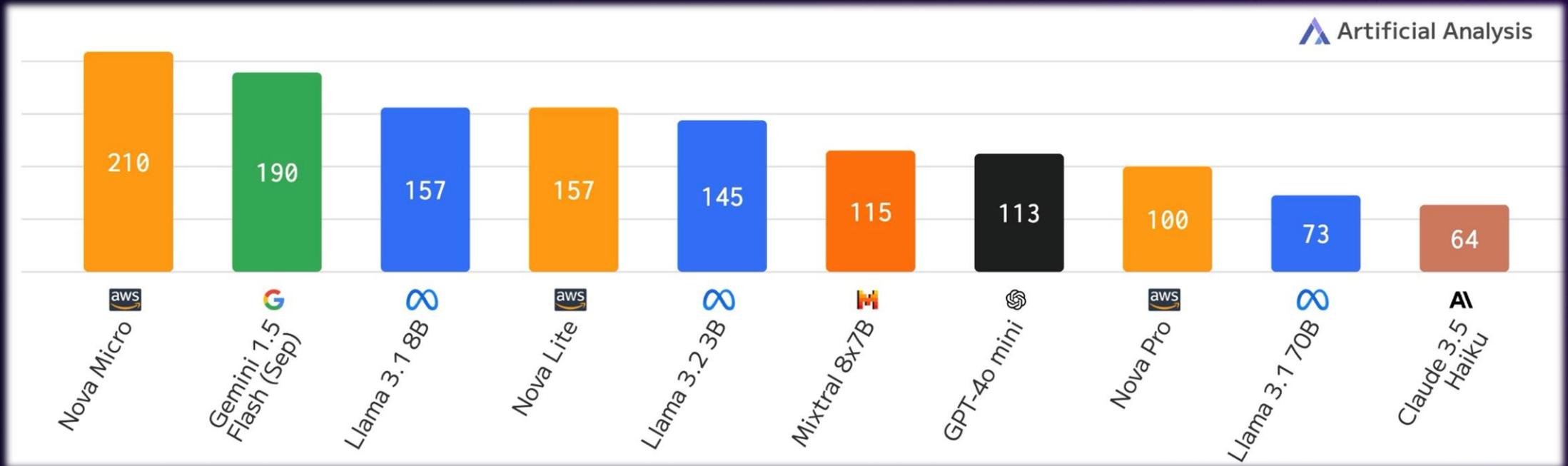
Fast inference for your use cases



Fast output generation is key for many real-world use cases

- ✓ Nova models are designed for low latency understanding applications
- ✓ Micro, Lite, and Pro are fastest models in their respective intelligence classes in Bedrock
- ✓ Coming Soon: *Latency Optimized Inference on Bedrock* for further speed up

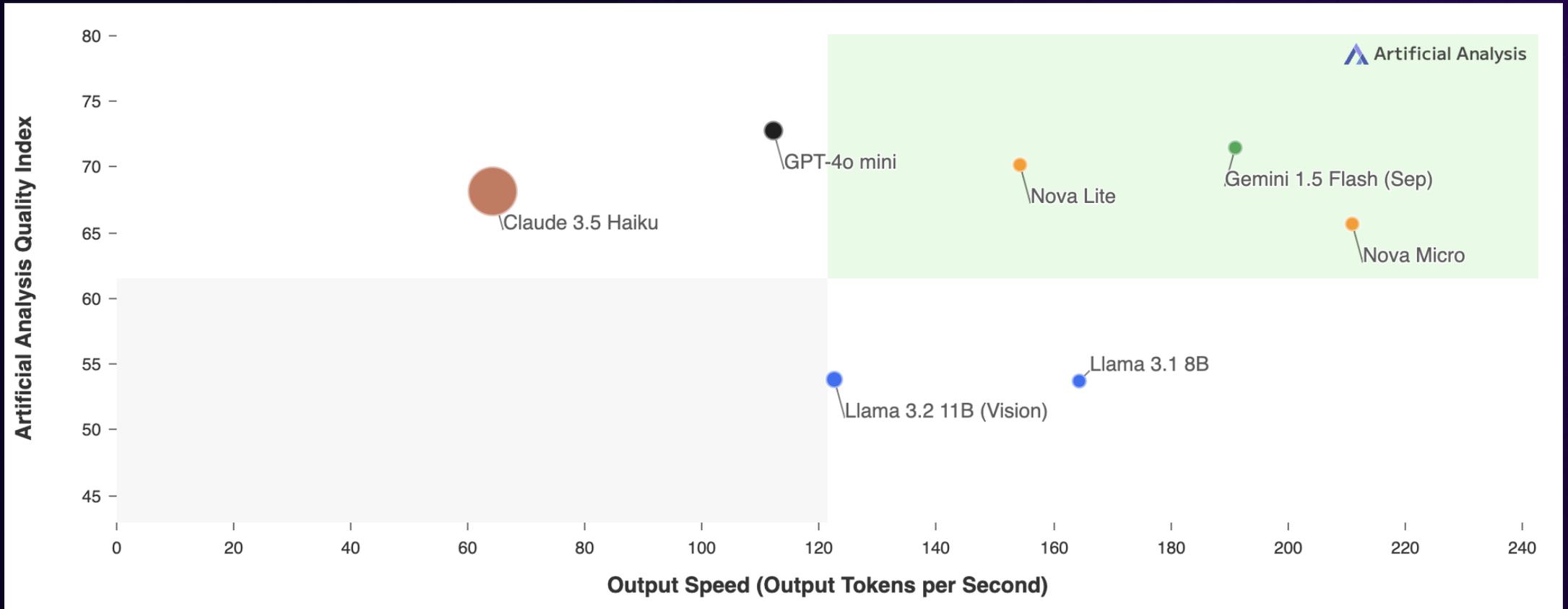
Output Tokens per Second



Think Fast, Think Nova

✓ Micro & Lite performing fastest on Amazon Bedrock

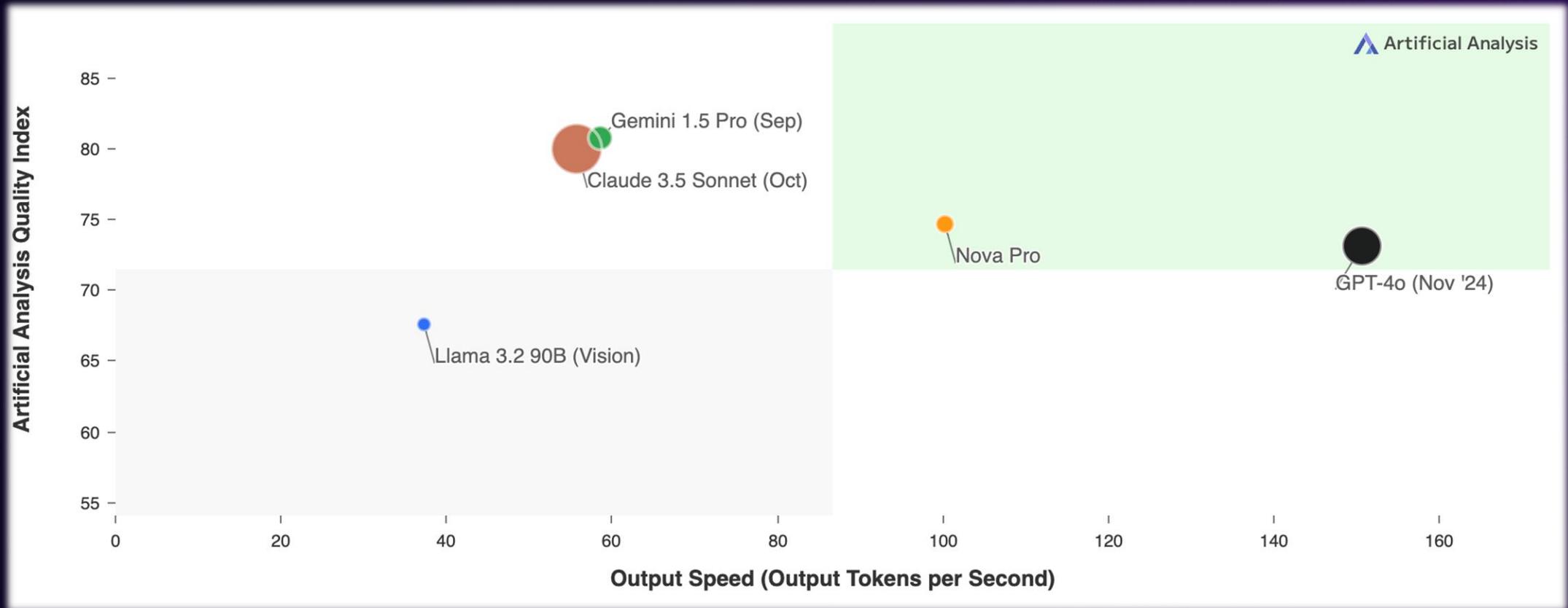
Size of circle indicating price



Think Fast, Think Nova

✓ Pro performing fastest on Amazon Bedrock

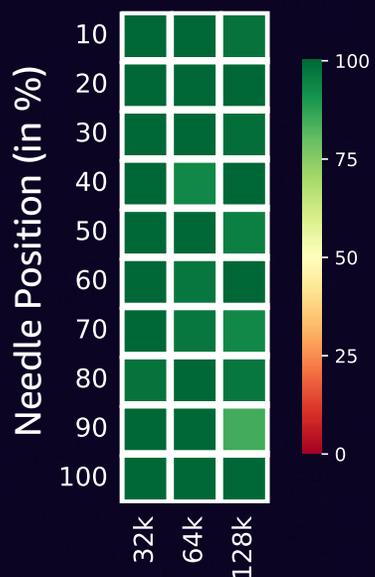
Size of circle indicating price



Amazon Nova's Context

High Accuracy across different context lengths

Amazon Nova
Micro **128k**



Needle-In-A-Haystack Accuracy at 128k

Micro	Lite	Pro
96.5%	98.4%	98.8%



Nova Long Context

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

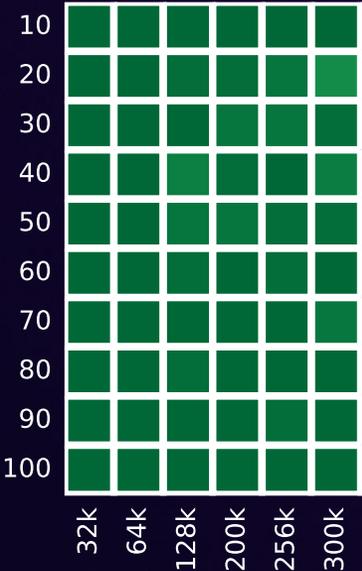
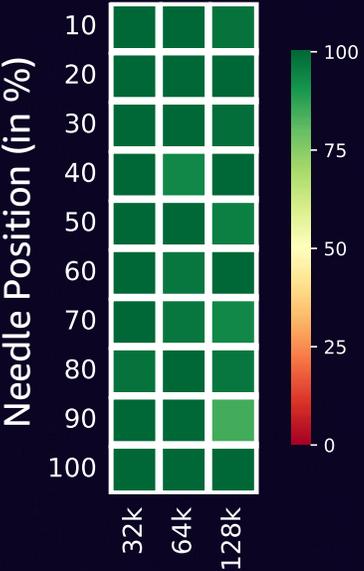
Amazon Nova's Context

High Accuracy across different context lengths

Amazon Nova
Micro **128k**

Amazon Nova
Lite **300k**

Amazon Nova
Pro **300k**



Amazon Nova's Context

High Accuracy across different context lengths

Amazon Nova
Micro **128k**

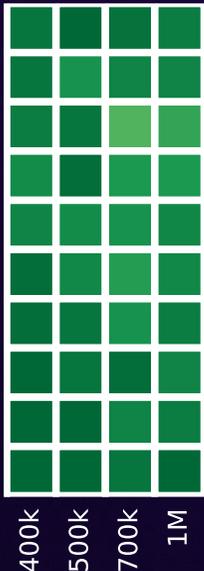
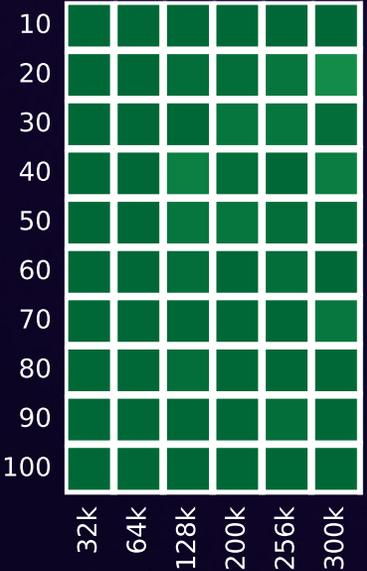
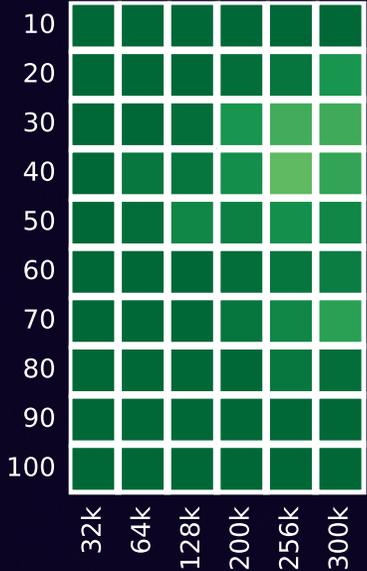
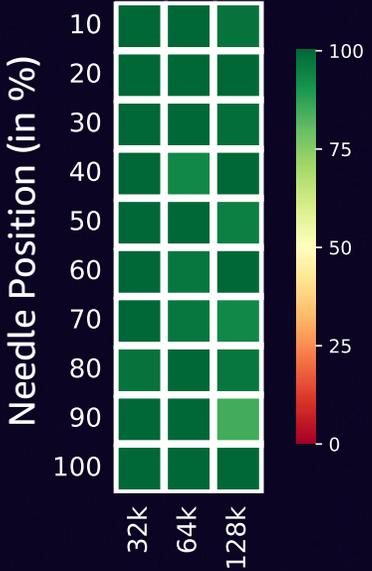
Amazon Nova
Lite **300k**

Amazon Nova
Pro **300k**



Amazon Nova
Pro **5M**

Coming Soon!



Get started today to build your long context gen AI solutions with Nova!

And stay tuned for 5 Million Token Context...



Customer Success Story

HEARST



Key Brands

- ✓ Fitch Ratings
- ✓ Cosmopolitan
- ✓ Good Housekeeping
- ✓ FDB
- ✓ MCG

\$12.8 Billion

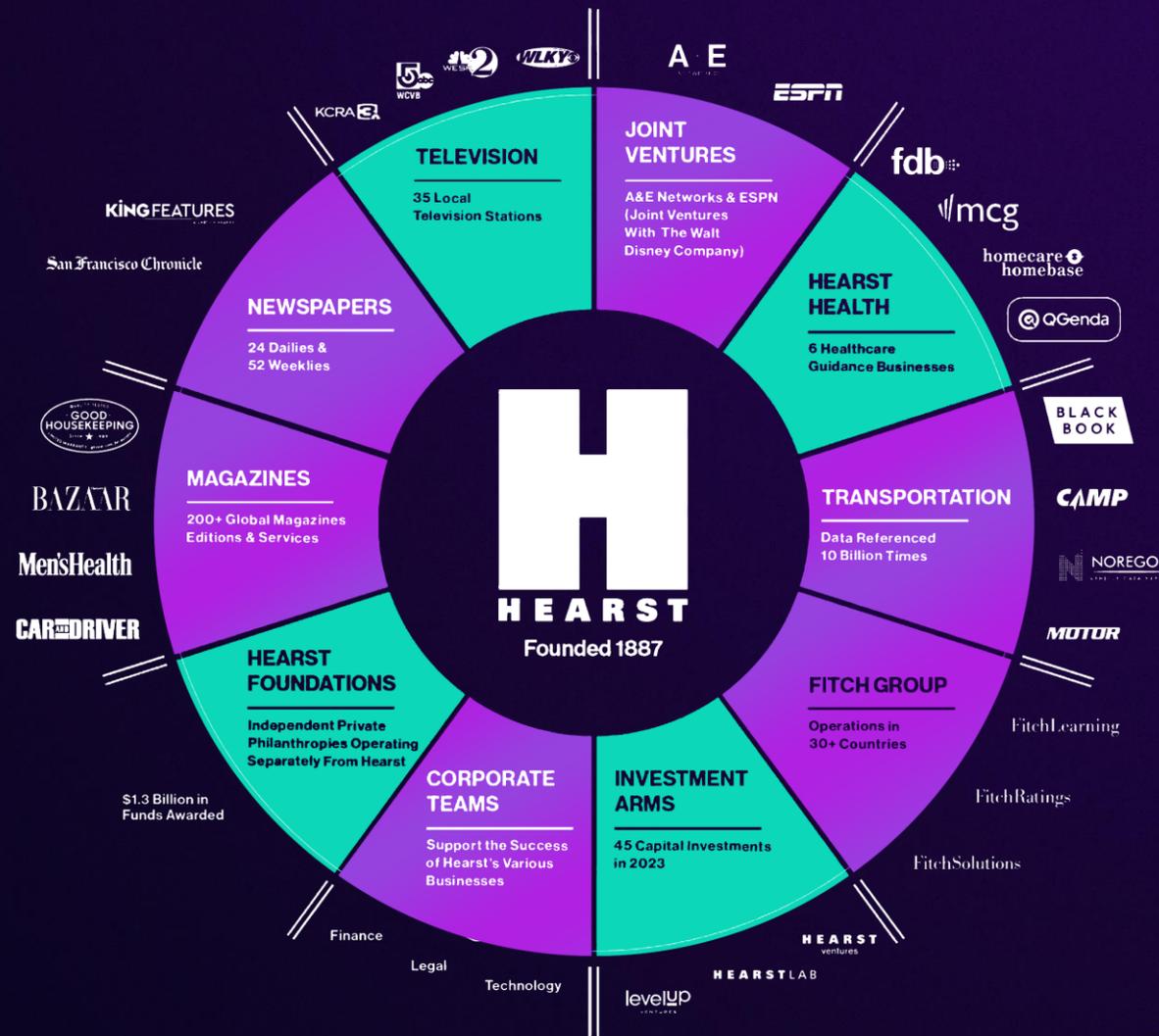
Projected 2024 Revenue

22,000+

Hearst Colleagues

6,500+

Colleagues Outside the U.S.



Generative AI at Hearst

Substantial Investments in Generative AI since late 2022



Training

Instructor-led Generative AI training for all Hearst employees

Launch

Launch of a Generative AI chat and GPT platform that allows use of all state-of-the-art models

Enhancing Productivity

Specialist development team working with businesses to enhance productivity

New Experiences

Multiple gen AI features developed and launched or launching soon to customers

Key Use Cases

FitchGroup

DOCUMENT PROCESSING

Optimize legal document processing by extracting and standardizing 150-200 vital data fields from hundreds of pages of detailed content.



WORKFLOW PROCESSING

Import health care authorizations across channels for insurance workflows and process aircraft maintenance records for verification and matching

COSMOPOLITAN

MEDIA & CONTENT ANALYSIS

Analyze video ad creative for design, messaging, and brand elements and generate insights from public meeting videos (transcripts, summaries, alerts)

Our Results

We used Amazon Nova to analyze video ads to assess likely customer response to ad placements on our media platforms. Amazon Nova was asked to summarize the ad content and to evaluate the ad across several metrics.

In our experiments the Amazon Nova

- ✓ **Generated a coherent and correct narrative from the video** Providing an accurate summary that can be used to understand the video without even watching it
- ✓ **Responded to steering prompts** Using defined personas to get different perspectives on the same video content
- ✓ **Made aesthetic and other abstract judgements** Provide feedback beyond factual narratives, that can be used for more complex analysis

Impact across Hearst Businesses



Visual Understanding at
Scale



Multitude of Opportunities
Identified Across Media,
Finance, Automotive, and
Health



\$10s of Millions in Estimated
Savings

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

Content Creation Models

- ✓ Amazon Nova Canvas
- ✓ Amazon Nova Reel

Coming soon

- ✓ Amazon Nova Premier
- ✓ 5 Million in Context Length
- ✓ Latency Optimized Inference capability

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 Nova

- ✓ [Run evaluations](#) comparing Amazon Nova in your Amazon Bedrock implementations
- ✓ Test Amazon Nova on [Bedrock Playground](#)
- ✓ Check out [Prompt Guidance](#) and [Quick Start Guides](#) for starters



Thank you!

Firat Elbey

[Linktr.ee/firatelbey](https://linktr.ee/firatelbey)

Imre Kiss

[linkedin.com/in/
imre-a-kiss](https://linkedin.com/in/imre-a-kiss)

Fabian Triefenbach

[linkedin.com/in/
triefen](https://linkedin.com/in/triefen)

Peter Goldstein

peter.goldstein@hearst.com
[linkedin.com/in/
petergoldstein](https://linkedin.com/in/petergoldstein)



Please complete the session survey in the mobile app