Generative AI in Action: From prototype to production

Dragoș Mădărășan

Solutions Architecture Team Lead,
AWS

SUPPLY CHAIN OPTIMIZATION

CUSTOMER SUPPORT

DEMAND FORECASTING

DEMAND FORECASTING

GENERATIVE AI

PILOT PROJECT

Proof of concept

PRODUCT DEVELOPMENT

PRODUCT DEVELOPMENT

Hundreds of thousands
of customers
experiment and explore

VIRTUAL ASSISTANTS

PILOT PROJECT

AUTOMATED REPORT GENERATION

SALES CONTENT

PERSONALIZATION

ACT 1

Proliferation of prototypes

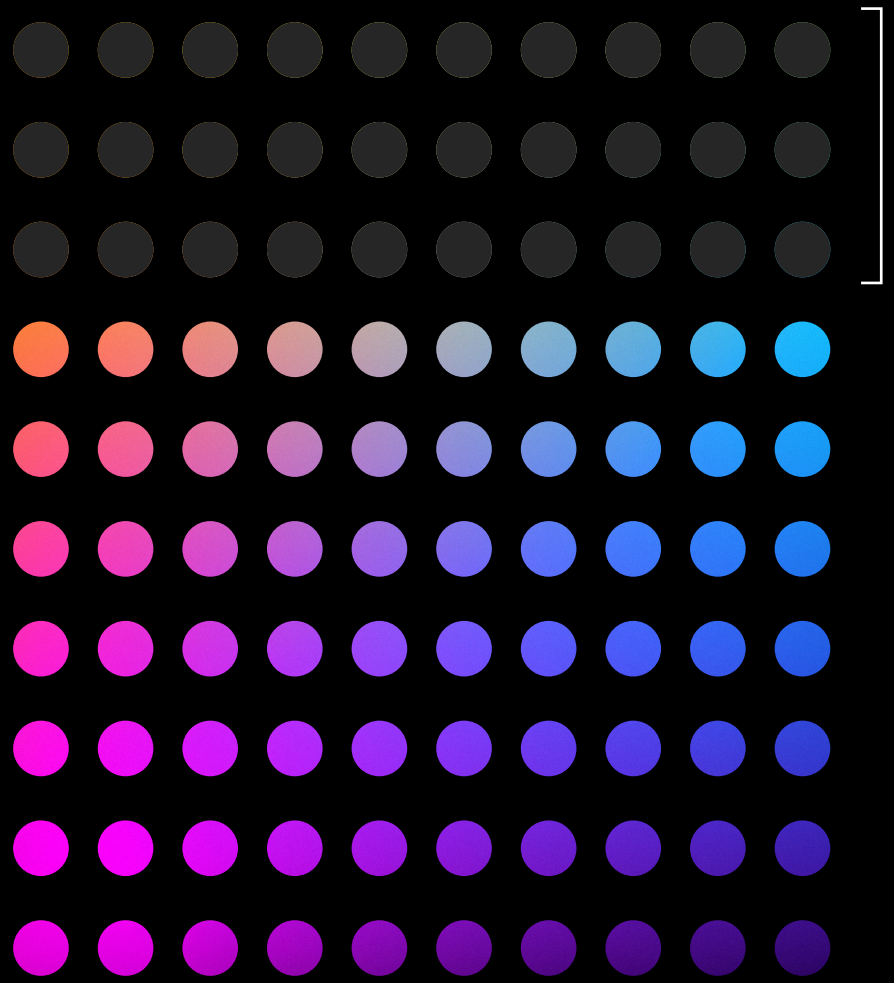
Generative AI

ACT 1

Proliferation of prototypes

ACT 2

Acceleration of production



Gartner® predicts

30%

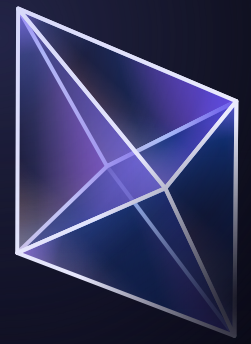
of generative AI projects will be abandoned after proof of concept by the end of 2025

Four key drivers for success

SUMMARIZATION

FAST

LOW COST



REASONING

No one model to rule them all

CODE
INTERPRETATION

INTELLIGENT

ANALYSIS

Amazon SageMaker AI

Build, train, deploy machine learning models for any use case with fully managed infrastructure, tools, and workflows

aws

NEW

Amazon Nova

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

Amazon Bedrock

The easiest way to build and scale generative AI applications

Amazon Bedrock

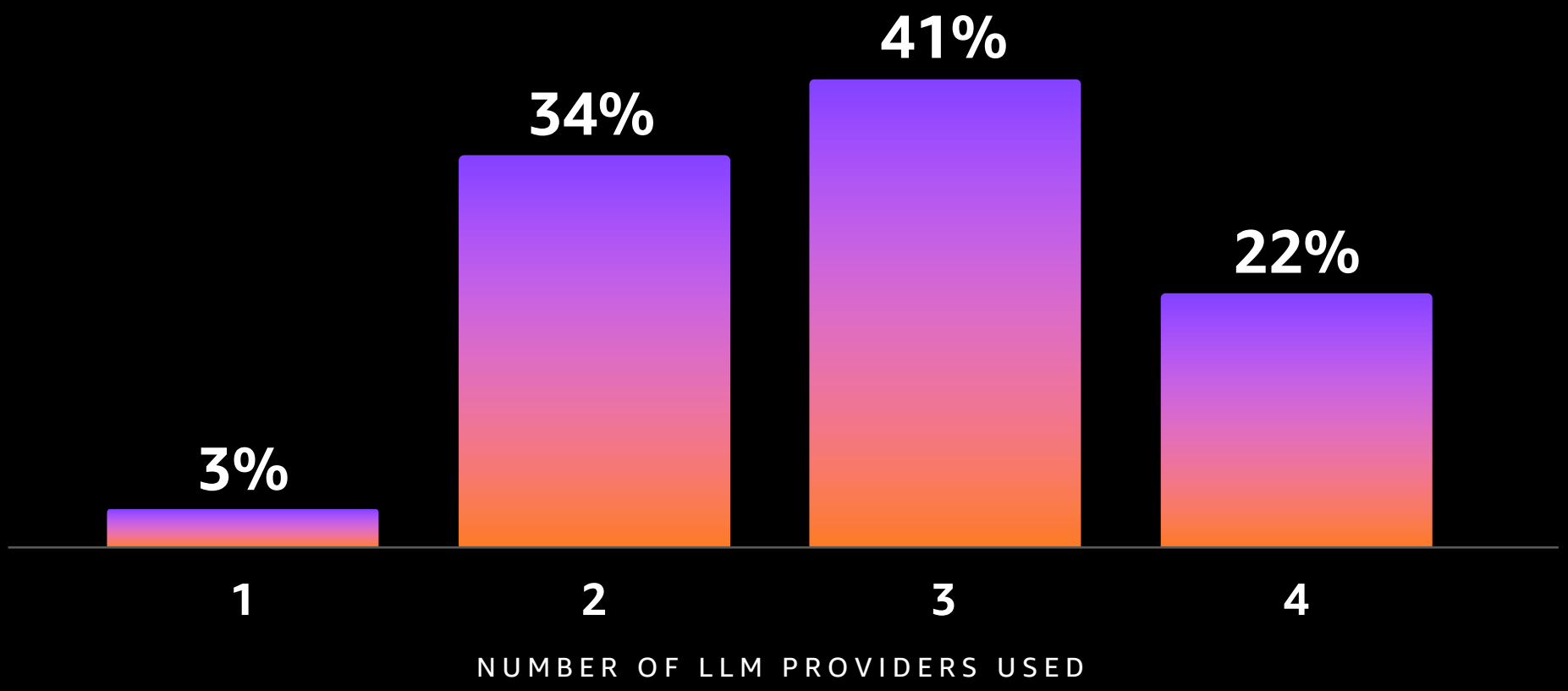
Broadest selection of models

CUSTOM MODEL IMPORT

Leverage your customized models on Amazon Bedrock

AI21 labs	amazon	ANTHROPIC	cohere	Luma	Meta	MISTRAL AI_	poolside	stability.ai
Effective reasoning & rapid analysis for long context windows	Frontier multimodal intelligence at low-latency, Agent & RAG Applications, high-quality image & video generation	Advanced reasoning & coding capabilities, including computer use skills	Multimodal search & advanced retrieval powering multilingual knowledge agents	High-quality video generation from text & images	Advanced image & language reasoning	Knowledge summarization expert agents, & code completion	Software engineering AI for large enterprises	High-quality AI image generation, easily deployable at scale
JAMBA	**AMAZON NOVA**	**CLAUDE**	**COMMAND EMBED RERANK**	**LUMA RAY 2**	**LLAMA**	**MISTRAL MIXTRAL**	**MALIBU POINT**	**STABLE DIFFUSION STABLE IMAGE**

Enterprises are deploying models from multiple model providers

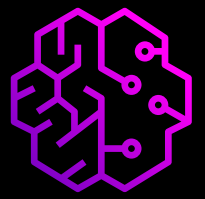


NEW

Amazon Bedrock Marketplace

Discover, test, and use over 100 popular, emerging, and specialized models in Amazon Bedrock

Cost



Amazon SageMaker HyperPod

Reduce time to train foundation models by up to 40% and scale across more than a thousand AI accelerators efficiently

Inference **cost challenges**

NEW

Amazon Bedrock supports prompt caching

Cache repetitive context in prompts across multiple API calls

AVAILABLE IN PREVIEW

NEW

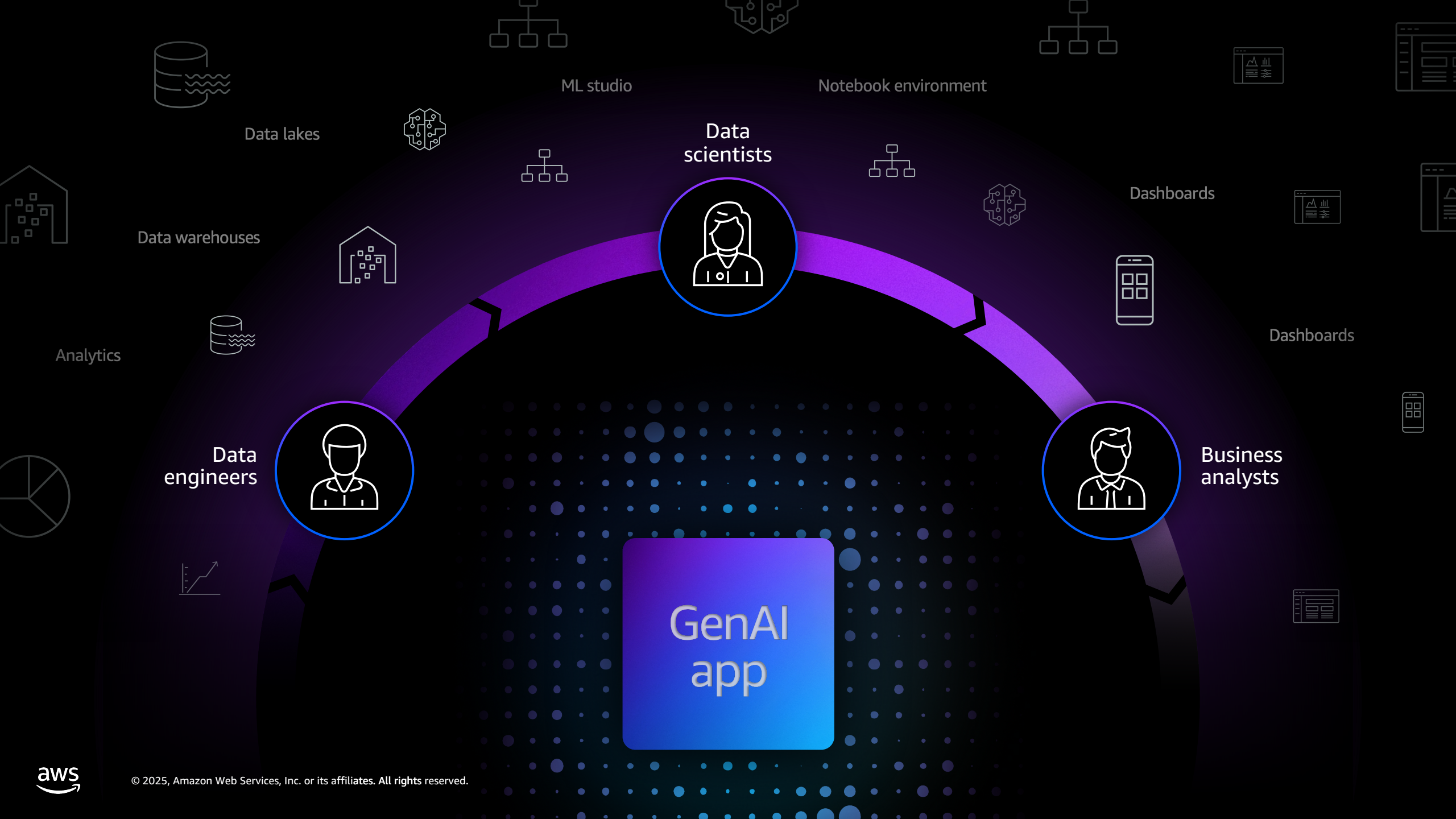
Amazon Bedrock Intelligent Prompt Routing

Automatically route prompts to models to optimize response quality and lower costs

AVAILABLE IN PREVIEW

Amazon Bedrock Knowledge Bases

Fully managed support for end-to-end RAG workflow



ISO42001

Amazon Bedrock

Amazon Q for Business

Amazon Textract

Amazon Transcribe

Accredited Certification

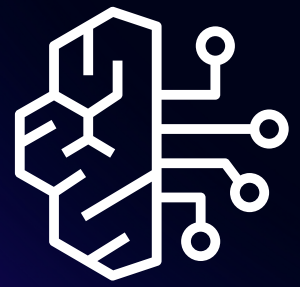
Up to

85%

More harmful content blocked with our safety features

75%

Hallucinated responses filtered



Amazon Bedrock Guardrails

NEW

Amazon Bedrock now supports multi-agent collaboration

Build powerful multi-agent systems—
faster, easier, and more cost-effectively

AVAILABLE IN PREVIEW

Four key drivers for Act 2

The best place to build your generative AI applications

Thank you!

Dragoș Mădărășan