Understanding the value of AI infrastructure

How industries, from healthcare and life sciences to media and entertainment, solve key business challenges by launching generative AI into production



Table of contents

Introduction 3

Use case 1: Healthcare and life sciences 4

Use case 2: Financial services 6

Use case 3: Media and entertainment 9

Use case 4: AI model training 11

Key questions 14

Next steps 15

INTRODUCTION

Equip your generative AI infrastructure to solve business challenges

Across industries, artificial intelligence (AI) is already improving productivity, bolstering the customer experience, optimizing data in new ways, writing code, and more. Businesses can rely on generative AI (gen AI) to empower employees by providing a more useful, personal, and engaging experience across every application, securely and resiliently—so they can complete work faster, automate mundane tasks, and be more innovative.

To realize this future, organizations need more than a powerful large language model (LLM) or chat assistant. They need a full range of capabilities to build and scale gen AI applications that are tailored to their businesses and use cases—including applications with built-in gen AI, tools to rapidly experiment and build gen AI applications, a cost-effective and performant infrastructure, and security controls and guardrails.

The use cases for gen AI underpin opportunities for organizations in a wide range of industries to innovate new products, services, experiences, and efficiencies. Across this eBook, we will explore some common gen AI use cases by industry.

$26B

expected spending in the Asia-Pacific region on gen AI in 2027[1]

95.4%

projected CAGR of gen AI spending in the Asia-Pacific region, 2022–2027[1]

Continue reading to explore:

- How to solve **key business challenges** with gen AI
- How leading **gen AI trends** may affect your deployments
- The benefits across **4 critical use cases**: healthcare and life sciences, financial services, media and entertainment, and AI model training
- **Business results** from AWS customers after deploying gen AI applications
- Real-world guidance on **evaluating your infrastructure** and exploring AWS infrastructure solutions
- Practical approaches to challenges you may face as your team moves from **model concept to production**



[1] "IDC: Generative AI Spending to Reach $26 Billion by 2027," IDC Press Release, April 2024

USE CASE 1: HEALTHCARE AND LIFE SCIENCES

Improve patient experiences and accelerate healthcare and life sciences innovations

Gen AI is transforming healthcare by unlocking high-quality data and insights for medical device companies, pharmaceutical providers, and medical centers, leading to faster discoveries and better clinical outcomes. In life sciences, gen AI is improving the speed and quality of drug discovery pipelines, which can help enhance patient outcomes and save lives.

Beyond simply meeting the compute, networking, and storage requirements of gen AI models, organizations in healthcare and life sciences (HCLS) must prioritize the protection of patient data. Amazon Web Services (AWS) infrastructure solutions abide by security and standards compliance, such as HIPAA and HITECH, to help ensure that patient data stays private and protected.

AI is used for:

- Virtual screening
- Biomolecular design
- Research and discovery
- Commercialization
- Clinical development
- Manufacturing and supply chain





PRIMARY USE CASES

Accelerating drug discovery and clinical development

Drug discovery is a lengthy and costly endeavor, typically taking many years and involving significant financial investment, with success rates generally being quite low. Gen AI has the potential to make notable progress by leveraging diverse data sources, simulations, and AI-powered insights to identify promising drug candidates and optimize their properties, automate research narratives, and enable efficient knowledge sharing throughout the development lifecycle.

Optimizing clinical workflows and improving patient engagement

Gen AI can help reduce provider burnout and improve patient experiences by automating and accelerating administrative and Electronic Health Record (EHR) tasks. This frees providers to increase their face time with patients—which today, makes up a minimal portion of the average clinician's day.

Driving operational efficiency and cost reduction

Gen AI streamlines regulated content development for providers and pharmaceutical companies by allowing workers to generate compliant first drafts through natural language and automation. Synthetic data generated by AI can accelerate research and even improve the scale and accuracy of AI models themselves—unlocking a host of new and potentially lifesaving solutions.

USE CASE 2: FINANCIAL SERVICES

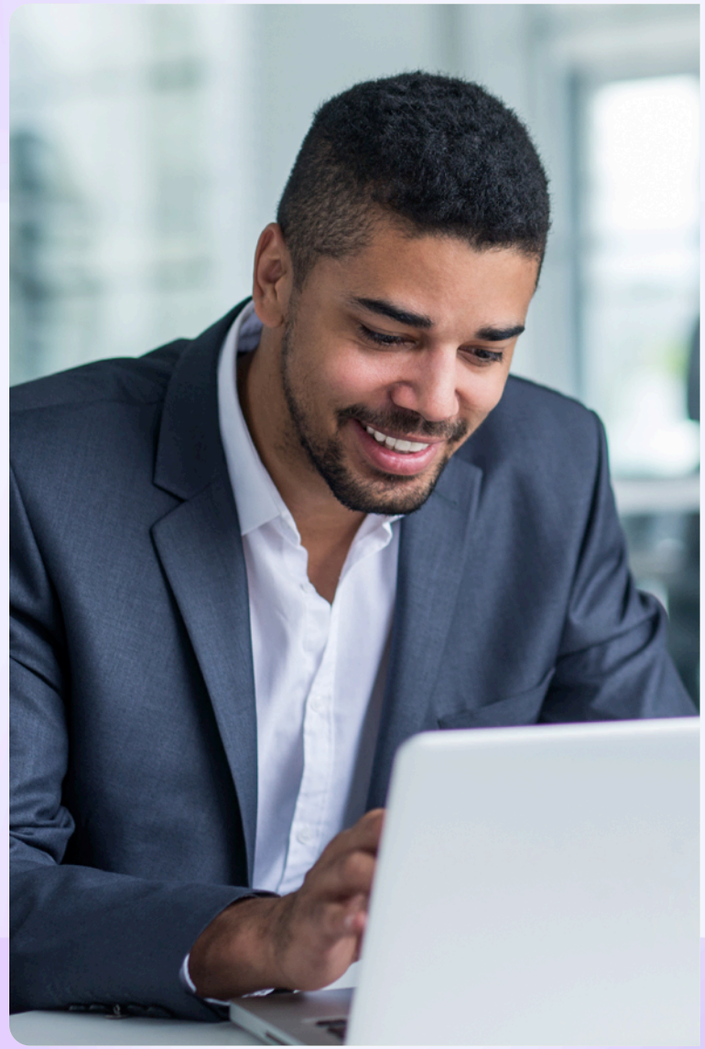
Reshaping banking, capital markets, insurance, payments, and more

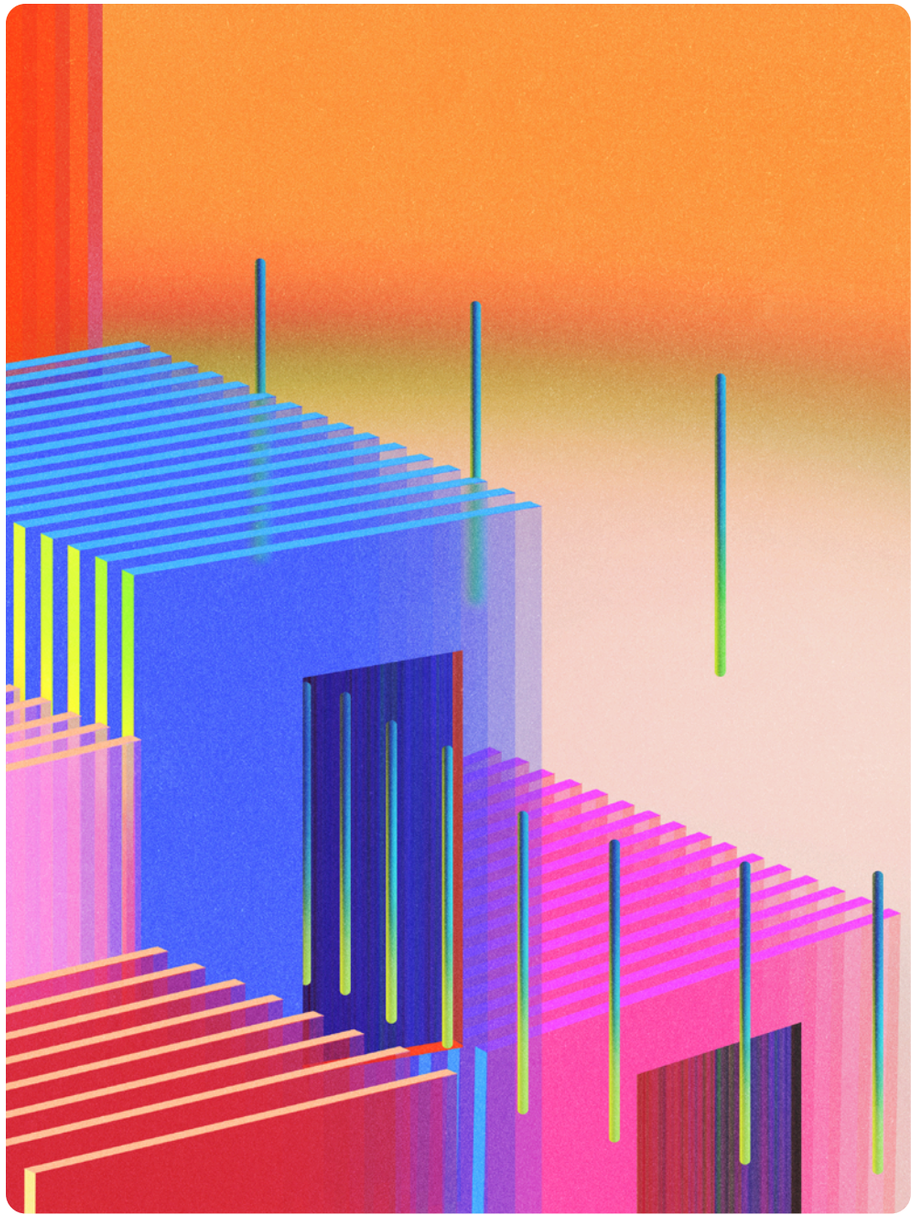
From core systems to client-facing services, gen AI is helping financial institutions increase efficiency, enhance analytics, and automate and personalize experiences. It is proving particularly beneficial for financial institutions in areas such as marketing content creation, customer service chatbots, summarization of text sources (such as analyst reports, financial news, and filings), scenario generation for risk management, and synthetic data creation for fraud detection modeling.

As financial providers embrace the transformative use cases identified here and beyond, they do so with a growing understanding of the compute, data, and storage resources required to make it all work. To ensure the building, running, and maintaining of AI systems is high-performing and cost-efficient, they need a broad set of accelerators to meet these demands.

AI is used for:

- Enterprise search
- Customer service
- Fraud detection and prevention
- Regulatory compliance





PRIMARY USE CASES

Generative recommender systems

Gen AI can analyze vast amounts of data from multiple sources to power personalized banking experiences, marketing optimization, and investment guidance.

Virtual assistants

Fintech startups and large international banks are expanding their use of gen AI to develop sophisticated virtual assistants that serve internal and external stakeholders, create hyper-personalized customer content, automate document summarization to reduce manual work, and analyze terabytes of public and private data to generate investment insights.

Data residency

Much of the data being managed by financial services companies is closely regulated by government agencies or the industries themselves. These regulatory policies have implications for the training, continued pre-training, and fine-tuning of models with data residency requirements.

crypto.com

Unique challenge

Crypto.com is a cryptocurrency exchange and comprehensive trading platform serving 100 million users in 90 countries. To provide its customers with market insights and help them make informed investment decisions, the company initially deployed off-the-shelf machine learning (ML) models alongside its own custom models for sentiment analysis. Unfortunately, these models struggled with multilingual accuracy and were expensive to host. Crypto.com turned to AWS, seeking a more effective solution to integrate and synthesize outputs from multiple ML models—both pre-trained and custom—to provide accurate, reliable, and comprehensive insights into the crypto market.

Solution overview

Within one month, Crypto.com implemented Anthropic Claude 3 Haiku models on Amazon Bedrock for sentiment analysis, analyzing crypto news in more than 25 languages. Crypto.com also used Amazon SageMaker AI to fine-tune custom models and quickly adapt to market changes.

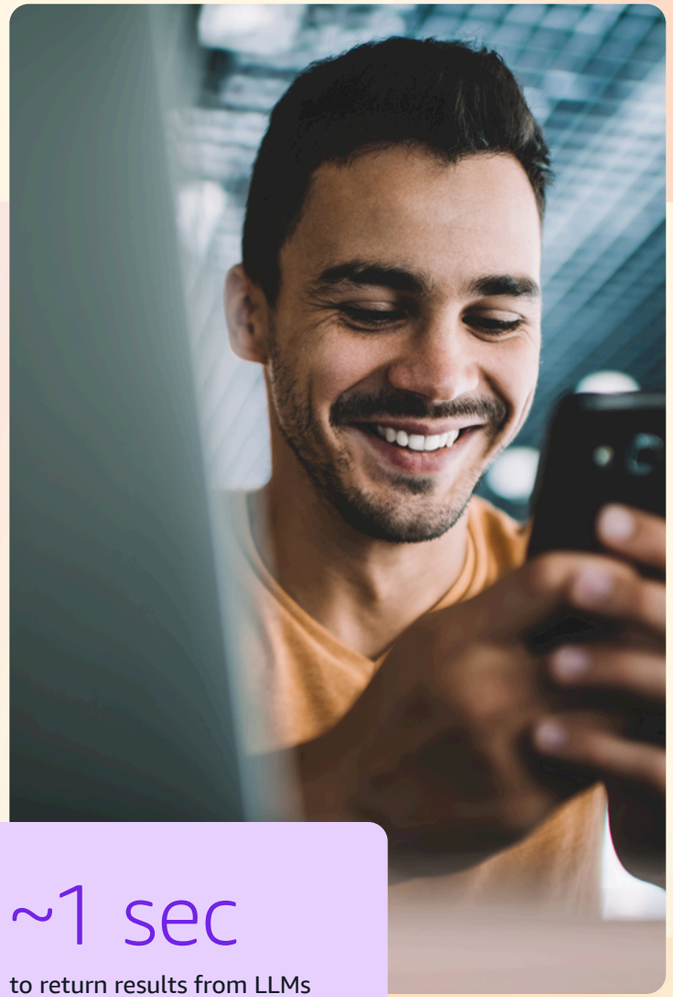
Use case in practice

Crypto.com found that LLMs on Amazon Bedrock returned results very quickly, typically within one second. By using Amazon SageMaker to customize models, such as Mistral AI and Meta Llama, with its own data, the company achieved high-quality analysis for established and emerging cryptocurrencies—and delivered intelligent, actionable, and timely financial insights to customers.

Real results

Leveraging gen AI on AWS, Crypto.com can now efficiently deliver accurate, comprehensive, and localized crypto market insights to its global user base. The integration of gen AI tools has improved model accuracy and helped automate insight generation, saving time and resources. Plus, AWS makes it easier for Crypto.com to test new models for gen AI use cases, which can help its users make informed and well-planned investment decisions.

Learn more about Crypto.com and AWS →



~1 sec

to return results from LLMs with gen AI on AWS

USE CASE 3: MEDIA AND ENTERTAINMENT

Delighting audiences and creating better content faster

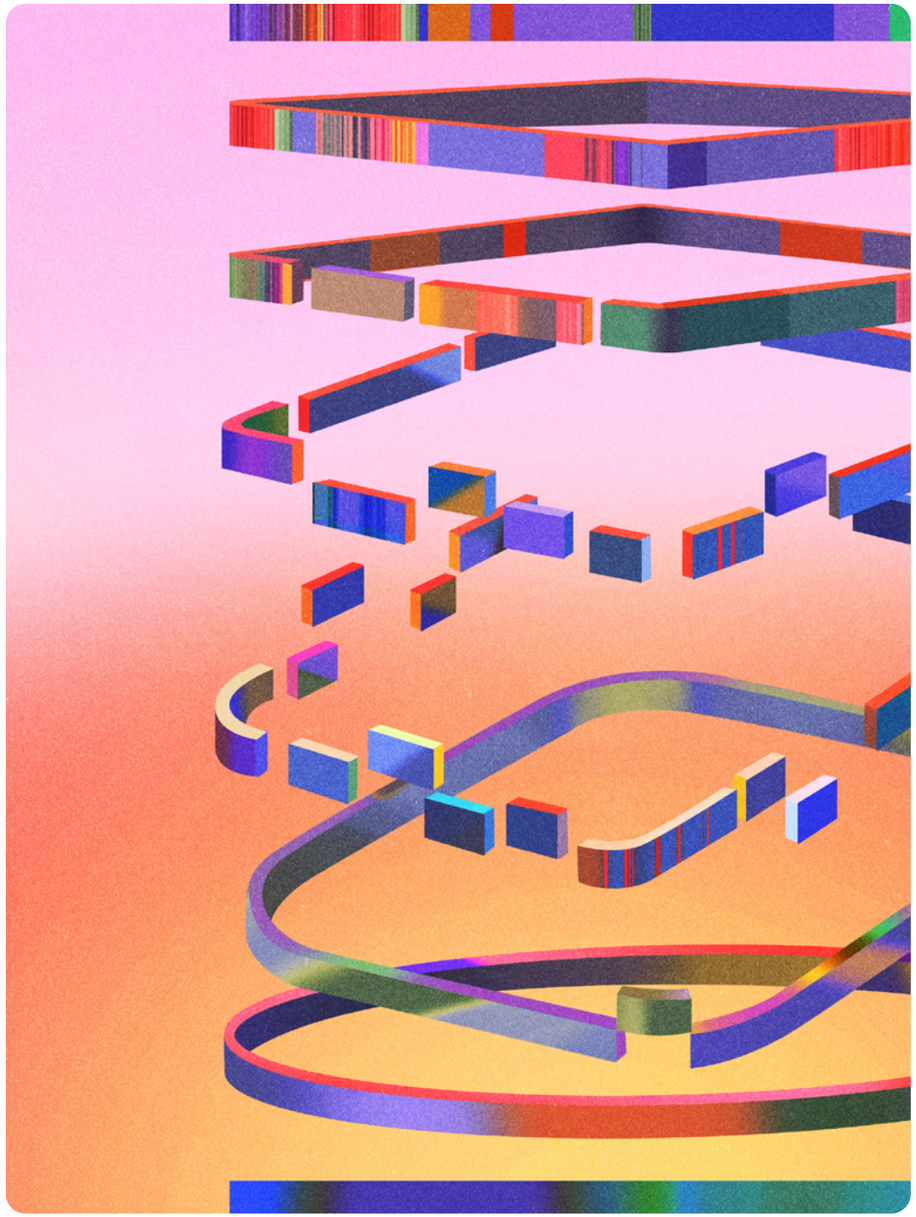
In a market where audiences have more choices than ever, gen AI enables media and entertainment (M&E) companies to deliver unique content and experiences that help set them apart. From video and music streaming to internet advertising and sports, gen AI is accelerating core processes for rapid content creation and curation, enhanced consumer engagement, and improved monetization options for M&E companies.

Across the industry, gen AI is unlocking productivity gains, maximizing audience engagement, and driving innovation. To achieve this, studios and broadcasters are training, customizing, and deploying GPU-accelerated gen AI models on premises and in the cloud.

AI is used for:

- Personalized recommendations
- Generation of 2D and 3D environments
- Curated advertising and content
- Video dubbing and video infilling
- Dialogue and voice generation





PRIMARY USE CASES

Content creation

Gen AI allows M&E companies to create text, images, videos, sound effects, music and voice-overs, 3D objects, and animations based on user prompts. This newly created content can be used for artistic exploration, data augmentation, style transfer, video synthesis and enhancement, animations, and more.

Data analytics and personalization

With gen AI, M&E companies can analyze vast amounts of data and tailor content to individual preferences and behaviors in real time. Through highly focused targeting and dynamically generated advertising, gen AI also improves M&E marketing efforts, driving greater revenue from enhanced audience engagement and increased subscription renewals.

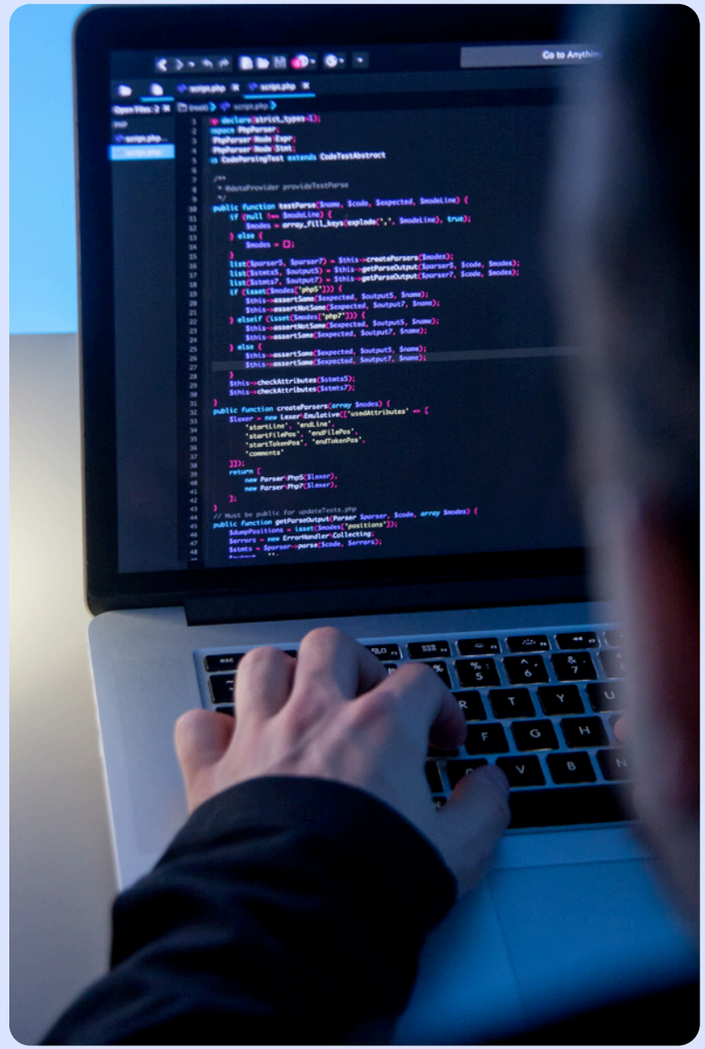
USE CASE 4: AI MODEL TRAINING

Build, train, and deploy your own models

Powered by foundation models (FMs) trained on data, gen AI can create new content and ideas, including conversations, stories, images, videos, music, and even software code, in response to a prompt. As these models continue to grow and adapt to unique use cases, their infrastructure needs change and evolve as well.

Businesses that are building models for gen AI need high-throughput, low-latency, and secure infrastructure to train and deploy their models in a reasonable time for inference—all while trying to continue to lower their costs and maintain the highest performance possible.

Businesses that are training and deploying models need a broad choice of cloud capabilities and specialized infrastructure that deliver performance while reducing costs across a secure and private experience.



NinjaTech AI

Unique challenge

In pursuit of its mission to enhance productivity by handling complex tasks with fast and affordable AI agents, NinjaTech AI began work on a multi-agent personal AI assistant: myninja.ai. But developing the solution required multiple specialized LLMs to work together seamlessly, and a scalable, cost-efficient solution for AI model training was critical.

Solution overview

NinjaTech AI used instances powered by AWS Trainium chips to build its cutting-edge productivity agent, NinjaLLM, which serves as the backbone of myninja.ai. A combination of powerful AWS instances and AWS ParallelCluster enabled the efficient parallel training needed to fine-tune models quickly and affordably, with each iteration completed in under three hours at a cost of less than $1,000. This rapid and cost-effective approach allowed NinjaTech AI to refine its models and integrate new updates swiftly.

Use case in practice

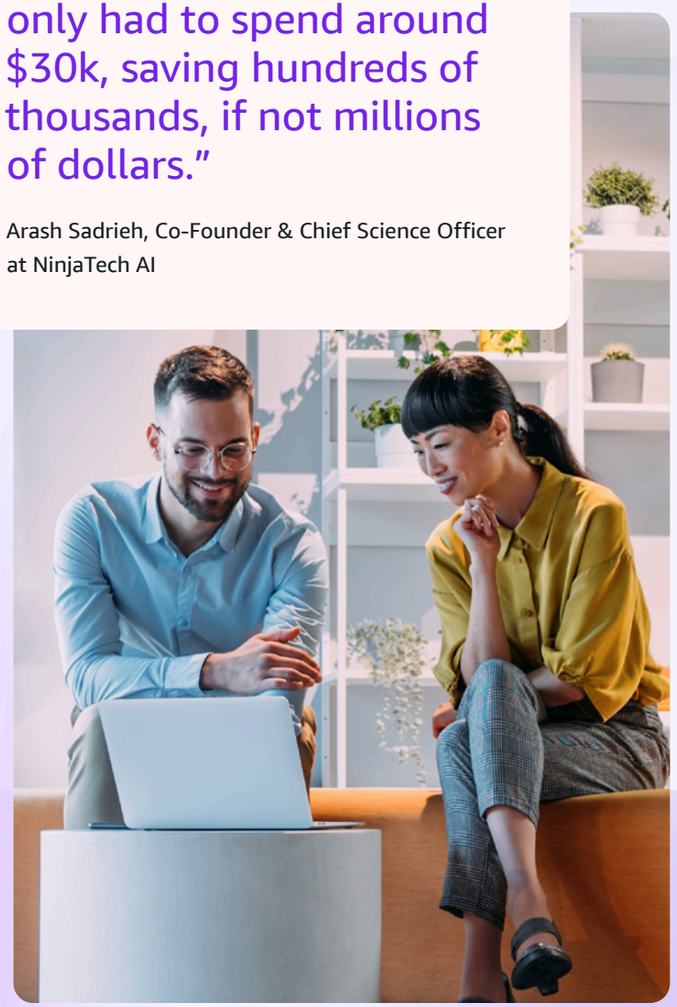
Myninja.ai can complete standard smart assistant tasks, such as scheduling meetings, conducting web research, generating code, and writing text. But the solution diverges from the pack with its ability to break down intricate multistep tasks, evaluate its own efforts dynamically, and continually learn and improve—all with minimal user input.

Real results

AWS solutions enabled NinjaTech AI to iterate quickly and enhance its AI models at a low cost. By the company's own estimation, using AWS instances powered by AWS Trainium chips reduced the project's costs by hundreds of thousands or more compared to traditional training accelerators.

Learn more about NinjaTech AI and AWS →

> "[By fine-tuning models with AWS Trainium,] we only had to spend around $30k, saving hundreds of thousands, if not millions of dollars."
>
> Arash Sadrieh, Co-Founder & Chief Science Officer at NinjaTech AI



perplexity

Unique challenge

Perplexity, a company building one of the world's first conversational search engines, uses gen AI to help users find relevant knowledge. Faced with the challenge of optimizing its models to provide rapid and accurate responses, which requires near real-time inference capabilities, Perplexity sought out a robust infrastructure solution.

Solution overview

With AWS, Perplexity gained greater flexibility in resource allocation by assigning the right tasks to the right instances. Amazon Elastic Compute Cloud (Amazon EC2) P4de instances deliver the highest performance for model training on Perplexity's substantial memory and bandwidth requirements. And with Amazon EC2 P5 instances, Perplexity can achieve the near real-time inference needed for rapid and accurate user query responses.

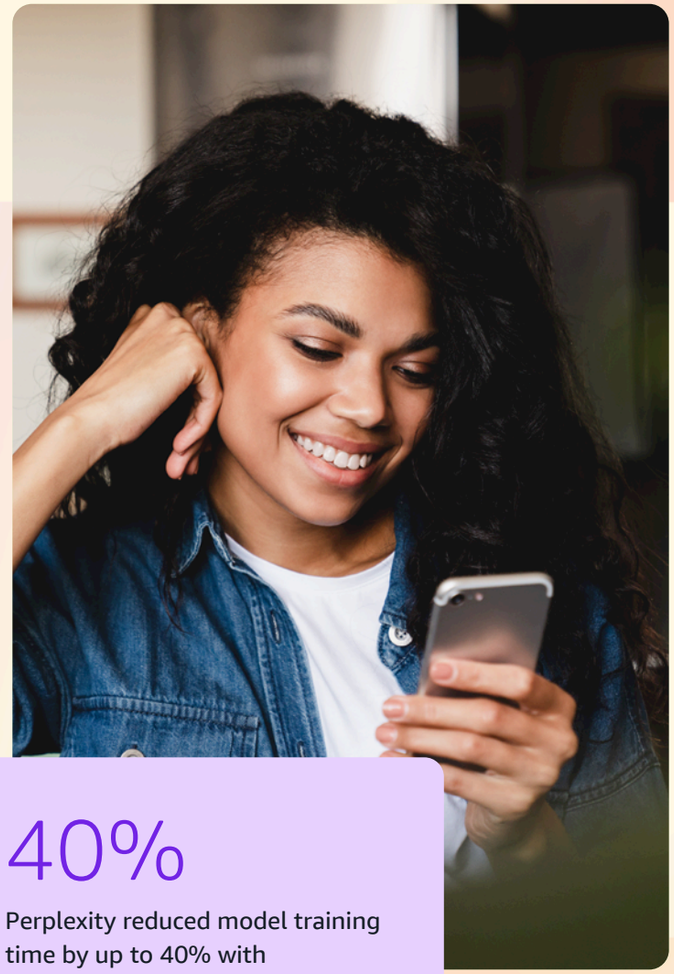
Use case in practice

Using Amazon SageMaker HyperPod, Perplexity accelerates the development of new products and customer experiences. AWS also helps enable Perplexity to handle massive spikes in concurrent users or queries without compromising latency.

Real results

Perplexity reduced model training time by up to 40 percent with Amazon SageMaker HyperPod. AWS instances allow Perplexity to manage spike periods of 10,000 concurrent users and over 100,000 queries per hour, all without affecting user experience.

Learn more about Perplexity and AWS →



40%

Perplexity reduced model training time by up to 40% with SageMaker HyperPod.

aws

KEY QUESTIONS

4 questions to ask as you embark on your generative AI project

Now that you've explored how different organizations power common gen AI use cases, you will want to decide on your best way forward. AWS experts recommend that you begin by answering four key questions intended to gauge your performance goals, cost barriers, frameworks, and management plans. From there, you will know how your infrastructure demands will need to evolve to support the use case.

1 What is your target performance?

Training is the amount of time it takes (how many computations per second) to train your model. Network performance is also important for distributed training clusters, as nodes communicate with updates to the model. For inference, this is both throughput (how many inference requests are processed per second) and round-trip latency (how quickly you can deliver results).

2 What are your cost barriers?

As model complexity increases, the cost to continually re-train your model for greater accuracy in predictions will become prohibitive. When you choose the infrastructure to train or deploy your model, the amount of underlying resources being optimally utilized also ties into your cost evaluation.

3 What type of machine learning framework support do you need?

Most organizations rely on ML frameworks, such as PyTorch and TensorFlow, so picking infrastructure and services that support your desired frameworks is essential.

4 How do you want to manage infrastructure?

As you select the right infrastructure for your gen AI project, you also need to consider management. AI infrastructure—including compute, networking, and storage—for training must be managed in a secure, cost-effective, and scalable manner. You can even use fully managed services, helping you and your team focus on what's most important.

NEXT STEPS

Propel your generative AI use cases into production

Organizations worldwide are finding new ways to drive business impact with gen AI. However, the training and deployment of gen AI models can be costly and time consuming, which is why it's a critical priority to find the right infrastructure to meet changing business needs.

AWS can help your business accelerate its gen AI projects with high-performing, cost-effective, and highly flexible infrastructure—along with easy ways to leverage solutions through fully managed services. AWS Trainium and AWS Inferentia offer the the highest-value infrastructure for your model training and inference needs. And Amazon SageMaker makes it easy to take advantage of infrastructure, tools, and workflows to provide an end-to-end solution for your AI lifecycle.

Discover why organizations across industries and geographies choose AWS to accelerate their gen AI journeys. With a broad range of infrastructure, services, and tools, the possibilities you can unlock with AWS are endless.

Start exploring AWS AI infrastructure today →



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.