WEBINAR

EMEA fintechs: unlock innovation with generative AI with AWS and NVIDIA

Key takeaways

aws | NVIDIA

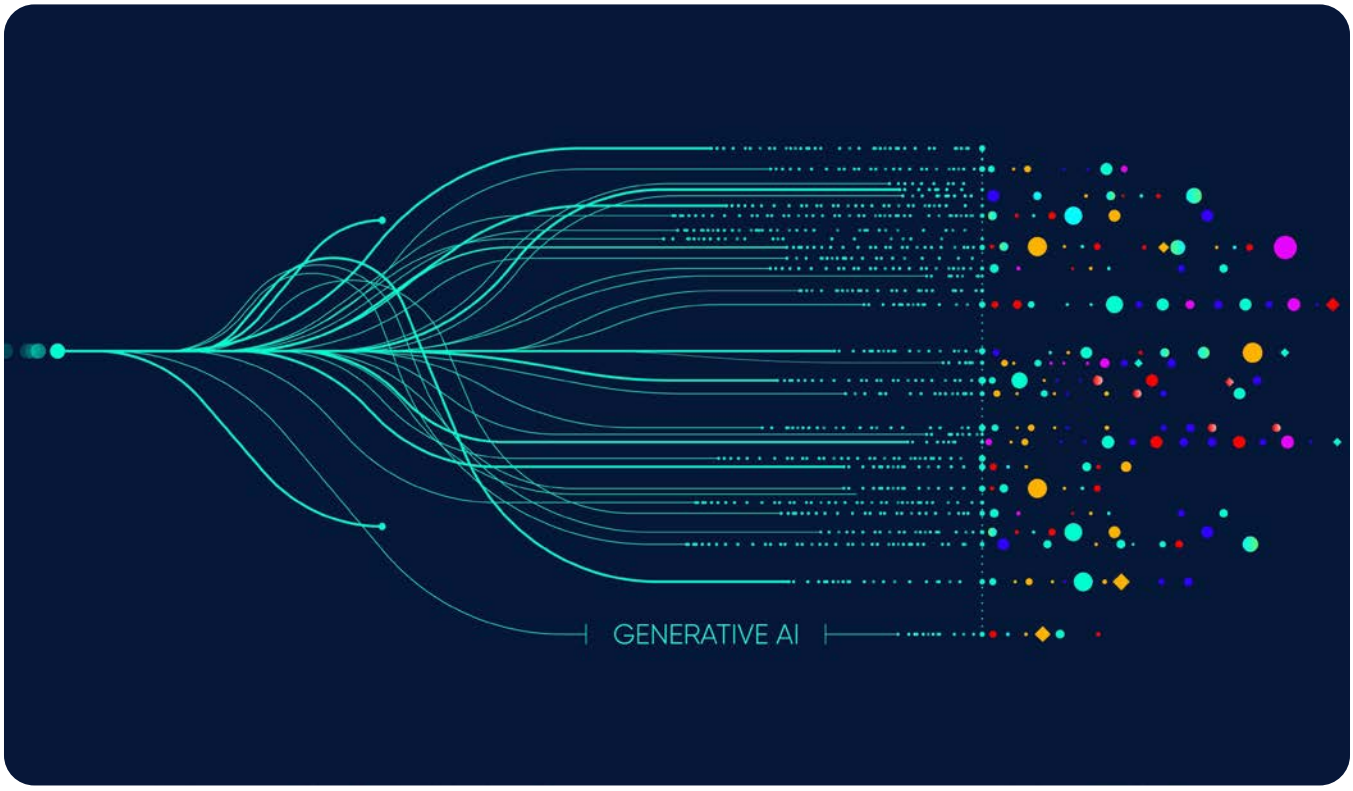
FINTECH FUTURES

Introduction

As funding and projected revenues increase, the fintech sector is rapidly evolving, largely fuelled by advancements in artificial intelligence (AI) and machine learning (ML). In the *EMEA fintechs: unlock innovation with generative AI with AWS and NVIDIA* webinar, available now on demand, industry experts from Amazon Web Services (AWS) and NVIDIA explore how generative AI enables fintechs to innovate quickly and stay one step ahead of the competition.

Discussion points:

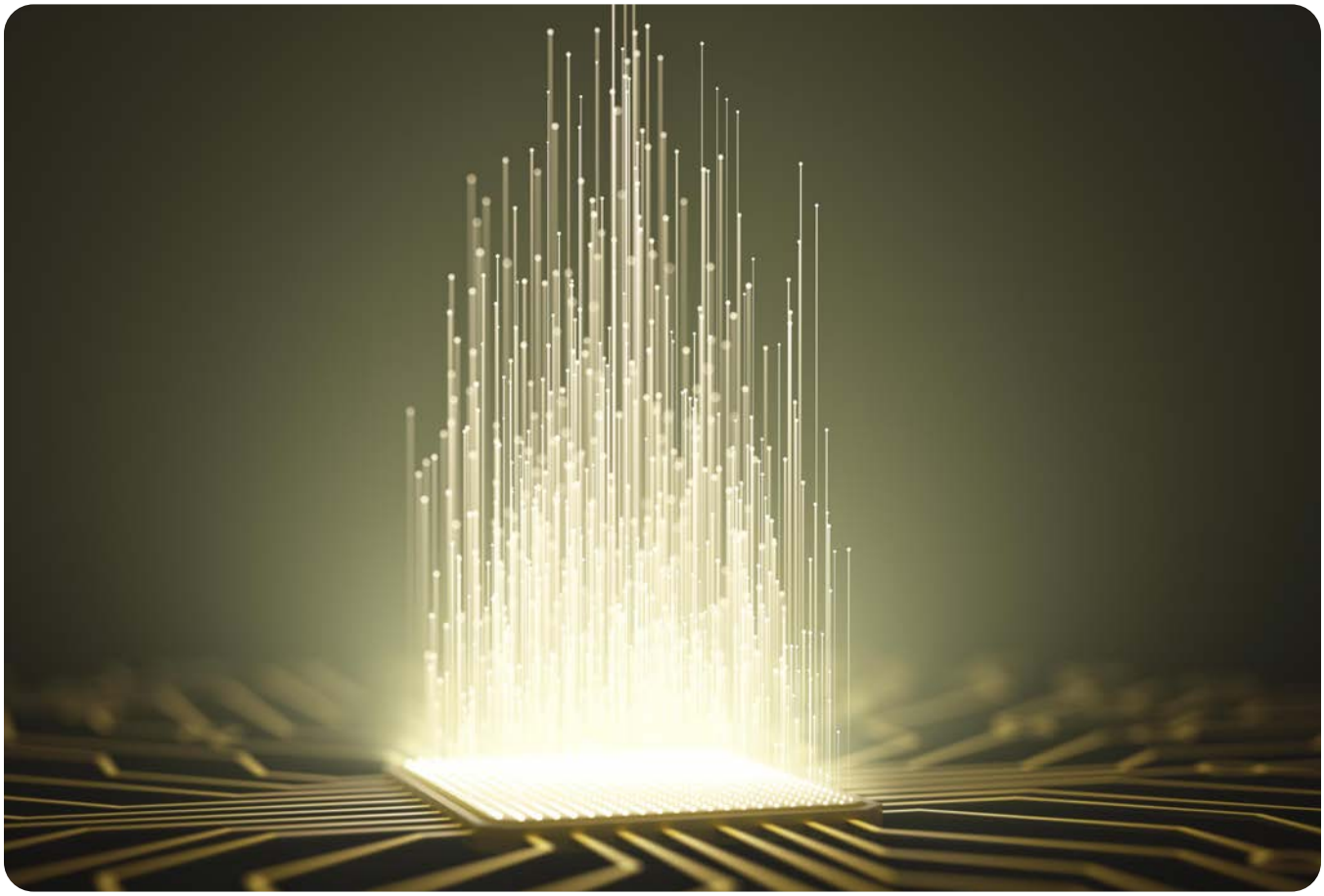
- An overview of generative AI in fintech and the segments that benefit most from AI.
- Key generative AI use cases including risk management, fraud detection, AI chatbots, personalised recommendations, algorithmic trading, and model backtesting.
- Key challenges faced such as managing complex data securely and scaling AI solutions.
- How to harness AI for smarter decision-making and enhance security with initiatives such as Open Banking which has made vast datasets accessible.
- How to bring community-built AI models into production using microservices.
- How accelerated compute from AWS and NVIDIA enables generative AI adoption.



Webinar speakers

Sam Edge, Global Head of Fintech, Amazon Web Services
Sam is the Global Head of Fintech at AWS, leading a team of ex-founders and investors to support the world's top Fintech startups and venture capital funds. Prior to AWS Sam has been a founder, investor, and consultant, with 15 years' experience in banking, capital markets, and fintech. Recently named by the UK's Evening Standard as a leading Global Fintech Influencer, Sam founded AWS' first ever Fintech accelerator, and is a mentor to startups through programmes such as Techstars, Plug & Play, and Startup bootcamp.

Dr. Jochen Papenbrock, Head of Financial Technology, EMEA, NVIDIA
Jochen is Head of Financial Technology EMEA / Lead Developer Relations Manager Banking Global at NVIDIA, the globally leading accelerated computing platform company. He has spent the last 25 years in various roles on the topic of AI in financial services. He works with executives, data scientists, developers, and partners in the global NVIDIA ecosystem. Jochen is a financial data scientist and received his degree and Ph.D. from the Karlsruhe Institute of Technology (KIT). Before NVIDIA he worked as a consultant, entrepreneur, and researcher with asset managers, banks, insurance companies, and central banks.



Watch the webinar on demand now

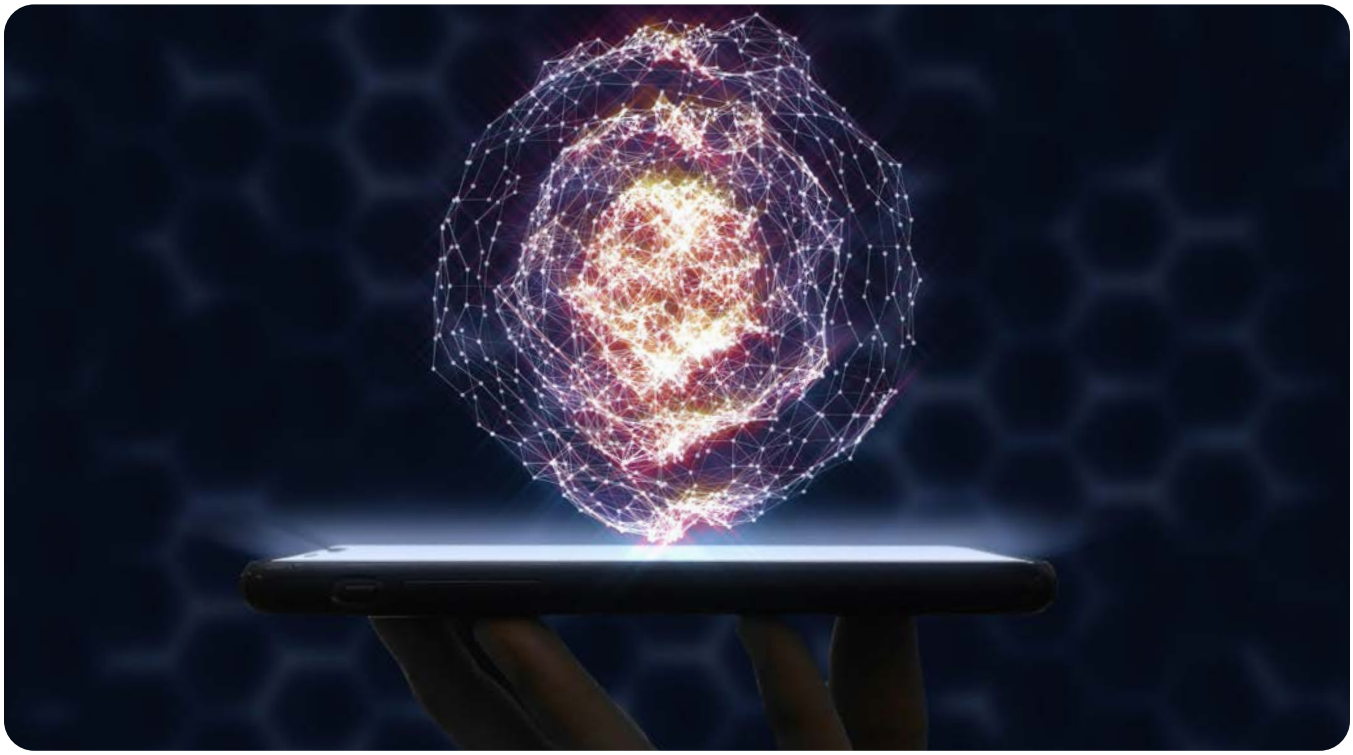
The *EMEA fintechs: unlock innovation with generative AI with AWS and NVIDIA* webinar is available to watch now on demand.

You can watch the full webinar by clicking here.

Who should watch:
- Fintech companies looking to start or optimise the use of generative AI.
- Cloud native start-ups that want to scale AI adoption.
- Financial firms who want to leverage their data and offer better customer service.

Benefits of watching:
- Learn how generative AI is being used in fintechs today.
- Discover the benefits of running generative AI in the cloud.
- Understand how AWS and NVIDIA work together, with joint customer examples.

The *EMEA fintechs: unlock innovation with generative AI with AWS and NVIDIA* webinar is brought to you by AWS and NVIDIA.



Key takeaways from the webinar

Generative AI has quickly become a major focus in the world of financial services, with businesses looking to utilise the technology to streamline processes, cut costs, and find innovative new use cases to gain a competitive advantage.

In this webinar brought to you by *FinTech Futures*, AWS and NVIDIA, industry experts Sam Edge, Global Head of Fintech at AWS, and Dr Jochen Papenbrock, Head of Financial Technology, EMEA at NVIDIA, explore how generative AI can enable fintechs to innovate quickly and drive business growth.

Here, we cover the key takeaways from the discussion, including how the technology is being used by EMEA fintechs today to power digital transformation, enhance customer experiences, and make smarter data-driven decisions. We also explore how AWS and NVIDIA are working together to drive innovation in the industry, with some examples of real-world customer stories.

The next big wave

During the webinar, Sam shared that ML and generative AI are driving the next phase of evolution in the industry, enabling a shift from manual processing and general solutions to a more automated financial services value chain that provides a much more personalised end result on behalf of the customer.

Due to generative AI's potential to transform industries, interest and excitement in the technology across financial services continues to grow, with McKinsey recently ranking banking as the number two industry globally (behind only tech) in terms of value generation from generative AI, predicting an annual potential of $200 billion to $340 billion dollars in value.

$200-$340bn

annual potential value generative AI could add to banking revenue

Source: McKinsey

"And so I think it's a very exciting time within financial services and fintech right now because of the combination of traditional AI and the new wave of generative AI," Sam says.

Fintech companies have been using traditional AI for years across the whole customer journey, with infrastructure use cases such as fraud detection and data protection; core processing use cases such as customer authentication, document processing, and analytics; and client-facing use cases such as customer experience and chatbots.

> *"It's a very exciting time within financial services and fintech right now because of the combination of traditional AI and the new wave of generative AI."*
>
> *Sam Edge, Global Head of Fintech, AWS*

Looking at generative AI in particular, Sam shared that the technology is currently being used in three key ways within fintech:

- **Customer experience:** Fintech companies are embedding generative AI into workflows around call centres and chatbots to better manage how they engage with customers
- **Productivity:** The technology can be used to boost the efficiency of knowledge workers through search, summarisation, and reporting tools
- **Innovation:** Fintechs are utilising AI to build better products and automate processes

In terms of innovation, Sam shares he is "really excited about this space in particular", because "there's been a big promise in fintech for years about the concept of self-driving money - everybody having a financial service provider in their pocket".

"I think if you look at the combination of new data sources coming through into the industry from Open Banking/Open Finance and combine that with the potential of autonomous agents leveraging generative AI, I think there's some really exciting things that can be done on the product innovation side of things," Sam says.

> *"In payments, it's about building systems that monitor these payment networks. For example, you could generate synthetic data leveraging generative AI tools, and also provide real-time fraud alerts."*
>
> *Dr. Jochen Papenbrock, Head of Financial Technology, EMEA, NVIDIA*

Jochen adds that companies across the financial sector are now looking to utilise the technology to build generative AI-powered banks, using AI models to recommend tailored solutions to customers, enhance customer service operations, and assess risk such as credit risk.

Additionally, Jochen shared generative AI is being utilised heavily in trading as companies look to drive outsized returns. For example, he says the tech is being used "for optimising quant finance applications, trading strategies, and also to revolutionise how capital markets do their forecasting".

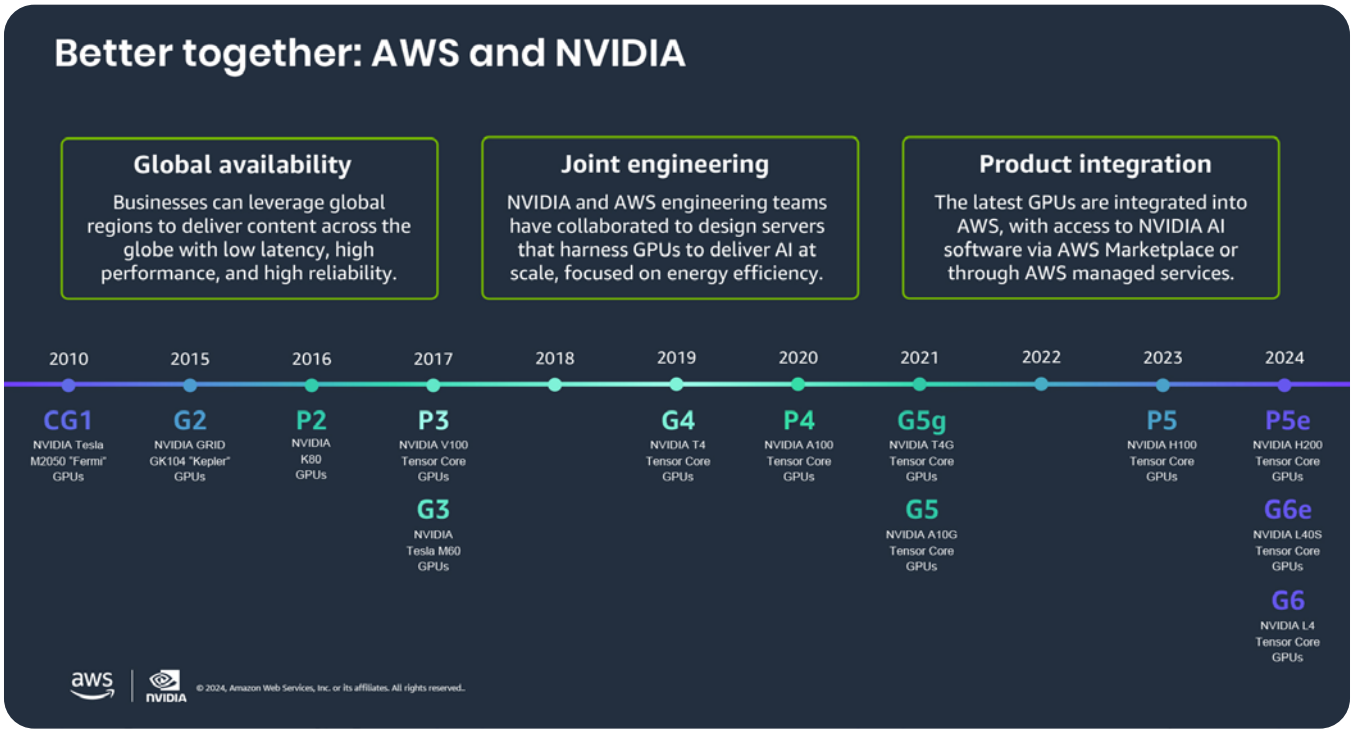
"In payments, it's about building systems that monitor these payment networks. For example, you could generate synthetic data leveraging generative AI tools, and also provide real-time fraud alerts," Jochen adds.

One thing's for certain: generative AI is having a transformative impact on fintech, and the long-term potential for the technology is enormous.

Better together: AWS and NVIDIA's partnership

AWS and NVIDIA have been collaborating since 2010 to deliver large-scale, cost-effective, and flexible GPU-accelerated solutions for customers spanning from the cloud, all the way to the edge and across infrastructure software and services, enabling customers to build a full stack tech offering.

"With GPU accelerated solutions available now in multiple AWS Regions, customers can access the compute power that they need at low latency, high performance, and high reliability, all of which are critical in a fast-moving, data-rich, and highly regulated space like financial services," Sam says.



The two companies have been expanding their generative AI offerings. Sam highlights that customers of AWS and NVIDIA in the Fintech space are now utilising generative AI for use cases including chatbots, document analysis, code generation, video and image generation, and speech recognition to improve customer service.

Sam notes that to realise the full value of these solutions, businesses need to customise AI using their own proprietary data, but notes that building models from scratch can be both "expensive and time consuming".

"With Amazon Elastic Compute Cloud (Amazon EC2) instances powered by NVIDIA GPUs, customers are able to accelerate training and inference for increasingly complex large language models (LLMs) and computer intensive generative AI applications," Sam explains.

"In addition to this, purpose-built tools from NVIDIA such as NVIDIA NIM microservices and NVIDIA Neural Modules (NeMo) microservices, which are part of NVIDIA AI Enterprise available in AWS Marketplace, enable organisations to unlock the full potential of generative AI and LLMs at scale."

> *"With Amazon EC2 instances powered by NVIDIA GPUs, customers are able to accelerate training and inference for increasingly complex LLMs and computer intensive generative AI applications."*
>
> *Sam Edge, Global Head of Fintech, AWS*

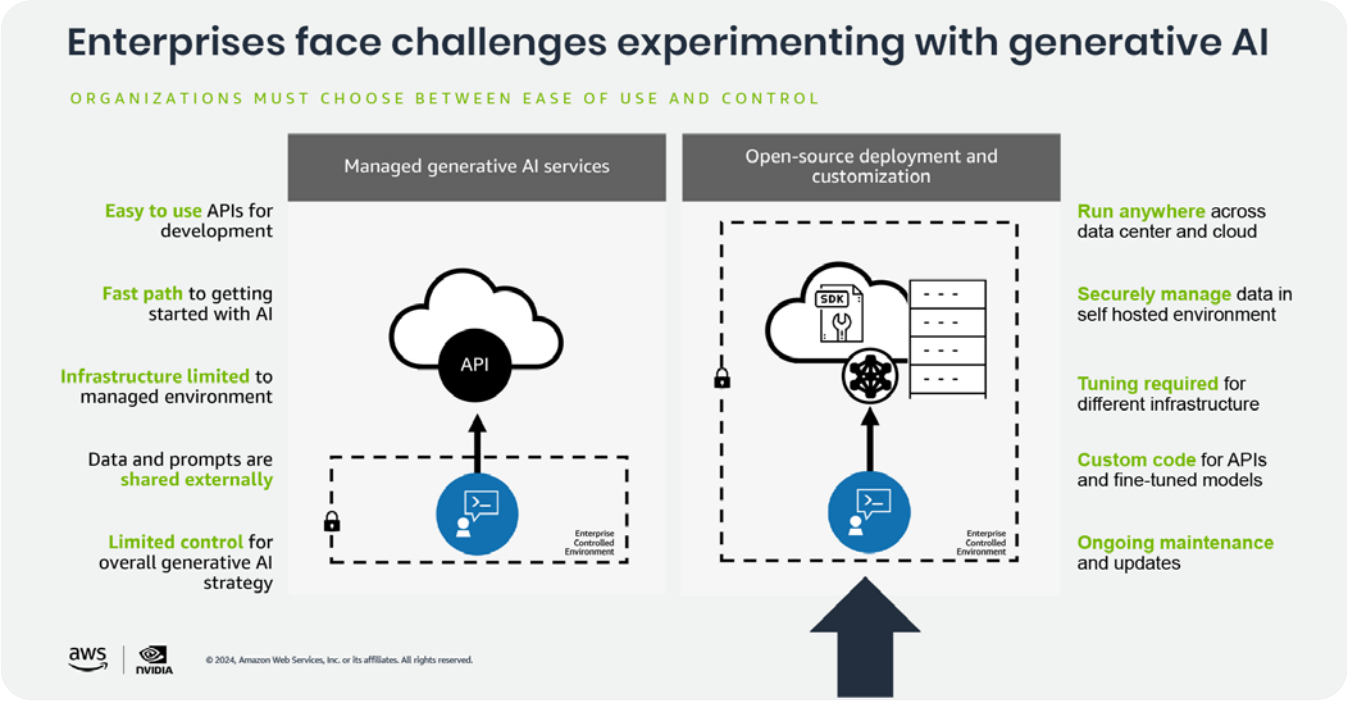
NVIDIA AI Enterprise, NVIDIA NIM and NVIDIA NeMo

Fintech organisations face a number of challenges in what is an ever-evolving landscape. Sam highlights six key challenges that he views as "evergreen" across the industry, namely increased fraud, the growing complexity of data, market volatility, rising infrastructure costs, security and recoverability, and shifting customer demands.

Generative AI can be a very effective tool in helping to tackle these challenges, for example enabling businesses to detect and prevent fraud at scale, as well as provide more personalised customer experiences to improve customer satisfaction and engagement.

But many companies face challenges when it comes to experimenting with generative AI. Jochen shares there are two ways to approach this. The first is to utilise API-based models and managed services where you can develop your applications with an easy-to-use API. While Jochen says this is the fastest way to get started, he also highlights drawbacks in that infrastructure is limited to the managed environment and you have limited scope to expand this further.

The alternative is to utilise open-source models that you can host yourself and fine-tune to meet your needs. However, Jochen says it's "not so easy to run, maintain, and self-host these models", as they require different infrastructure, and you also need ongoing maintenance and updates.



To solve this, NVIDIA provides a number of frameworks, tools, and blueprints to push these open-source models to enterprise level, which can be done using the NVIDIA computing stack on AWS.

NVIDIA AI Enterprise offers end-to-end AI software that includes over 100 frameworks and AI models, all certified and optimised for production on AWS. Jochen describes it as an "operating system for AI", enabling companies to utilise NeMo and NIM to build, fine-tune, train, and run generative AI models in production.

The company offers a range of different API-based modules that can be combined to cover an end-to-end generative AI strategy, from data processing to training, and model customisation to influencing and deployment.

NIM are a set of easy-to-use inference microservices for accelerating the deployment of foundation models on any cloud or data centre.

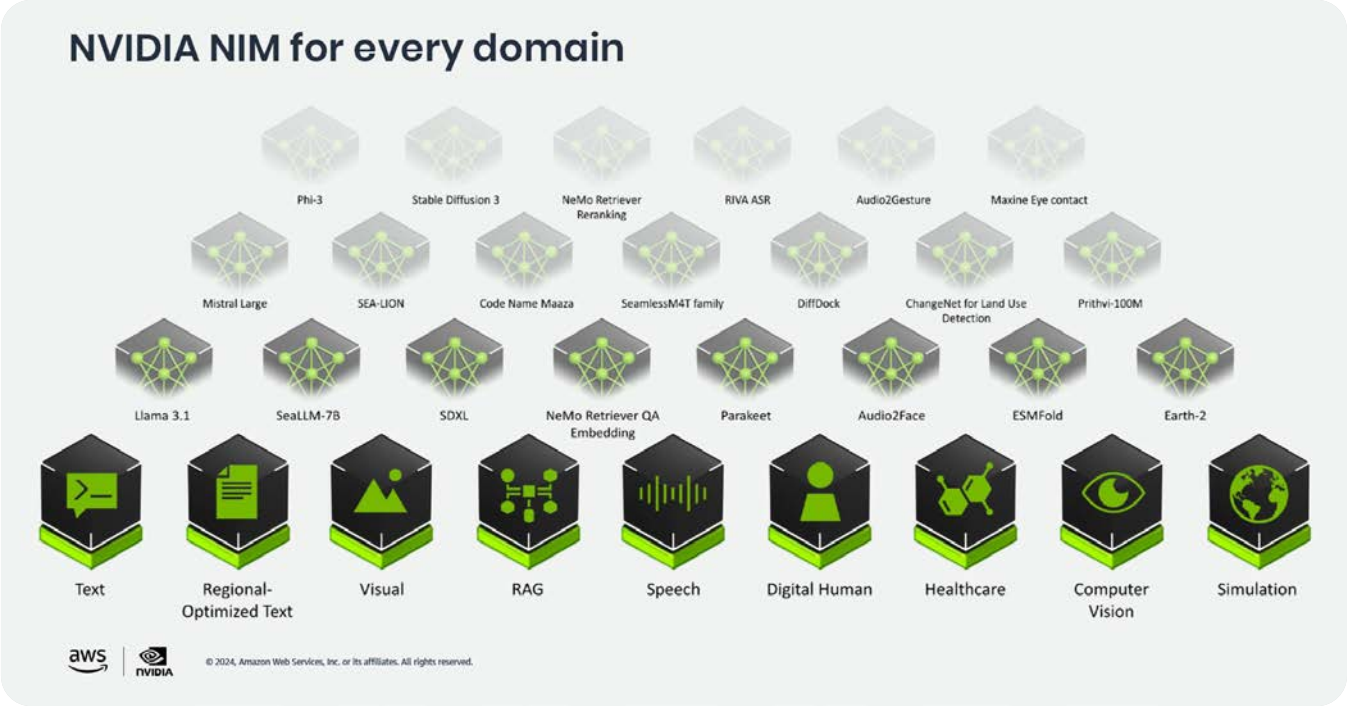
Jochen says all these microservices can be used to "build a blueprint for AI agents, similar to digital workers, that could be trained, updated, and guard-railed in production", adding that this is "the fastest way to deploy AI models on an accelerated infrastructure".

Jochen explained NIM provide a container that can hold a specific model, for example an open-source LLM, which can then be easily customised and deployed at scale.

In terms of customisation, businesses can take a base model and then fine-tune it with their own corporate data or language to create specific LLMs designed for purpose. NVIDIA also offers tools through its NeMo framework to clean and curate the data businesses need for their fine-tuning.

Businesses can even swap models in and out of the container to test newly released models or compare models to see which one has the lowest hallucination, which provides the highest accuracy, and so on.

Jochen highlights that the company has NIM for "every domain", covering not just text generation, but also a range of other use cases, as shown in the image below:



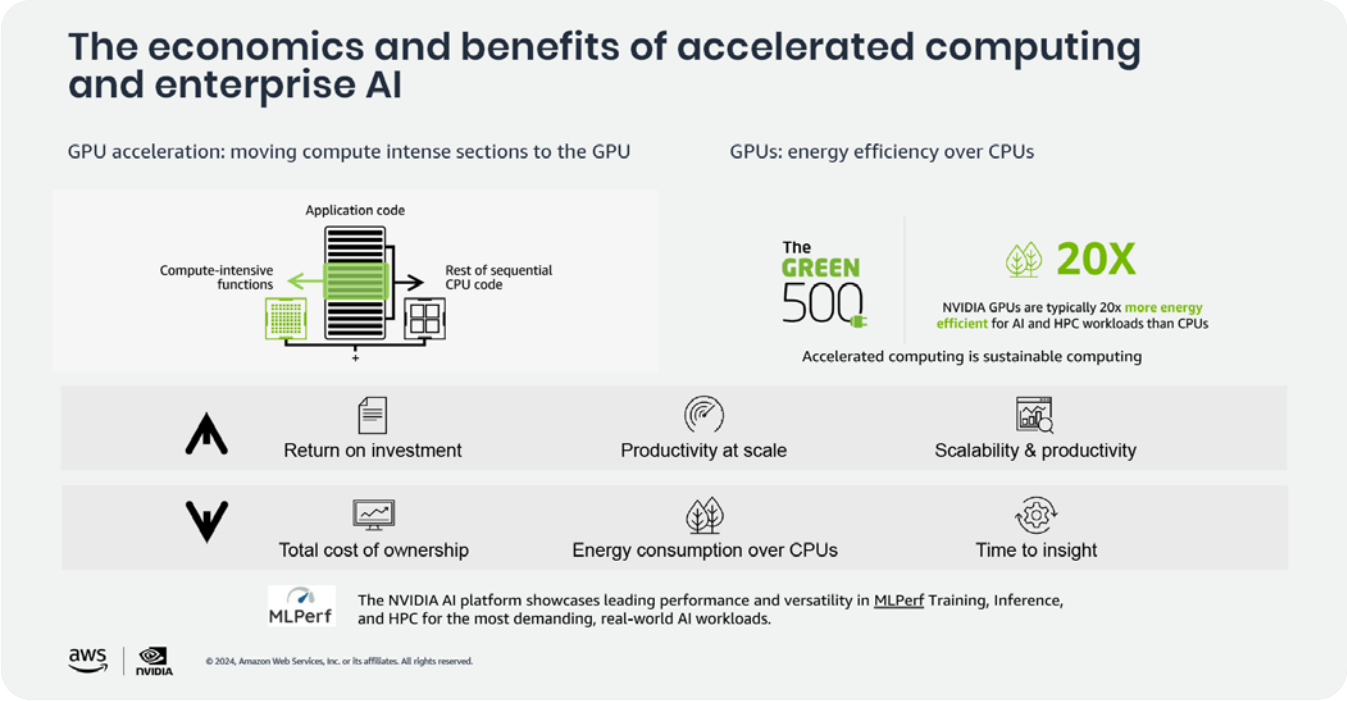
Leveraging NIM enables fintech companies to quickly build and deploy generative AI applications at an industrial scale, covering a wide range of use cases and with increased cost efficiency, customisation options, and performance metrics, allowing companies to effectively manage the entire AI life cycle.

Accelerated computing

As shown above, generative AI is more than just language models, with numerous other modalities such as audio, images, and video also available, enabling companies to utilise a number of combinations to achieve their desired goals.

Fintech firms have huge amounts of data that can be accessed and utilised to create effective generative AI models based on these elements, however a powerful framework is needed to build them.

Jochen says NVIDIA has turned to accelerated computing to streamline this process. With accelerated computing, you can "compute, train, and process the data and even run these models at scale, at speed, in massively parallel ways".



"People sometimes call it sequential computing or supercharged computing, and in the end it results in higher return on investment, and the productivity of accelerated computing users is increased because they can train thousands of models per day. They can select the model that is most trustworthy, most predictive, and then deploy this at scale."

Jochen also highlights the need for low latency model inferencing, which essentially means getting immediate responses from LLMs when you interact with them.

"Users are not very motivated to use these tools when the reaction time of those models is above 300 milliseconds. It needs to be way below that to be able to be perceived as attractive and useful, same with fraud detection methodologies," he says.

As such, for organisations looking to embrace AI and generative AI, the ability to utilise accelerated computing to improve AI model performance and to quickly train, customise, and update models at industry scale is essential.

"Accelerated computing helps us to quickly retrain, customise, update, and monitor our models, introducing guardrails, tests, and observability tools to really validate the quality of these models. It also improves the trustworthiness and reduces the hallucination of these models at rates which are industry accepted so that we can use them in production," Jochen adds.

Customer stories

A number of companies across the globe have partnered with AWS and NVIDIA to power their generative AI offerings, and three customer stories in particular were shared by Sam and Jochen in the webinar.

The first is a personal finance startup that has raised approximately $120 million in funding. The company utilises ML to connect customers with personalised financial products based on their needs, and it wanted to train its ML models quickly and effectively to meet changing customer expectations.

The company leveraged Amazon EC2 P3 Instances powered by NVIDIA GPUs to improve both efficiency, training times for its ML models was reduced from months down to days, and cost, training costs were reduced by around 75%.

The second customer story focused on a financial company that was building a homegrown LLM to create a new finance AI offering. The company assigned close to 1.3 million hours of training time for the LLM on NVIDIA's A100 GPUs running on AWS, combing through 345 billion documents that were sourced from publicly available press releases, news pieces, and public filings and resulting in the creation of a new LLM with domain expertise.

The third story centred on an investment management services company that wanted to accelerate time to market for their ML models. The company selected the NVIDIA AI inference platform to deploy its models on NVIDIA GPUs, facilitated by the NVIDIA Triton Inference Server, on AWS, which reduced model deployment times from several months to 15 minutes.

These success stories highlight the potential generative AI has to truly transform business operations, from cutting costs to drastically reducing time to market. They also highlight the importance of collaboration when leveraging emerging technologies like generative AI, and how partnering with industry specialists such as AWS and NVIDIA and tapping into their knowledge and expertise can enable companies to reap the full benefits of this transformative technology.

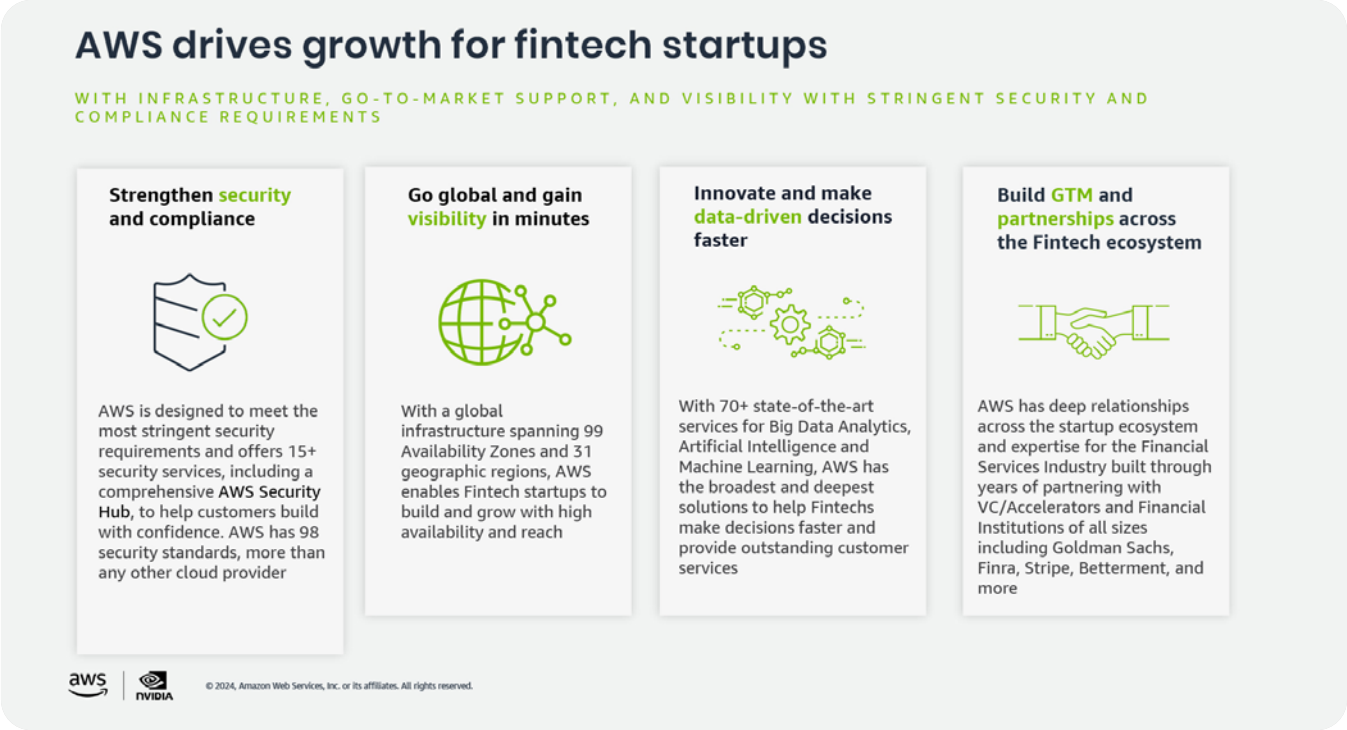
AWS supporting the startup ecosystem

AWS' global startup team aims to help startups such as new fintechs grow on AWS. Sam explains that the company has touchpoints across the ecosystem, "with investors at all stages of maturity, from very early-stage investors and angels, and seed investors, through to accelerators, incubators, universities, and finally late-stage investors including venture capital funds and private equity funds".

"We work with these organisations to help support their cohorts and their portfolio companies, to help them to grow and help them to find more success," Sam adds.

The company also distributes a range of startup and founder support packages, including AWS Activate, which provides access to expertise, technical support, and other resources to assist businesses at the early stage of their journeys. AWS also works with events such as Money20/20 to provide a platform for its partner fintech companies to tell their stories.

AWS has a global cross-functional team of business developers, account managers, and solution architects from a range of specialist backgrounds who work with the highest potential fintech startups to help them find success.

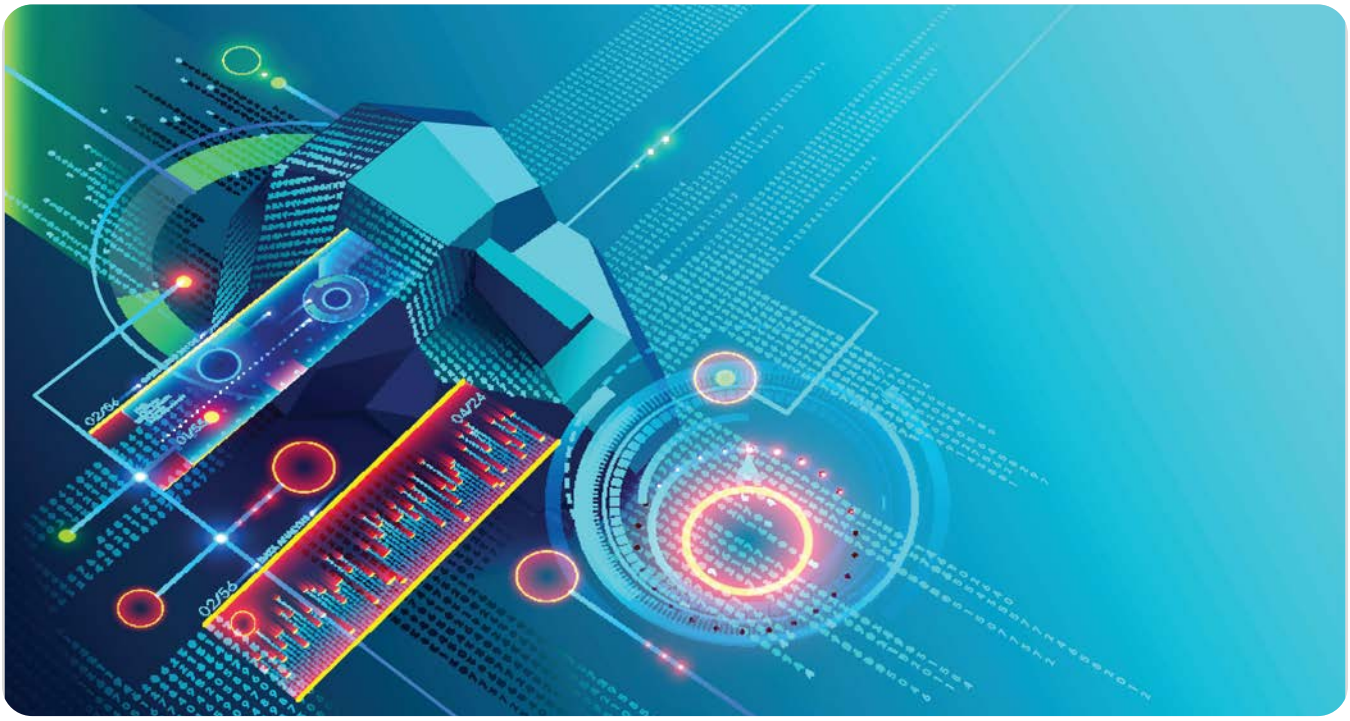


Sam highlights that some of the most common needs AWS supports fintech startups with include "making sure they're embedding security and compliance into their solutions from day one, and increasingly as they scale", helping later stage startups to expand into global markets and quickly gain visibility, and assisting B2B companies with their go-to-market strategies and partnerships.

> *Data has become 'the new oil' and is the 'lifeblood' of financial services.*
>
> *Sam Edge, Global Head of Fintech, AWS*

With the abundance of data in the fintech industry, there's also an increasing trend of startups wanting to make better use of their data. Sam says that data has become "the new oil" and is the "lifeblood" of financial services. As such, AWS is utilising ML technology and data analytics to help companies provide more personalisation for their customers, automate processes, manage their risk effectively, and improve and accelerate their decision making to spur growth.

AWS has been utilising AI for years, and it aims to use the knowledge it has gained through this experience to help fintech companies utilise the technology and grow their offerings.



Conclusion

The use cases for generative AI in fintech continue to grow, and companies such as AWS and NVIDIA are making huge strides in providing the framework and blueprints fintechs can leverage to unlock the full potential of the technology as they look to innovate and scale.

Alongside AWS' startup support initiatives, NVIDIA also offers a wider ecosystem for AI developers, which includes expert forums, training resources, and educational courses to help developers leverage its generative AI offerings effectively.

Through their collaboration, AWS and NVIDIA are empowering global fintechs to build, train, customise, and deploy generative AI at scale, enabling companies at all levels of maturity to boost customer service, streamline processes, and innovate, enabling the sector as a whole to grow and to continue redefining what is possible with technology.

Giving his advice to budding fintech startups looking to utilise generative AI and grow in the industry, Sam says: "My advice is don't work backwards from the technology because I've seen, especially at the application layer, that that can be damaging."

"I think there's still a lot of opportunity within fintech, you know, I think a lot of problems have been solved, but there's still loads of jobs to be done that are calling out for talented, innovative founders."

"So, identify the problem. The bigger the problem, the bigger the opportunity. Don't get distracted. Find the right people. Work with the right support network, including AWS and NVIDIA. And go and build groundbreaking solutions."

"I think it's a great time right now."

In summary

- Generative AI is currently being used in three key ways within fintech: to improve **customer experience**, to boost **productivity**, and to drive **innovation** by enabling companies to build better products and automate processes.
- AI, including ML, can be a very effective tool in helping to tackle industry challenges such as increased fraud, the growing complexity of data, market volatility, rising infrastructure costs, security and recoverability, and shifting customer demands.
- When it comes to experimenting with generative AI, companies can either utilise API-based models and managed services, where you can develop your applications with an easy-to-use API, or companies can utilise open-source models that you can host yourself and fine-tune to meet your needs.
- Through their partnership, AWS and NVIDIA are empowering global fintechs to build, train, customise and deploy generative AI at scale.
- NVIDIA's AI Enterprise solution offers end-to-end AI software that includes over 100 frameworks and AI models, all certified and optimised for production on AWS. Jochen describes it as an "operating system for AI", enabling companies to utilise NVIDIA's NeMo and NIM services to build, fine-tune, train and also run generative AI models in production. The company offers a range of different API-based modules that can be combined to cover an end-to-end generative AI strategy, from data processing to training and model customisation to influencing and deployment.
- AWS provides infrastructure and go-to-market support to help drive growth for fintech start-ups looking to utilise generative AI. The company also distributes a range of start-up and founder support packages, including AWS Activate, which provides access to expertise, technical support, and other resources to assist businesses at the early stage of their journeys.
- Alongside AWS' start-up support initiatives, NVIDIA also offers a wider ecosystem for AI developers, which includes expert forums, training resources, and educational courses to help developers leverage its generative AI offerings effectively.

Why AWS and NVIDIA

AWS and NVIDIA have collaborated since 2010 to continually deliver large-scale, cost-effective, and flexible GPU-accelerated solutions for customers. Spanning from the cloud to the edge, these innovations extend across infrastructure, software, and services to offer a full-stack solution that accelerates time to solution when building and deploying AI into production. With GPU-accelerated solutions available in multiple AWS Regions, customers can access the compute power that they need to achieve low latency, high performance, and high reliability.

For more information, visit our partnership page

Resources

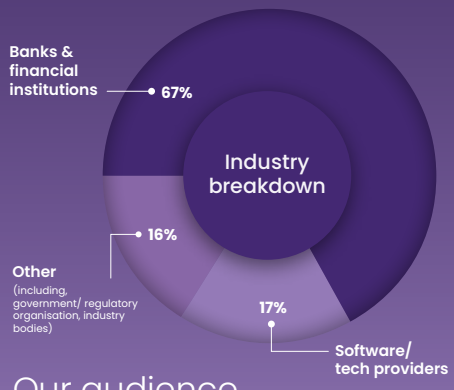
- Learn More
- Ask an Expert
- AWS Activate
- AWS Marketplace
- NVIDIA NIM APIs
- NVIDIA Developer Program
- NVIDIA Deep Learning Institute (DLI)
- NVIDIA Inception

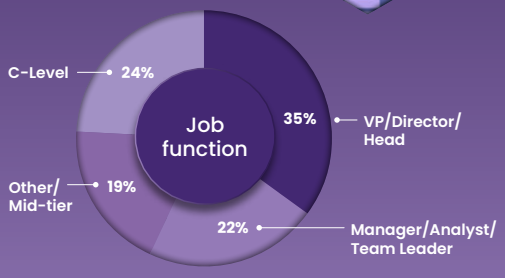
FINTECH FUTURES

fintechfutures.com

FinTech Futures FinTech_Futures

With over 40 years of media expertise, FinTech Futures is the leading global provider of independent intelligence and insight for professionals in the fintech sector.





Our audience

Our portfolio

We have a range of advertising & sponsorship opportunities available To find out more, talk to:

Sam Hutton
Head of Sales
sam.hutton@fintechfutures.com
+44 208 052 0434

Kate Stevenson
Business Development Manager
kate.stevenson@fintechfutures.com
+44 782 593 0099