aws | Deloitte.

Building an enterprise approach for generative AI:

A guide for boards and business leaders

Introduction

Around the world, businesses are turning to AI to boost productivity, improve interactions with consumers and citizens, and create better products, services, and experiences. Over the last 12 months, generative AI (gen AI) in particular has shown considerable potential to transform how businesses interact, engage, and operate. Whilst there are many articles in the media around generative AI at the moment, it is often hard to distinguish hype from reality. Many of the use cases and tools highlighted in these articles are focused on consumer applications and it is often hard to know how these translate into enterprise scale. This white paper is designed to help executives with this challenge and provide practical approaches to help adopt generative AI successfully in your business, leveraging the combined experience of Deloitte and Amazon Web Services (AWS).

What is generative AI?

According to AWS, generative AI is a type of artificial intelligence that can create new content and ideas, including conversations, stories, images, videos, and music. Like all artificial intelligence, generative AI is powered by machine learning models – very large models that are pre-trained on vast amounts of data and commonly referred to as Foundation Models (FMs). Apart from content creation, generative AI is also used to improve the quality of digital images, edit video, build prototypes quickly for manufacturing, augment data with synthetic datasets, and more.

What is the market potential of generative AI?

Research from Deloitte, for example, indicates that generative AI can improve the front-office productivity of the top 14 global investment banks by up to 35 percent, resulting in additional revenue of USD $3.5 million per front office employee by 2026. Here, investment banking divisions can boost productivity by using generative AI to help generate initial deal structures and conduct due diligence, compliance, and valuation. Generative AI can also accelerate trading-related activities such as understanding market sentiment, catching anomalies, and placing orders at scale.

Rich potential benefits such as these are spurring more businesses everywhere to embrace this new technology. Analyst firm Gartner predicts that by 2024, 40 percent of enterprise applications will have embedded conversational AI, up from less than 5 percent in 2020, whilst by 2026, generative AI will automate 60 percent of the design effort for new websites and mobile applications.

How are companies using generative AI today?

Businesses and government organizations around the world are already successfully using generative AI to drive innovation, enhance experiences, and transform industries.

- **Revolution in retailing:** A leading e-commerce platform used generative AI to create personalised shopping experiences for its users. The generative AI model generated product recommendations tailored to individual tastes by analysing user behavior, preferences, and previous purchases. The result? A 30 percent increase in sales and improved customer satisfaction.
- **Healthcare innovation:** A large healthcare institution used generative AI to analyse massive datasets of patient records to predict potential health risks, allowing for early intervention and personalised treatment plans. This proactive approach resulted in improved patient outcomes and lower healthcare costs.
- **Media content creation at scale:** A large media company deployed generative AI to automate the content creation for its digital platforms. The generative AI model streamlined content production, ensuring the timely delivery of news summaries, personalised playlists, and more to millions of users.
- **Financial services fraud detection:** A large financial institution implemented a generative AI-based model to improve its fraud detection mechanisms. The model identified and flagged suspicious activities in real time by analysing transaction patterns and behaviors, reducing fraud incidents by 40 percent and increasing customer trust.
- **Design and prototyping in manufacturing:** An automotive manufacturer used generative AI to help design and prototype new vehicle models. Based on market trends, user feedback, and sustainability criteria, the generative AI model created innovative design concepts. This accelerated the design process, shortening time-to-market, and generated designs that appealed to target audiences.
- **Designs in jewellery:** A high value gems and jewellery manufacturer is working with Deloitte to create designs through generative AI. If a customer requests a product design that is not available in the manufacturer's catalogue, generative AI creates and shares some potential designs with the product team for feasibility, as well as the customer for preference.

Generative AI delivering value today for Chorus New Zealand

Chorus is at the forefront of connecting New Zealanders to the world. It built, and now operates the majority of the country's fibre internet network, ensuring high-speed, reliable, and world-class internet connections for homes and businesses.

Their architecture repository contains a complex technology landscape. However, this repository poses challenges for new staff, who find it difficult to navigate, and for anyone seeking quick answers to their queries.

In collaboration with their partner, Deloitte, Chorus initiated the first phase of their generative AI journey. They embarked on a gen AI proof of concept using Amazon Bedrock. The primary objective was to simplify the process of querying the repository and accelerate the retrieval of answers. The team successfully developed a gen AI tool that efficiently ingested a select repository of data and provided a user-friendly chat interface for seamless querying.

"The POC yielded outstanding results for Chorus. The gen AI tool delivered a 75% productivity increase compared with new staff performing the same activity prior to the tool and gave a significantly better user experience. Buoyed by this success, Chorus is now actively exploring additional areas where gen AI can expedite operations and streamline processes."

— Michael Brosnahan, Portfolio Architect, Chorus New Zealand

Across the Asia-Pacific and Japan regions, business leaders are carefully evaluating the promise and realities of generative AI. Key questions on their minds include: What are the true capabilities of this emerging technology? Is generative AI a substantiated breakthrough or hype? Will its abilities continue improving rapidly? Which business functions and processes could be transformed through applied generative AI? What is the total cost profile, both current and future, for utilising this technology? What investments in human talent are required to successfully adopt generative AI and capture concrete value? Finally, should organisations build in-house generative AI expertise or source it as a managed service?

Though generative AI holds much disruptive potential, it also surfaces complex strategic choices for enterprises. Thoughtful consideration must be given to match promise with practical adoption. By prudently exploring these questions, Asia-Pacific and Japan business leaders aim to chart an optimal course to leverage generative AI for competitive advantage.

This whitepaper, prepared by Deloitte and AWS, explains the opportunities and challenges associated with generative AI and provides a practical guide for executives and leaders to address them.

How to build an enterprise approach for generative AI

Start with a strategy

Due to the transformational nature of generative AI, to realise its true potential, many organisations will need to make significant changes to the culture, processes, technology platforms, and even their products and services.

Organisations looking to adopt generative AI need to start by developing a comprehensive strategy. This will establish the vision, guardrails, and roadmap for leveraging this transformational technology.

Key components of a generative AI strategy include building organisational capability, identifying high-value use cases, determining the optimal adoption model, selecting the right technology platforms, and understanding and mitigating risks. On the capability front, organisations must assess their current skills and talent, and invest in developing the human capital required to effectively deploy generative AI. The strategy should outline a plan for hiring data scientists and engineers, upskilling employees, and fostering an innovative culture open to testing and learning. Another critical step is rigorously evaluating potential use cases and prioritising those that align to core strategic goals and can generate clear returns. Focusing generative AI on the highest-impact levers will allow organisations to demonstrate value early on.

- Organisations also need to decide between building in-house capabilities or leveraging external partners and managed services. This adoption model choice has implications for control, cost, speed, and access to leading generative AI solutions. Choosing the right enabling technologies, including AI cloud platforms, data foundations, and modern IT architecture is equally important.
- As is proactively assessing and governing for risks associated with leveraging generative AI.

Building your organisational capability

In order for organisations to adopt generative AI and scale it to deliver meaningful business value, they will need to develop new capabilities across their organisation.

- **Build AI fluency**

  To successfully leverage generative AI, boards and executive leaders need to build a level of AI fluency across their organisations. AI fluency means understanding what AI is, why the technology is valuable to the business and how it can be applied to specific domains. Because generative AI is more decentralised than traditional AI solutions, almost every worker in a business needs to understand how they can obtain value from generative AI and what their responsibilities are in relation to it, as well as the risk it may pose and how they can mitigate any potential issues. To build their fluency in generative AI, businesses should adopt tailored learning programmes that address the needs of different personas.

  Deloitte leads the development of AI fluency with boards and executive teams, known as the AI Fluency Greenhouse experience, where organisational leaders can attend tailored workshops to understand what AI is, how it works, the current level of maturity at their business, governance and risk, and the ways other participants in their industry are using the technology. The AI Fluency Greenhouse experience provides a pragmatic and practical opportunity to interact with the technology and understand its application across multiple dimensions and use cases. The session typically generates collaborative discussions within the group about how to introduce AI into areas such as the business strategy. This brings boards and executives onto the same page to understand their common aspirations, and the actions that need to be undertaken to realise the state they want to get to.

  In addition, AWS offers a range of business and technical enablement for customers looking to build skills around generative AI on AWS. This enablement can span from self-paced, high level on-line training that allows business users to understand the basic concepts around generative AI, through to in-person hands-on labs, designed to give technical practitioners the experience they need to start building their own solutions.

- **Adopt continuous learning**

  The pace of change surrounding generative AI today is dramatic. New models are being released every month and new use cases are being identified every day. Even for the most experienced AI practitioners, staying up-to-date with the latest capabilities can be a challenge. For organisations looking to build AI fluency across their business, they cannot treat training their employees as a 'one off' exercise. Successful organisations will require leaders to establish a culture of continuous learning, where they look to create tools and processes that allow their staff to undergo ongoing education in the latest generative AI use cases and techniques. From there, they can also build a mindset of experimentation that can help drive further improvements and efficiencies by deploying generative AI across a range of new use cases.

- **Encourage a culture of experimentation**

  A culture of continuous experimentation and improvement means encouraging leaders to move beyond expectations set by traditional software projects to understand what a typical generative AI project looks like, what success rates to expect, and how to make investment choices around certain use cases. It also means establishing the psychological safety for teams to experiment and balance risk versus reward. Once teams and individuals generate ideas for new solutions, these can be passed to a team that can prototype them quickly. If teams don't adopt this approach, the business may end up too focused on improving processes whilst missing out on the chance to develop innovative new solutions that could transform the way they work.

- **Embrace open communication**

  Another important element in the successful adoption of generative AI in any organisation is communication. Employees often become concerned when businesses start adopting AI solutions, fearing that it may negatively impact their role. Organisations looking to adopt generative AI should look to communicate with their employees around why the organisation is adopting this new technology and how they can be involved to build new skills and take on new, enhanced roles, as a result.

Choosing the right use cases

When organisations are considering how to navigate their generative AI journey, it is often difficult for them to identify the right use cases. Whilst the specific use cases may vary by organisation and industry, it is imperative that organisations focus on use cases with real business value and senior stakeholders who will sponsor the projects to deliver them. Too often organisations start with use cases that are technical proofs-of-concept and deliver little business value. These projects typically do not have the business case to progress into production. We suggest that use cases are prioritised by balancing: value delivered and complexity to deliver, ensuring you start with use cases that are valuable but achievable.

The ways businesses can achieve value with generative AI typically fall into two categories: Increasing productivity and efficiency in existing processes and creating new products and services. Today, most organisations are focused on the former scenario as driving increased efficiency in your processes and your existing teams is often an easier way to get started, with lower risk, whilst you learn how to successfully productionise these solutions.

When we consider the types of generative AI use cases that are most common today, these typically fall into four broad categories:

1. The first category includes classic content generation use cases. These use cases are where generative AI is used to accelerate the creation of new content, such as marketing copy, emails, or software code. This approach is typically used by an expert user to generate a document first draft much quicker than what could be done manually. This approach uses the AI solution to do the undifferentiated heavy lifting, allowing the expert user to then refine into something more differentiating for the organisation.
2. The second use case category is not one that gets a lot of attention, but rather the use of these models occurs behind the scenes, where you may never even know that you are working with one of these generative models. These models are now very effective at driving improvements to search, personalisation, and summarisation, so in instances where internal and external parties search data, make recommendations, or even provide simplified summaries of lengthy data, these models can improve the efficiency and effectiveness of these processes.
3. The third area is around enabling new domain area exploration. This is the situation where you have a customer or employee that wants to learn about a product or process by 'chatting' with the system.
4. The fourth area is around expert decision-making where LLMs are used to provide information to expert users that enables them to make better and faster decisions. A great example of this is in the customer contact centre where generative AI can assist agents during customer calls by analysing call transcripts in real-time to make suggestions to the agent on how to address questions without placing the customer on hold to look up answers.

'Generative AI presents opportunities for transformational innovation across all industries, and within every line of business of an organisation.'

— AWS

Technology

Choosing the right technology platform is crucial for successfully implementing generative AI within an enterprise context. At AWS, we recognise that adopting new technologies like generative AI requires more than just access to capable models.

Delivering business value with generative AI requires a platform with capabilities that go well beyond data science. At AWS our integrated cloud platform provides over 200 services, across areas such as data and analytics, database, security, and IoT, that can be quickly integrated to develop modern business applications that are secure, reliable, and scalable.

When it comes to generative AI, AWS also offers the flexibility for customers to choose from a range of foundation models and model providers to deliver generative AI capabilities across text, image, and software code modalities. Businesses can select the model that best meets their needs for a given use case. Amazon Bedrock provides models with a range of sizes to suit customers' use cases, accuracy, latency, and cost requirements. Customers can also choose whether the model is open source or from a proprietary model provider.

As new models are being developed and improved at a high velocity, customers need to develop a strategy for how they will assess new models against their existing ones and how they will decide when to upgrade their solution to a new model. Whilst this is not an exact science, the rate of development of new models means that customers looking to avoid continuous churn should consider new models at designated upgrade cycles, if their existing solution is no longer meeting their current needs or won't support future use cases.

Amazon Bedrock offers several capabilities to support security and privacy requirements and is compatible with common compliance standards including GDPR and HIPAA. With Amazon Bedrock, your content is not used to improve the base models and is not shared with third-party model providers. You can use AWS PrivateLink with Amazon Bedrock to establish private connectivity between your foundation models and on-premises networks without exposing your traffic to the internet.

As generative AI solutions leverage very large AI models, they require significant compute power to run, which creates both cost and performance considerations. AWS provides a range of compute capabilities to develop and deploy generative AI solutions designed to minimise cost and accelerate performance. AWS Trainium is a machine learning accelerator that AWS purpose-built for training large AI models. AWS Inferentia2 is a purpose-built ML accelerator for running inference on large language models. AWS Trainium and AWS Inferentia2 provide high performance and the lowest cost in Amazon EC2.

Real-world applications of generative AI are the best way to comprehend its power. Amazon Web Services' pre-built generative AI–powered solutions, such as Code Whisperer, provide concrete examples of how the technology can transform industries, improve customer experiences, and spur innovation. Amazon CodeWhisperer is a generative AI powered coding companion that can generate software code suggestions in real time from natural language comments in your code helping to accelerate software development.

> *'Our mission is to make it possible for developers of all skill levels and for organisations of all sizes to innovate using generative AI. This is just the beginning of what we believe will be the next wave of ML powering new possibilities for you.'*
> ***— AWS***

Determining your approach to leveraging generative AI

Deciding how to implement generative AI requires customers to evaluate their needs and resources. Organisations must determine whether it is best to build custom solutions internally or leverage solutions from vendors.

For those seeking to develop their own generative AI solutions, tools like Amazon Bedrock provide options for customers to consume an existing foundation model or fine tune a foundation model with the corporate data. Alternatively customers could choose to leverage tools like Amazon SageMaker to build their own foundation model from scratch.

- **Model consumers** can achieve quick wins by deploying pre-trained models from model providers such as Anthropic, AI21 Labs, Cohere, Stability AI, Meta, and Amazon. This allows tapping into powerful generative AI capabilities, without requiring extensive data science expertise.
- **Model tuners** take it a step further by customising foundation models for specific use cases or industries by fine tuning the base foundation models on their own corporate data, allowing more tailored performance.
- **Model building** is recommended for those with ample data science resources and unique needs not addressed by existing models.

Regardless of approach, customers should weigh the trade-offs of speed, cost, complexity, and performance as they determine the best strategy to introduce generative AI into their organization.

The approach you choose for adopting generative AI can have a significant impact on the cost of the solution. When consuming existing foundation models, the model provider incurs the cost of building and training the model. Customers using that model will typically pay a charge that is based on a price per token. (With Amazon Bedrock, a token is comprised of a few characters and refers to the basic unit that a model learns to understand user input and prompts to generate results. For image generation models, you are charged for every image generated.) With this approach, the cost of using generative AI is dependent on how much text or how many images you put in the model and how much is output. This pricing model is particularly attractive for use cases that are low volume or sporadic in nature.

If customers wish to fine tune an existing foundation model on their data, in addition to the usage charges, there will also be a cost associated with running the fine tuning job and hosting the customised model. As customers are typically tuning the pre-trained model on a small amount of corporate data, the cost to fine tune the model is significantly less than training a new model.

Where customers elect to build their own foundation model, they will need to train the full model across petabytes of data, leveraging very large volumes of compute resources to train the model. This approach requires significantly more effort and expertise and the costs can run into the millions of dollars. The customer will also bear the costs associated with hosting the trained model for running the inference on the model as well. This approach is only recommended where a company sees significant business value from developing their bespoke model and where they have the necessary volume of proprietary data to train the model.

Understand and manage the risks

The rapid advancement of generative AI introduces exciting new capabilities, but also new challenges we must thoughtfully navigate. As enterprises adopt AI-powered technologies, they need to be mindful of risks around bias, fairness, transparency, and accuracy of responses.

Generative models can introduce additional risks that need to be managed, such as the propensity for them to make factually incorrect statements (hallucinations), or breach privacy through memorisation of proprietary data. Organisations need to understand if a model they have chosen will use the data it receives within prompts to retrain the model, compromising their corporate IP in the process.

As generative AI leverages foundation models built with pre-trained datasets obtained from third party model providers (that vary from model to model), consumers should understand what data a foundation model was trained on to ensure that the source is acceptable.

Another challenge associated with large language models is that they can perform a wide range of tasks and can potentially respond to questions from users on unintended topics. Organisations need to put guardrails in place to ensure that the models are only responding to the intended scope of requests and are not used for illegal or inappropriate purposes.

As generative AI provides the ability to create new content, consideration should be given to moderation of content generated for adherence to communication policies established by an organisation to ensure that standards are met for the working environment.

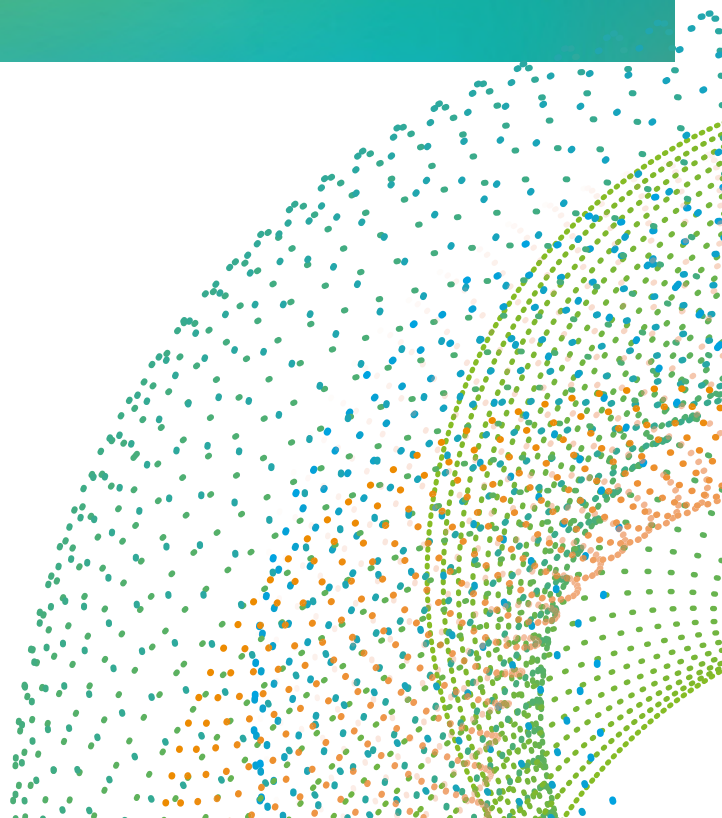
Regulatory compliance is another key consideration around implementing generative AI solutions. Regulations around generative AI are still being developed in many geographies and differ between countries. Organisations should put in place checks to ensure that they are compliant with existing regulations as well as forming internal policies around areas where regulation may not yet exist, to ensure that they have their own guidelines around acceptable usage. Whilst regulations governing generative AI are still being developed in many jurisdictions, businesses can look to guidance or frameworks from authorities or industry bodies in the interim.

Addressing responsibility issues is a key step in getting started with the technology. The Deloitte Trustworthy AI Framework is one tool businesses can potentially use as a guide here. Trustworthy AI™ requires governance and regulatory compliance throughout the AI lifecycle from ideation to design, development, deployment, and machine learning operations (MLOps) anchored on the seven dimensions in Deloitte's Trustworthy AI™ framework – transparent and explainable, fair and impartial, robust and reliable, respectful of privacy, safe and secure, and responsible and accountable.

Responsible AI practices will allow enterprises to deliver helpful, harmless generative AI systems to customers. With thoughtful implementation, we can unlock benefits of the technology's benefits whilst safeguarding trust. A commitment to responsible innovation is crucial as we develop this powerful new computing paradigm.

'While generative AI powers new possibilities, it is also powering a debate on its risk challenges. Leaders will need to weigh the risks and rewards to maximise trust and use generative AI responsibly to unlock its full value.'

*— **Diego Saenz,** Managing Director | AWS Lead Alliance Partner (Data + AI), Deloitte*



Get the right support

Whilst many companies are experimenting and exploring generative AI capabilities internally, companies who want to accelerate their journey should look to collaborate with experts who can provide the skills and capabilities to scale their adoption of generative AI.

By combining robust in-house expertise with effective collaborations with technology companies, consulting partners, academic institutions, and AI start-ups, businesses can increase their chances of success with generative AI.

In April 2023, Deloitte announced a new generative AI practice to help clients harness the technology to enhance productivity and fuel innovation. The new practice combines the world-class services, AI talent, and deep industry experience that enterprise leaders need as they design their generative AI strategies and leverage the disruptive new technology to create innovative AI-fueled applications. At the core of the practice is a Generative AI Market Incubator with a dedicated team of engineers focused on the rapid development of generative AI pilot programs, demos and proofs-of-concept and most valuable practices, and an R&D team working alongside Deloitte alliance partners to train and tune foundation models.

The AWS Generative AI Innovation Center provides customers with expert guidance on successfully adopting generative AI techniques to drive business value. Through this program, customers partner directly with leading scientists and strategists with AWS to imagine and identify new applications for generative AI tailored to their needs. The Innovation Center helps customers select the most promising use cases based on business priorities, providing estimations of feasibility and value to inform decisions. Customers then receive hands-on help developing scalable solutions, integrating generative AI securely into existing systems and workflows.

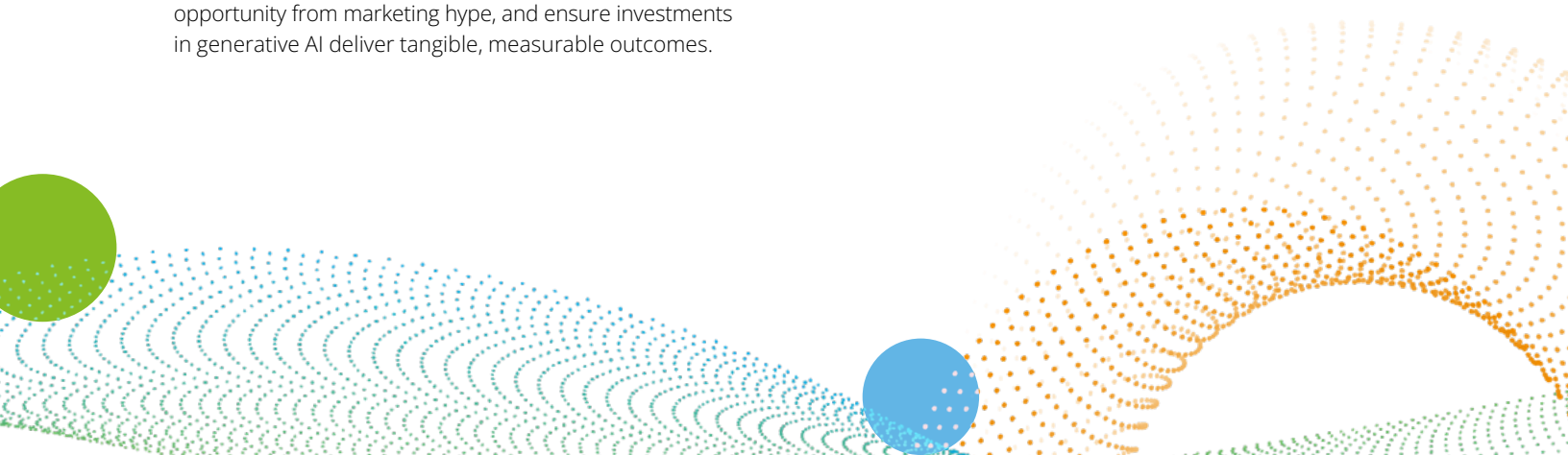
Start experimenting today

We recommend customers begin exploring and adopting generative AI today to develop competitive advantage rather than deferring initiatives. Generative AI represents a disruptive technological shift that will substantially impact businesses and industries in the coming years.

Organisations that start accumulating experience and building capabilities with generative AI now will have a significant competitive edge versus those who delay adoption until the technologies are more mature. They will move faster up the experience curve, understand how to best leverage generative AI for business value sooner, and have more time to build trust with customers. Only by taking generative AI solutions through to production can businesses determine their value and understand how to manage the risks. This can enable leaders to distinguish opportunity from marketing hype, and ensure investments in generative AI deliver tangible, measurable outcomes.

Deferred adoption also brings opportunity cost in the form of potential revenue growth, cost savings, and other benefits that could have been realised sooner. We have already seen early adopters of generative AI gain advantages in areas like marketing content creation, product ideation, and customer service automation. Customers who wait risk missing growth opportunities and playing catch up later.

Early adopters are reimagining business models, products and services. Starting the experimentation phase now is allowing these organisations to build expertise, refine strategies, and develop best practices. This proactive approach ensures that when generative AI becomes mainstream, these enterprises will be not just be participants but leaders in the field.



Conclusion

Generative AI represents a transformational opportunity for enterprises across Asia Pacific and Japan to boost productivity, enhance customer experiences, and develop innovative new products and services. However, to realise the full potential of this emerging technology, businesses must take a thoughtful approach grounded in strategy.

Key steps include building organisational capability through training and fostering a culture of experimentation, carefully evaluating use cases for alignment to business goals, implementing responsible AI practices to manage risks, and collaborating with partners to accelerate learning.

Though generative AI introduces new complexities, its capabilities far outweigh the challenges. By proactively building capabilities today, organisations position themselves to lead in leveraging this technology for competitive advantage.

The future is bright for enterprises ready to shape it. With pragmatic adoption, generative AI can propel businesses to new heights of performance.

Contributors

Christopher Gregory Lewin
Asia Pacific GenAI Lead, Deloitte
chrislewin@deloitte.com
Mobile: +65 6232 7128

Jae Lee
Asia Pacific AWS Lead Alliance Partner, Deloitte
jaelee1@tohmatsu.co.jp
Mobile: +81 80 4599 4826

Takafumi Moriya
Japan AWS Lead Alliance Partner, Deloitte
tmoriya@tohmatsu.co.jp
Mobile: +81 80 4366 5330

Damian Harvey
New Zealand AWS Lead Alliance Partner, Deloitte
dharvey@deloitte.co.nz
Mobile: +64 27 558 5559

Kuldeep Singh
Principal Global PDM AI/ML,
Analytics, Database, AWS
ksingh@amazon.com
Mobile: +1 925 699 2341

Brad Ryan
Analytics & Machine Learning
Partner Lead APJ, AWS
brdryn@amazon.com
Mobile: +6 141 493 9247

Deloitte.

aws

Since 2006, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud. AWS has been continually expanding its services to support virtually any workload, and it now has more than 240 fully featured services for compute, storage, databases, networking, analytics, machine learning and artificial intelligence (AI), Internet of Things (IoT), mobile, security, hybrid, media, and application development, deployment, and management from 105 Availability Zones within 33 geographic regions, with announced plans for 18 more Availability Zones and six more AWS Regions in Malaysia, Mexico, New Zealand, the Kingdom of Saudi Arabia, Thailand, and the AWS European Sovereign Cloud. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—trust AWS to power their infrastructure, become more agile, and lower costs. To learn more about AWS, visit aws.amazon.com.