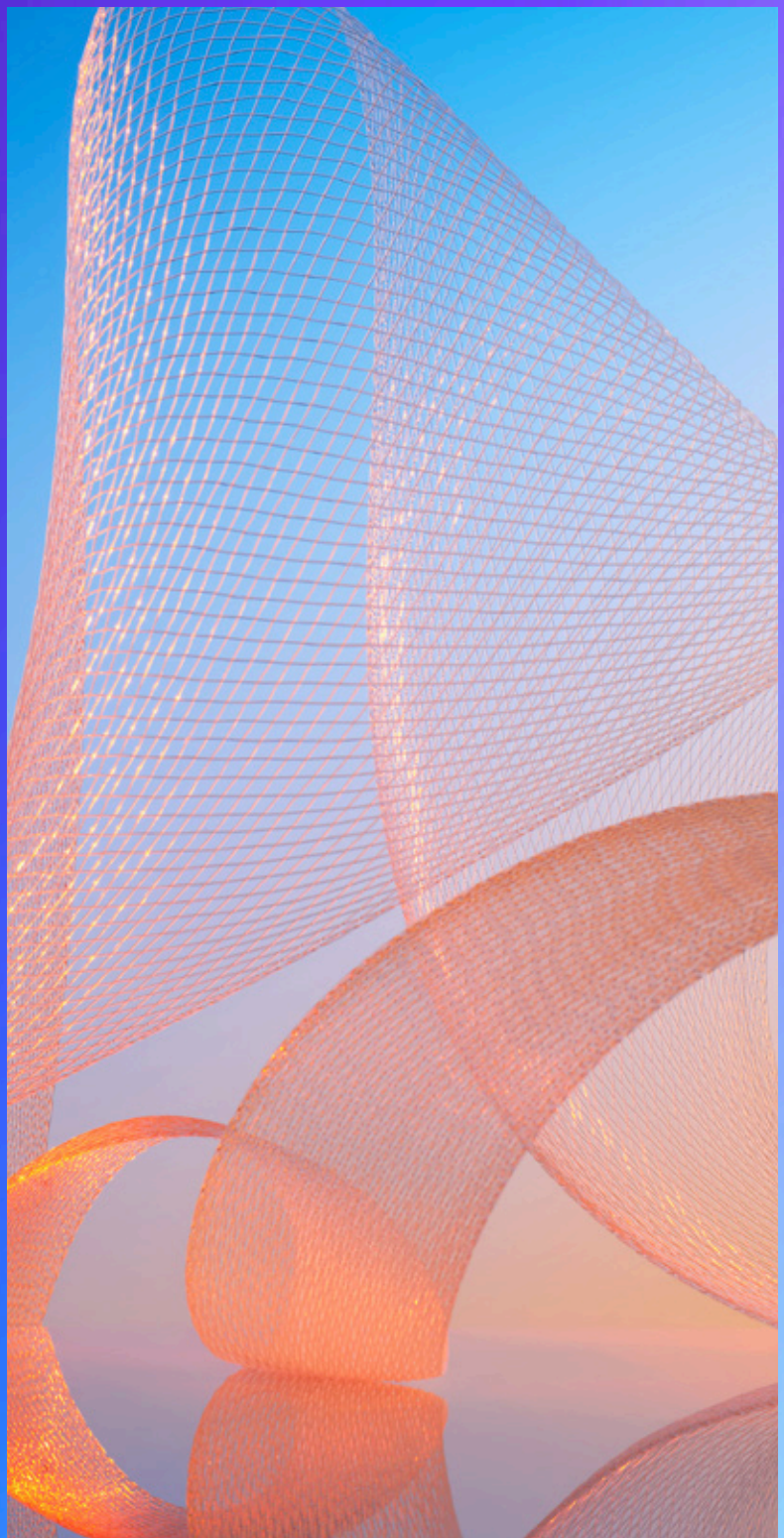
BUSINESS INNOVATION BRIEF

Harness generative AI to accelerate growth



Executive summary

A surge of excitement across consumers and businesses—combined with broad, easy access to the technology—has led us to an inflection point for generative artificial intelligence (AI).

Most leaders recognize the importance of this moment and the need to quickly develop a strategy to explore generative AI. But many may still have questions about it, including:

- What is generative AI?
- How is generative AI different from AI or ML?
- What are the main business use cases?
- How can we use generative AI to integrate with our applications?
- How should I start?
- What are the risks, and how can they be mitigated?

This business innovation brief provides an overview of generative AI, outlining its capabilities, use cases, and business value. It also offers valuable insights from Amazon Web Services (AWS) subject matter experts, leveraging our extensive knowledge and experience in AI and machine learning (ML) technologies.

Who is this content for?

This innovation brief is designed for business leaders in software companies seeking to better understand generative AI—and learn how they can leverage it to achieve business goals.

Table of contents

INTRODUCTION

A new world of intelligence

Consumers and businesses alike are fascinated by generative AI's ability to create sophisticated content, generate code, answer questions, and more—all from simple natural language prompts, often within seconds.

While a lot of attention has been given to how consumers are using generative AI, there is an even bigger opportunity in how businesses will use it to deliver amazing experiences for their customers and employees. The true power of generative AI goes beyond a search engine or chatbot and will transform every aspect of how companies and organizations operate.[1]

McKinsey estimates that 75 percent of the value generative AI could deliver will come from four key areas: customer operations; marketing and sales; software engineering; R&D. The direct positive impact on software engineering productivity could range from 20 percent to 45 percent.[2]

[1] "Rate of generative AI adoption in the workplace in the United States 2023, by Industry," Statista, December 2022
[2] "The economic potential of generative AI: The next productivity frontier," McKinsey, June 2023

Seizing the opportunity

Companies across industries are racing to seize the economic opportunities that generative AI presents. If leading financial projections prove accurate, the rise of generative AI is likely to usher in a new era of the global economy and pave the way for software developers to help power the next stage of their customers' growth.

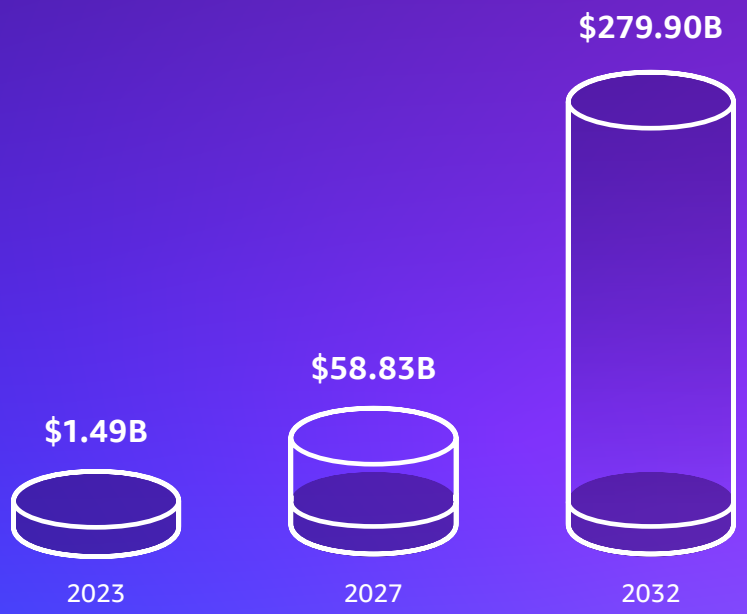
According to research by Goldman Sachs, generative AI could increase global GDP by as much as 7%, or roughly $7 trillion, over the next 10 years.[3]

These lofty financial projections are not driven by consumer interest alone. The potential of generative AI to improve business productivity and outcomes accounts for just as much, if not more, of the excitement and enthusiasm surrounding the technology.

For businesses of every size and across every industry, generative AI is a revolutionary technology that is beginning to drive considerable value—and has the power to fundamentally transform the business landscape.

Expected annual growth for generative AI

Market forecast to grow at a CAGR of 34.2%[4]



[3] "Generative AI could raise global GDP by 7%," Goldman Sachs, April 2023
[4] "Generative AI to Become a $1.3 Trillion Market," Bloomberg, June 2023

The strategic imperative

Businesses across industries and around the world are looking to leverage generative AI to optimize costs, accelerate innovation for enhanced customer experiences, and increase productivity.

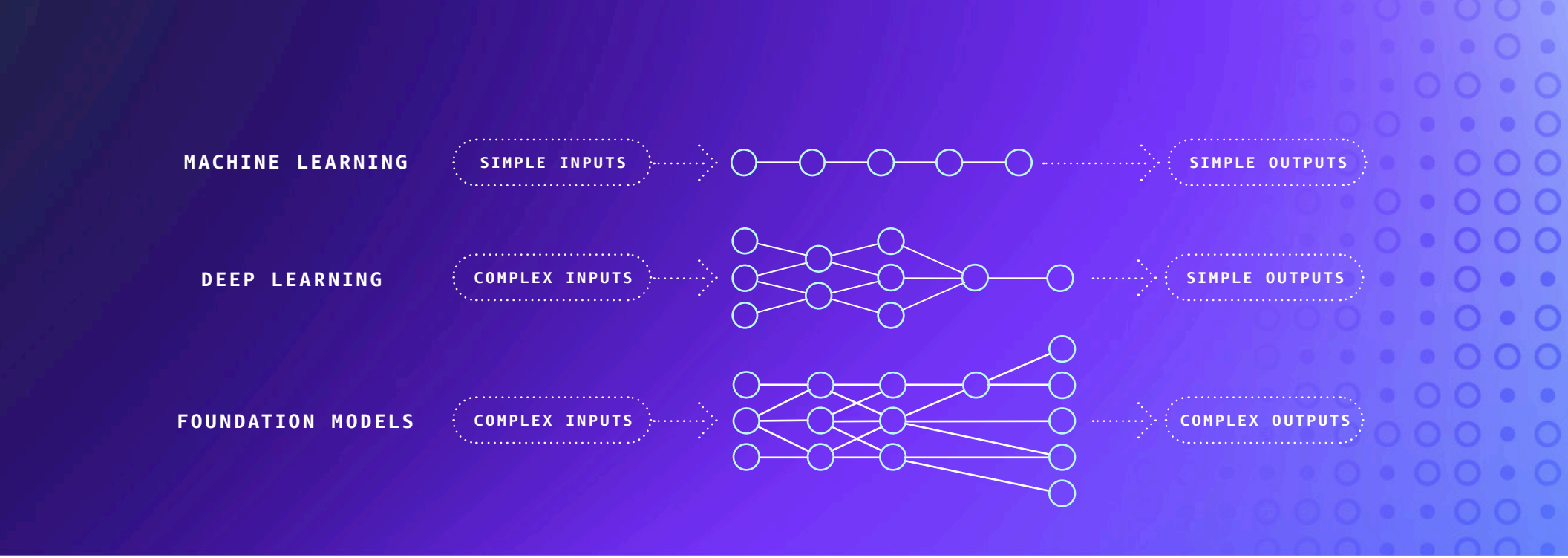
For most businesses, however, the path to achieving these benefits remains unclear.

Many business leaders are aware that generative AI can help achieve better outcomes with fewer resources. This is of particular interest for software companies, as development efficiency is crucial. Software engineering, for example, could see a 20 percent to 45 percent increase in productivity, thanks to its use.[5] This would be reached by reducing time spent on activities like generating initial code drafts, analyzing root-cause or creating new systems design.

Business leaders are mindful that they need to take advantage of generative AI quickly, otherwise their competitors may take the lead. However, few have succeeded in developing strategies for how they will adopt the technology, where they will put it to use, or how they will achieve and measure their results.

Read on to learn how your company can start realizing the business value of generative AI today—so you can keep pace with the market and leapfrog your competition.

[5] "The economic potential of generative AI: The next productivity frontier," McKinsey, June 2023



Understanding generative AI

Before your business can fully unlock the business value of generative AI, it's important to have a fundamental understanding of how the technology works.

Generative AI is a term used to describe algorithms that can create new content and ideas, including conversations, stories, images, videos, music, and code. Generative AI is powered by extremely large ML models that are pretrained on vast amounts of data. These are commonly known as **foundation models (FMs)**.

Traditional forms of ML allowed us to take simple inputs, like numeric values, and map them to simple outputs, like predicted values. With the advent of deep learning, we can take complicated inputs, like videos or images, and map them to relatively simple outputs, for example, if the image contains a cat or not.

With generative AI, you can leverage massive amounts of complex data to capture and present knowledge in more advanced ways—mapping complicated inputs to complicated outputs, like summarizing a long document and extracting the key insights.

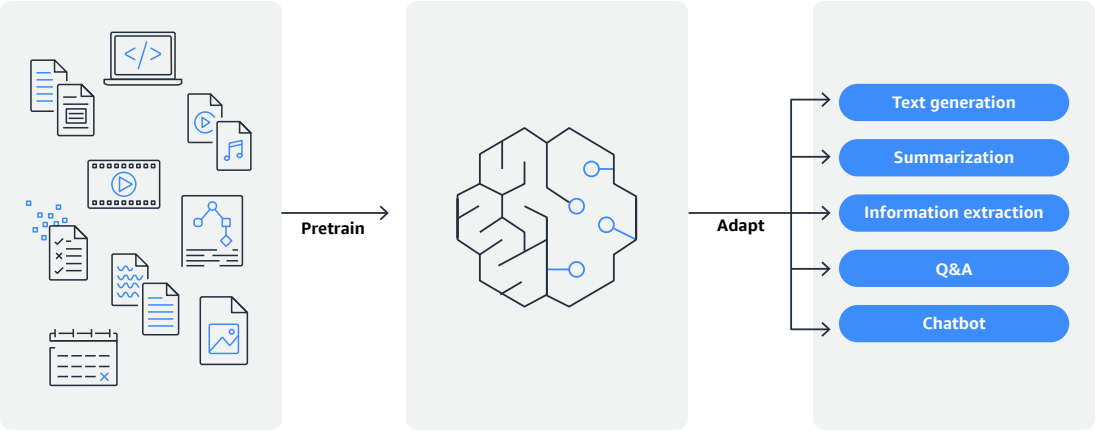
Text-based generative AI systems use a specific type of FM called a **large language model (LLM)**. LLMs can perform a wide range of tasks that span multiple domains, like writing code, creating documentation for developers, enhancing software testing, identifying bugs, and enhancing customer service.

Make data your differentiator

When you want to build generative AI applications that are unique to your business needs, your business's data is your strategic asset. FMs can be customized and fine-tuned with your business's proprietary data to deliver a more differentiated experience compared to an "out-of-the-box FM." For example, a large grocery chain that tracks shopper preferences can customize an FM to produce a better recommendation engine that is highly differentiated from competitors' offerings.

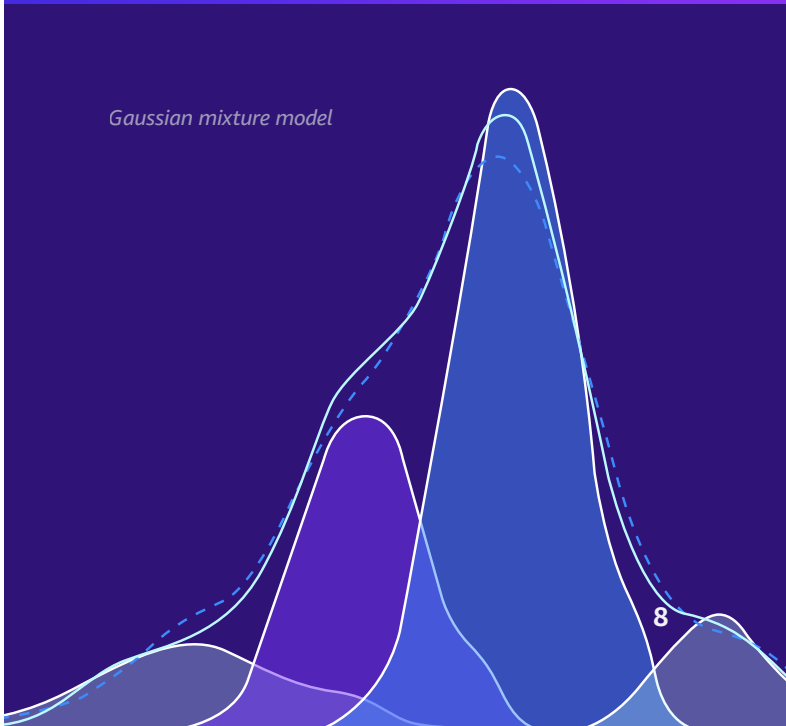
Businesses can also use customized FMs to easily create unique content that embodies their brand's tone and style. For instance, a financial firm that needs to auto-generate a daily activity report for internal circulation can customize an FM with proprietary data, including past reports. The FM could then learn how these reports should read and what data was use to generate them – to deliver a report that better reflects the needs of the business.

Now that you have a general understanding of how the technology works, let's begin exploring how you can put generative AI to work for your business.



Moments in generative AI history:

Today's FMs that are used to create generative AI applications are built atop a long history of AI innovation. Two of the earliest models with generative AI capabilities are the hidden Markov model (HMM) and the Gaussian mixture model (GMM), both developed in the 1950s. HMMs use known data to make educated guesses about unknown data (for example, predicting whether a card player is cheating based on their results). GMMs can examine a group of data (such as a music playlist) and subgroups within that data (for example, genres) to infer unknown information (such as "this is a rap song"). Both are still used today.



Business capabilities of generative AI

Across industries, businesses are using generative AI to enhance productivity and create business value in many ways, including:

Code generation
Improve developer productivity by 57% with AI coding companion **Amazon CodeWhisperer** [6]

Personalization
Improve personalized recommendations and generate tailored content

Optimize back-office tasks
Reduce operational costs, minimize human error, and increase overall efficiency

Content generation
Create text, images, videos, and music

Design and creativity
Increase design variety and innovation, rapid prototyping and faster design cycles, optimize processes and workflows

Virtual assistants
Enhance customer experience with human-like responses

Decision making
Refine strategies, products, and solutions with contextual insights

Moments in generative AI history:

Another early example of generative AI is ELIZA, a chatbot (or "chatterbot," as they were previously known) developed by an MIT professor from 1964 to 1966. Like its namesake, Eliza Doolittle of *Pygmalion* and *My Fair Lady*, the program grew more sophisticated by "learning" from human interactions. ELIZA was most famously used to mimic the behavior of a therapist conducting an initial psychiatric interview, with the user playing the role of the patient.

[6] Data collected from a "productivity challenge," conducted by Amazon during the Amazon CodeWhisperer preview

Business considerations for generative AI

As you work to identify the capabilities of generative AI that are most useful to your company—and develop a strategy for implementing them into your business processes—you will need to determine which FMs to use in creation of generative AI applications to suit different needs.

You should also carefully consider the infrastructure you will be using to support your FMs. Your models will benefit from a cost-efficient infrastructure that meets your requirements for performance.

When evaluating FMs used to create generative AI applications, look for models that offer:

1. Easy ways to build and scale generative AI applications with security and privacy built in
2. Performant, low-cost infrastructure to train your own models and run inference at scale
3. Generative AI-powered applications to transform how work gets done
4. Data as your differentiator

Responsible AI, security, and privacy

With their vast size and open-ended nature, FMs raise new issues in defining, measuring, and mitigating responsible AI concerns across the development cycle, such as accuracy, fairness, intellectual property (IP) considerations, hallucinations, toxicity, and privacy. For example, looking at the issue of fairness, can we ask an LLM to assign male and female pronouns at the same rate in reference to a doctor? Does that still apply if the prompt describes the doctor as having a beard? And should we do the same for other professions? With the rise in popularity of women's football after the 2023 FIFA Women's World Cup, for example, can AI tell the difference between men's and women's teams with the same name? You can see that simply defining fairness in the context of an LLM is challenging and requires new approaches and solutions.

Generative AI technology and how it is used will continue to evolve, posing new challenges that will require additional attention and mitigation. To tackle these challenges and foster innovation, **academic, industry, and government partners** are working together to explore new solutions and concepts to ensure that generative AI continues to evolve in a responsible, private, and secure way.

Data privacy and security are also critical to scaling generative AI responsibly. When it comes time to customize and fine-tune a model, businesses need to know where and how their data is being used. They need to be confident their private data is not being used to train a public model and that customer data remains private. Businesses need security, scalability, and privacy to be baked in from the start to be viable for their business applications.

Read the blog post *Responsible AI in the Generative Era*
Learn more ›

Moments in generative AI history:

In 2014, the development of the first generative adversarial network (GAN) marked one of the biggest breakthroughs in generative AI. In a GAN, two models (a "generator" and a "discriminator") compete in a zero-sum game. The generator manufactures content that appears increasingly "genuine," while the discriminator analyzes its opponent's techniques to better identify fakes. This novel approach of using AI to train other AI proved revelatory, while GANs themselves unlocked a new era for digital imagery.

In the following section, we feature an AWS industry leader who shares his experiences and strategies for cloud best practices, cultural change, organizational agility, and transformation with generative AI.

Insights on generative AI from business leaders

Like any new technology that becomes mainstream, generative AI has a learning curve. Business leaders are asking about where generative AI fits or doesn't, how to use it effectively, and about the non-technical considerations.

History teaches us that we see profound, positive change only when people, processes, skills, and culture are addressed alongside technology. Based on these learnings and from the thousands of customers we talk to, our advice for those intrigued by generative AI is simple.

First, be curious. Learn what generative AI is, why it has captured people's imaginations, and what problems it can solve. Dive deep into areas such as the security of your data when customizing models. And encourage others to learn these things, too, rather than just delegating them to your technical team members.

Second, think big and work backwards from the customer. This is a standard way of thinking here at AWS! Really understand the opportunities in your business—whether it's an opportunity to create efficiencies, improve productivity, invent new apps, or accelerate innovation and time-to-market. Fall in love with the problem before you jump to a solution, looking for areas that reduce costs, increase resilience, or further growth. And think big about the opportunity; small thinking becomes a self-fulfilling prophecy.

Finally, start now. Most business initiatives take time to get traction, so start experimenting quickly. You will learn more from this than the endless planning and waiting for the hypothetical perfect time that is typical of many business technology adoptions.

Phil Le-Brun
Director,
AWS Enterprise Strategy

Joining AWS in 2019, Phil's experience implementing technology at scale—including a 25-year career spearheading digital transformation efforts at McDonald's Corporation—has allowed him to learn a variety of practical lessons. He shares this knowledge with enterprises to help them achieve their cloud-based technology goals, such as supporting organizational agility and increasing customer centricity.

Common generative AI use cases

Engineering productivity

- **Natural language low-code generation:** Integrating natural language interfaces with low-code platforms enables developers to use human language commands or descriptions to define application features, logic, and functionality. Generative AI enhances this process by interpreting and translating natural language input into underlying code components, automating the creation of application elements without extensive manual coding. This approach accelerates development, making it accessible to a broader range of users, including those with limited programming experience.
- **Tools to generate, debug, and test code:** Automate code generation, bug detection, and testing.
  - **Code generation:** Generate code snippets, templates, or even entire modules autonomously, saving time and reducing manual coding errors.
  - **Bug detection:** Analyze code for potential bugs, vulnerabilities, or deviations from coding standards, helping to identify issues early in the development process.
  - **Code Refactoring:** Suggest code refactoring options to improve code quality, readability, and maintainability, assisting developers in optimizing their codebase.

Product development

- **Improve and accelerate the product lifecycle:** Generative AI enables product software developers to unlock new levels of creativity and innovation. With generative AI leveraging large amounts of data, it can identify patterns, generate design variations, and propose solutions that humans may not have considered—leading to breakthroughs in product development. Generative AI also enhances design collaboration by enabling multidisciplinary teams to exchange and explore ideas. It supports easy sharing and refinement of AI-generated designs, fostering inclusive teamwork.
- **Generative design:** Generative AI streamlines product design by quickly producing and assessing design options, leveraging user input and data analysis to enhance and propose optimized solutions. This accelerates development cycles, enabling quicker product launches without compromising quality. As new cases are added to the dataset, generative AI can continuously refine its understanding and summarization capabilities, adapting to evolving trends and needs.

Customer engagement

- **Natural conversation for customer communications:** Generative AI can automate customer query responses and interactions without requiring direct human intervention at every step. These interactions can take place across various channels, such as chatbots, virtual assistants, or email responses. Additional benefits of automated customer communications include: 24/7 availability, instant responses, consistent and standardized customer experience across channels, scalability, freeing up human agents to focus on more complex and value-added interactions.
- **Personalized responses and self-service:** Automated systems can gather and analyze customer data during interactions, enabling businesses to understand customer needs and preferences better. By analyzing customer data, automated interactions can provide tailored responses and recommendations, improving relevance and engagement. It enhances self-service experiences for customers by providing automated and personalized assistance, enabling users to find answers, solutions, and information independently, through interactive FAQs, troubleshooting guidance, personalized recommendations, and process automation. Generative AI ensures accurate responses, supports complex queries, and provides consistent support around the clock, leading to improved user satisfaction and efficient problem-solving.

Data insights

- **Extract information from large amounts of data:** Extract valuable insights from large volumes of data quickly. In software development, data insights from generative AI support identifying code improvements, detecting bugs, optimizing performance, enhancing security, and offering data-driven design recommendations. These insights contribute to more efficient development, improved user experiences, and enhanced software quality.
- **Identify patterns and trends:** Generative AI automates analysis—enhancing accuracy, and offering new perspectives and deeper insights and predictions. Generative AI can identify complex patterns by learning underlying data structures, recognizing hierarchical and non-linear relationships, analyzing multiple dimensions and time-series data, detecting anomalies. It adapts its understanding over time and is valuable for business decision making, better strategies, and a deeper understanding of various domains—including insight into potential for new revenue streams. By leveraging generative AI to find patterns in data, software developers can make more informed decisions, optimize code, enhance user experiences, and address potential issues, ultimately leading to more efficient and effective software development.

Buying experience

- **SEO-optimized copy for landing pages, blogs, and social media posts:** Optimize SEO (Search Engine Optimization) copy for digital content in a number of ways. It can generate relevant keywords and phrases that have high search volume and relevance. It can also generate additional content sections, paragraphs, or bullet points that expand upon key topics, enhancing the depth and breadth of the content. SEO-friendly meta titles, meta descriptions, and header tags can also be generated.
- With **Natural Language Integration**, keywords and phrases can be integrated naturally into the content, avoiding keyword stuffing and making the content more reader-friendly. AI can suggest an optimal content structure that follows best practices for readability and SEO, such as using subheadings, bullet points, and numbered lists.

Human resources

- **Deliver great employee experiences:** Generative AI enhances employee experiences in software development companies by offering personalized support, fostering professional growth, and improving work processes, including personalized learning and development, and enhanced onboarding. Training materials, videos, and interactive modules tailored to individual learning styles will help new hires ramp up quickly.
- **Project Assistance:** AI-powered tools can provide guidance and suggestions during the software development process, enhancing the quality of work and reducing errors.
- **Reduced Administrative Tasks:** Automate routine administrative tasks, freeing up employees' time to focus on more creative and strategic aspects of their roles.
- **Enhanced Collaboration:** Promote collaboration by assisting in communication, suggesting suitable team members for projects, and improving overall team dynamics.

Moments in generative AI history:

The 2017 introduction of a new type of deep learning model, the transformer, set the stage for modern generative AI. Unlike older models, which break down input data, process it, then put the pieces back together, transformers process the entire input all at once. This makes them ideal for natural language processing (NLP), where understanding the full context of the input is critical.

In 2018, OpenAI took the technology further by creating the first Generative Pre-trained Transformer (GPT). From there, OpenAI developed its GPT-2 engine in 2019—which it then used to power ChatGPT, introduced in late 2022.

How AWS can help you succeed with generative AI

You can unlock the full business value of generative AI for your business with AWS. Reinvent your applications, create entirely new customer experiences, drive unprecedented levels of productivity, reduce operational costs, and ultimately transform your business.

Experience and expertise

One of the key advantages of AWS lies in a rich AI heritage built over two decades of focused investment. In fact, more than 100,000 customers currently use AWS for AI.

Amazon, the driving force behind AWS, harnesses ML capabilities to power its ecommerce recommendations engine, optimize robotic picking routes in fulfillment centers, and much more. Further, ML informs Amazon's supply chain, forecasting, and capacity planning.

Deep learning is also employed in the Amazon Prime Air drone delivery system and the computer vision (CV) technology behind Amazon Go, the innovative retail experience that allows customers to select items and leave the store without traditional checkouts. And Alexa, which is supported by more than 30 different ML systems, helps customers with a wide array of tasks billions of times each week.

With thousands of dedicated ML engineers, AI and ML are deeply ingrained in the heritage of Amazon and AWS—continuing to shape our future.

More than

100,000

customers currently use AWS for AI

Why build with AWS?

Companies of all shapes and sizes choose to build generative AI and other AI and ML applications on AWS for many reasons. Here are some of the top advantages of building on AWS, according to our customers:

The easiest way to build and scale generative AI applications with security and privacy built in

Amazon Bedrock is the easiest way for customers to build and scale generative AI–based applications using FMs. Bedrock makes **Amazon Titan** FMs and models from leading AI companies such as AI21 Labs, Anthropic, Cohere, Stability AI, and Meta accessible via an API. Customers using Bedrock can leverage the benefits of AWS, which is architected to be the most flexible and secure cloud computing environment available today. **Agents for Amazon Bedrock** is a fully managed capability that makes it easier for developers to create generative AI applications that can deliver up-to-date answers based on proprietary knowledge sources and complete tasks for a wide range of use cases.

The most performant, low-cost infrastructure for generative AI

For years, AWS has invested in developing silicon that delivers the highest levels of performance and cost optimization for AI and ML workloads. The results—**AWS Trainium** and **AWS Inferentia**—deliver the lowest costs for training models and running inference in the cloud. AWS has also developed **Amazon Elastic Compute Cloud** (Amazon EC2) instances to help you take advantage of these capabilities. For example, **Amazon EC2 Trn1** instances powered by Trainium save you up to 50 percent on training costs,[7] while **Amazon EC2 Inf2** instances powered by AWS Inferentia2 deliver up to 40 percent lower cost per inference.[8]

Data as your differentiator

With AWS, it's easy to use your business's data as a strategic asset to customize FMs and build more differentiated experiences. Data is the difference between a general generative AI application and one that truly knows your business and your customer. And with the most comprehensive set of data and AI services, you can securely customize an FM on AWS with your data and build a model that is an expert on your business, your data, and your customers.

Generative AI-powered applications for the enterprise to transform how work gets done

AWS is building powerful new applications that transform how our customers get work done with generative AI. Boost productivity with purpose-built conversational agents that streamline coding in the enterprise with **Amazon CodeWhisperer**, simplify business intelligence with **Amazon QuickSight Generative BI**, and improve clinical efficiency for healthcare organizations with **AWS HealthScribe**. With security, privacy, and responsible AI at the forefront, easy customization, and integration into your existing data sources and applications, enterprises can quickly take advantage of generative AI without the heavy lifting.

Further reading on responsible AI:

AWS responsible AI resource hub ›

eBook: *Democratized, Operationalized, Trusted: The 3 Keys to Successful AI Outcomes* ›

[7] Over other comparable Amazon EC2 instances
[8] Compared to prior generation AWS Inferentia–based instances

AWS generative AI services

Facilitate your generative AI applications with a range of AWS technologies, including:

Amazon Bedrock ›

Build and scale generative AI applications with FMs. Bedrock supports a variety of FMs, including:

- **Amazon Titan**: For text summarization, generation, classification, open-ended Q&A, information extraction, embeddings, and search
- **AI21 Labs Jurassic-2 Multilingual LLMs**: For text generation in various languages
- **Anthropic Claude 2**: LLM for conversations, question answering, and workflow automation based on research into training honest and responsible AI systems
- **Stability AI Stable Diffusion**: Generates unique, realistic, high-quality images, art, logos, and designs
- **Cohere Command + Embed**: Text generation model for business applications and embeddings model for search, clustering, or classification in over 100 languages
- **Meta Llama 2**: Pretrained and fine-tuned LLMs for natural language tasks like question and answering and reading comprehension

AWS Trainium: Train models faster with up to 50 percent cost savings[9] using this ML model accelerator

AWS Inferentia2: Run high-performance FM inference with up to 40 percent lower cost per inference using this accelerator[10]

Amazon CodeWhisperer: Enjoy 57 percent faster application development[11] while helping to ensure security with this AI coding companion, which is at no cost for individual use

Amazon QuickSight Generative BI: Transform traditional multistep business Intelligence (BI) tasks into intuitive and powerful natural language experiences with Generative BI capabilities in Amazon QuickSight

Amazon SageMaker: Build your own FMs with managed infrastructure and tools to accelerate scalable, reliable, and secure model building, training, and deployment

Amazon SageMaker JumpStart: ML hub that provides access to algorithms, models, and ML solutions so you can quickly get started with ML. With SageMaker JumpStart, ML practitioners can choose from a broad selection of **publicly available FMs**. ML practitioners can deploy FMs to dedicated SageMaker instances from a network-isolated environment and customize models using SageMaker for model training and deployment

[9] AWS Trainium delivers up to 50 percent cost-to-train savings over comparable Amazon EC2 instances
[10] AWS Inferentia delivers up to 40 percent cost per inference over comparable Amazon EC2 instances
[11] Data collected from a "productivity challenge," conducted by Amazon during the Amazon CodeWhisperer preview

Next steps

Now that you have a better understanding of generative AI, what it can do, and its potential business benefits, the next step is to clearly define your objectives and identify use cases for leveraging it. It's best to start with smaller experiments and simple, precise goals. Once you've achieved some quick wins, you can begin scaling your efforts upward and outward.

Collaboration with experts is highly recommended to ensure you consider factors such as data availability, data quality, and ethical implications related to generative AI. Furthermore, infrastructure considerations should not be an afterthought, as they can significantly impact costs, scalability, and energy consumption. Engaging with AWS experts can provide valuable guidance throughout the decision making process and stages of implementation.

The time is now

The dramatic rise of generative AI brings us to a tipping point. FMs grow more sophisticated and powerful every day. For companies it's the power to transform business by creating entirely new customer experiences, with enhanced or new solutions, and driving unprecedented levels of efficiency and productivity.

All of which leads to an indisputable fact: To compete in this new era of profound technological advancement, every business, and particularly software companies, need to consider making generative AI a part of its innovation road map.

With the most cost-effective cloud infrastructure for generative AI; a host of AI products, services, and solutions; and years of trusted AI expertise, AWS can help turn the promise of generative AI into results for your organization.

Partner with AWS to accelerate your generative AI journey today.

Get started ›