

Moving Forward with Machine Learning and Generative AI

Learning Objectives

After completing this chapter, you will be able to:

- 4.1 Explain the relationship between big data and machine learning, and how this growing trend is impacting various fields.
- 4.2 Trace the major developments in the history of language models, explaining the key concepts and innovations that led to the creation of powerful language models.
- 4.3 Explain the strengths and weaknesses of ChatGPT.
- 4.4 Explain the functionalities of other large language models (LLMs) in an industrial setting.
- 4.5 Explain how AI is integrated into mobile apps to enhance the user experience.
- 4.6 Analyze how AI is being used in the Apple Intelligence Platform.

Data is at the heart of modern business today. It helps businesses understand customers, make better decisions, manage inventory, and monitor competitors. Over the last 20 years, companies have had to handle increasing amounts of diverse data, known as big data. By analyzing vast amounts of information, a company can gain valuable insights and make decisions based on the trends uncovered from the data. Although big data provides a vast reservoir of information, it is like having a library full of books without knowing how to read them. The raw data itself does not automatically reveal trends or patterns. This is where machine learning comes into play. Machine learning can automatically spot patterns across large volumes of data and even learn from its successes and mistakes, giving organizations the ability to take their big data analytics initiatives to the next level.

Netflix, a popular movie streaming service, is a prime example of a company that analyzes massive amounts of data. Netflix uses big data and machine learning to deliver personalized recommendations (**Figure 4-1**). The company collects vast amounts of data about its users, including viewing history, ratings, language, location, search history, and device information. Using machine learning algorithms, Netflix analyzes this data

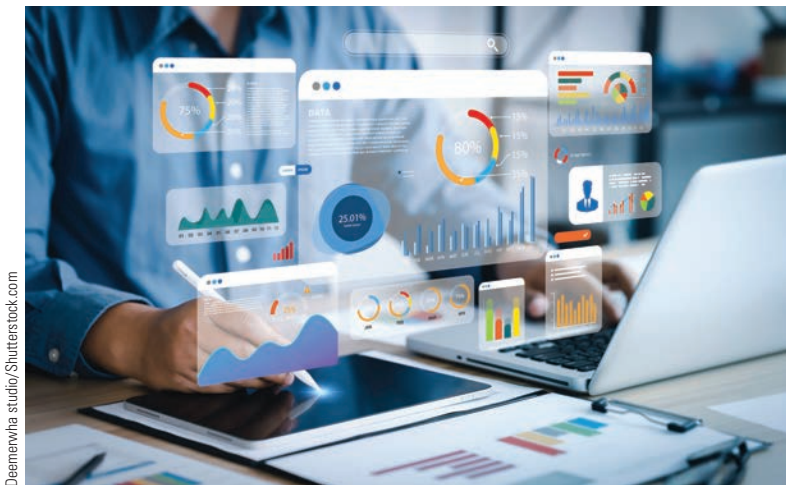
Figure 4-1: Netflix uses big data and machine learning to find your favorite new shows



to identify patterns and to suggest shows and movies that a user is likely to enjoy. For example, the algorithms might discover that users who enjoyed *Stranger Things* also liked *Dark*. Netflix can determine the best placement for specific content on its platform, increasing the likelihood that users will discover and watch it. This personalized approach enhances the user experience and informs Netflix's content creation decisions.

4.1 The Rise of Big Data and Machine Learning

Figure 4-2: Every business sector collects and analyzes big data



The era of big data and machine learning is usually associated with modern artificial intelligence (AI) development; however, its roots are in the Internet's widespread adoption in the 1990s, which boosted data growth. As more people connected to the Internet, they generated vast amounts of data through online activities such as browsing websites, using email, shopping online, and participating in social media. This data was stored on servers around the world, creating an enormous digital repository. Businesses and governments began to collect and store data on a large scale and were aided by the introduction of inexpensive cloud data storage and more powerful computer hardware. Today this data is used for a variety of purposes, including market research, scientific analysis, healthcare patient diagnoses, financial services, inventory management, and business decision-making (**Figure 4-2**).

The Emergence of Big Data and Data Centers

The term “big data” began to gain traction in the early 2000s. Big data initially referred to extremely large **datasets** that are too complex to process using traditional data management tools. A dataset is basically a collection of information. Imagine you have a notebook where you write details about all the places you would like to travel, including the destination, activities, local foods, and places to stay. Each page in your notebook represents a collection of information, and when you put all these pages together, you have a dataset. If you had a page about a trip to Paris, the Eiffel Tower, crepes, and a fun place to stay named Ducks Eiffel Tower Hostel, that would be one row in your dataset (**Figure 4-3**). When you add more trip ideas, you build a larger dataset of your travel history in a predefined format called **structured data**, which would be organized by destination and other information. The

structured data is typically stored in a database, where each piece of information is assigned to a specific field or column.

When these datasets become exceptionally large, complex, and difficult to process using traditional data management tools, they are classified as big data. Big data refers to datasets that meet these five key criteria:

1. **Volume:** The sheer amount of structured data is massive, often measured in petabytes (one quadrillion bytes).
2. **Velocity:** Data is generated and collected at a rapid pace, making it challenging to keep up to date.
3. **Variety:** Data comes in many formats, including text, numbers, images, videos, and more.
4. **Veracity:** This criterion refers to the data's accuracy ("truthfulness") and how much it can be relied on.
5. **Value:** Although all business data has value, the true worth of data must often be uncovered by techniques that expose its hidden value.

All of this big data must be stored in computer farms or on computer servers located in one or more warehouse locations called **data centers**. A data center (**Figure 4-4**) is a physical facility designed to house and operate computer systems and related components, such as servers, storage devices, networking equipment, and security systems. Data centers provide the necessary infrastructure, including considerable amounts of power, cooling, and security, to support the operation of computer servers. These centers offer the infrastructure necessary to handle the five characteristics of big data.

Data centers can be categorized into two main types: on-premises and cloud. **On-premises data centers** are owned and operated by an individual organization such as a corporation like Meta; a government data center housing tax records and national security information; or a university supporting research, teaching, and student data. They provide complete control over the infrastructure, but require significant upfront investment and ongoing management fees. The size and complexity of these data centers vary depending on the needs of the organization. A **cloud data center** is operated by a cloud service provider, offering various services over the Internet to millions of customers. One of the largest cloud service providers in the world, Amazon Web Services (AWS), operates a huge network of data centers across the globe (**Figure 4-5**). Although Amazon is most well known for its online retail store, AWS is the cloud platform offering a wide range of services to businesses that rely on powerful data centers to store and process the data for AWS services.¹ Other major cloud service providers include Microsoft Azure, Google Cloud Platform, Alibaba Cloud (a major player in the Chinese retail industry), IBM Cloud, Oracle Cloud Infrastructure, and Salesforce Cloud.

AWS has a vast array of services used by companies across various industries. **Table 4-1** contains several examples of how major companies have leveraged AWS.

Figure 4-3: Dataset created by a collection of information such as a list of dream trips

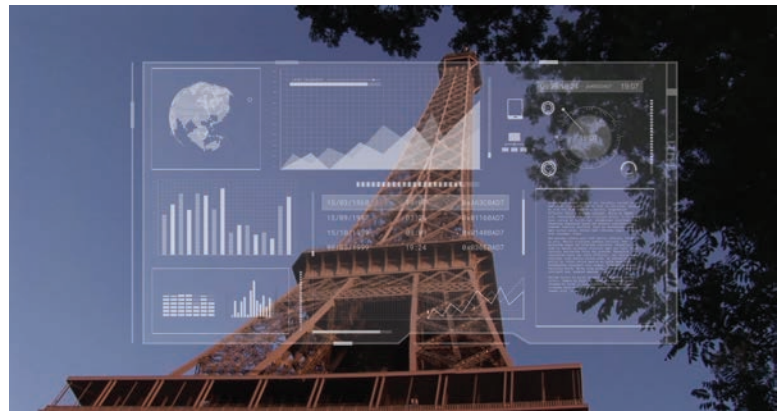


Figure 4-4: Data center servers

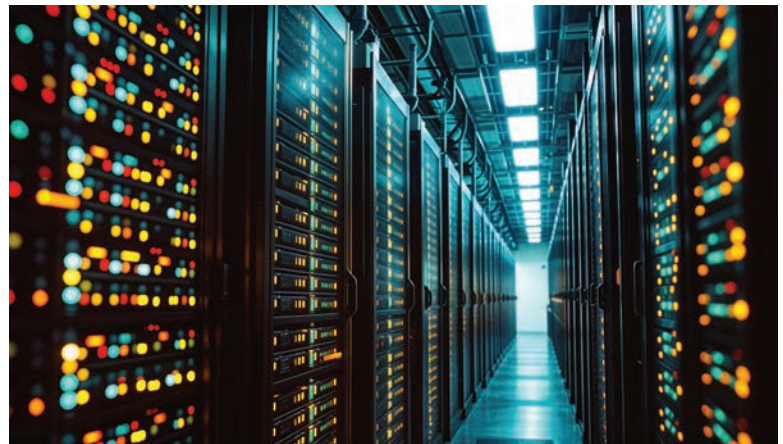


Figure 4-5: Amazon Web Services website

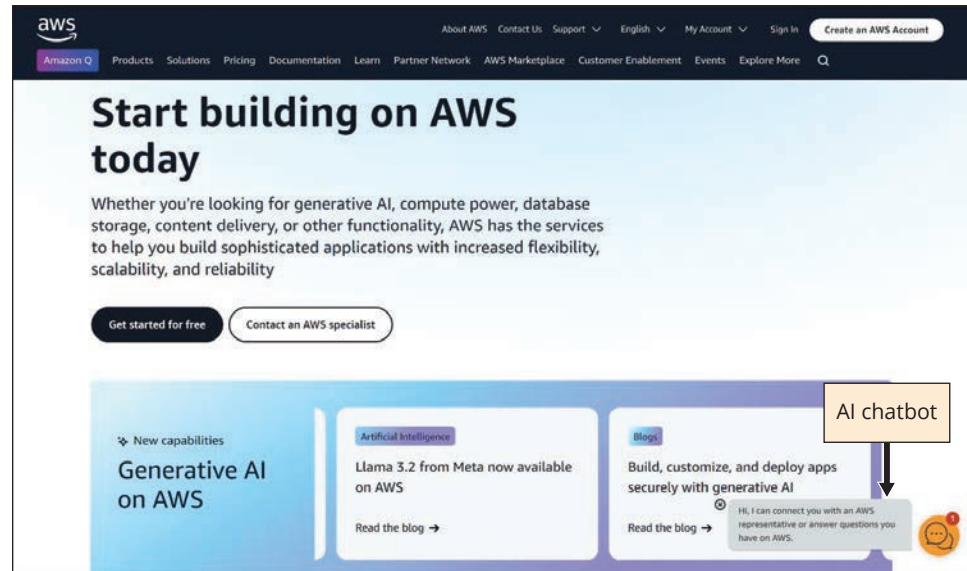


Table 4-1: AWS services for major companies

Company	Service category	AWS services
Amazon online store	E-commerce service	Amazon is a major user of its own AWS platform. In fact, AWS was originally built to support Amazon's growing e-commerce business. Canvas, a popular LMS, uses the AWS platform.
Lyft	Ride service	Lyft uses AWS to manage its ride-hailing platform, including real-time location tracking and payment processing.
Airbnb	Home sharing service	Airbnb leverages AWS for its database of homes and rooms, dynamic pricing, search functionality, and global operations.
Hulu	Streaming media service	Hulu uses AWS to stream movies and TV shows globally, scaling its infrastructure to meet peak demand.
Walmart	Physical stores and e-commerce services	Walmart employs AWS for its e-commerce platform, supply chain management, and data analytics.
Sephora	Physical stores and e-commerce services	Sephora leverages AWS for its beauty subscription service, inventory management, and customer analytics.
Capital One	Financial services	Capital One uses AWS for its cloud-native banking platform, fraud detection, and customer analytics.
Cerner	Medical records services	Cerner uses AWS to power its electronic health record (EHR) systems, medication, and healthcare analytics.

Geographic location can influence the size of a data center due to factors such as power availability, climate, and connectivity. Although data centers can be found in various regions around the world, certain areas have the greatest concentration, including northern Virginia in the United States, as it is close to the capitol in Washington DC; Dublin, Ireland, due to its favorable tax regime in Europe; Singapore, as the major financial hub of Asia; and other major hubs including Sydney, Australia, Sao Paulo, Brazil, and Johannesburg, South Africa. Many data centers are located in remote areas (**Figure 4-6**) to minimize the risk of natural disasters, such as earthquakes, floods, and hurricanes. These areas often have lower land costs and fewer regulatory restrictions than more populated areas. To protect sensitive information, data centers are surrounded by high-security fences, often without signage to deflect unwanted attention.

Data centers require significant amounts of power. They are often located near power plants to ensure a reliable and efficient power supply. Presently, there are over 10,000 data centers around the world, and more are being built every day, driven by factors such as increasing AI and the growing demand for data storage and processing capabilities.

A recent trend has seen Microsoft, Google, and others partner with companies to reopen closed nuclear power plants to provide a clean source of energy for their data centers. Microsoft has signed a 20-year agreement with Constellation Energy to reopen the Three Mile Island power plant in Pennsylvania, which was the site of a commercial nuclear accident in the 1970s. Restarting one unit of this plant will produce 800 megawatts of power, which is more than the amount of hydroelectric power supplied by the Hoover Dam.

Figure 4-6: Data center in a rural area



Make more Aerials/Shutterstock.com

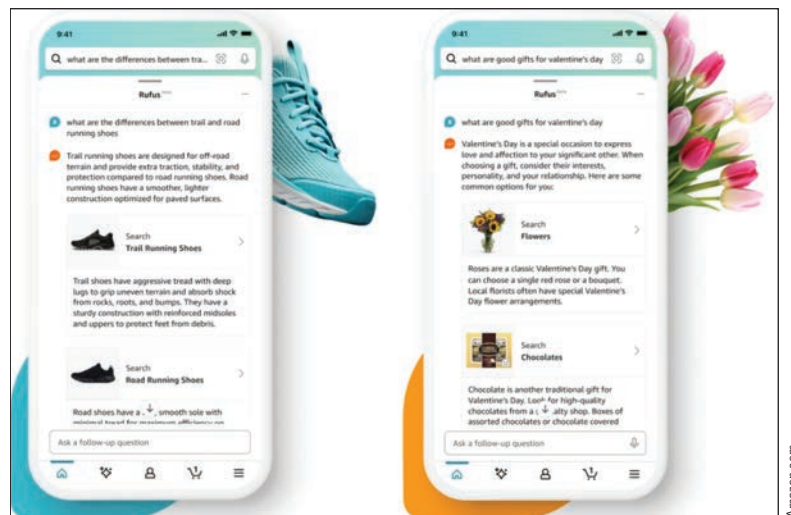
The Intersection of Big Data and Machine Learning

Although data storage is a crucial component of handling big data, it alone does not address how machine learning helps to process all that information. The real value of big data comes from analyzing it to extract meaningful insights. Many big data applications, such as fraud detection or product recommendation systems, require real-time data processing with machine learning. These applications demand not just storage but also high-speed data processing infrastructure. For example, when you shop on Amazon.com, an AI assistant named Rufus can help with your shopping. You use a conversational interface to interact with Rufus to find products, evaluate options, and receive personalized recommendations. A customer can ask Rufus to compare products, shop by occasion, or even recommend what to look for while shopping for specific products, like identifying what to consider when buying headphones. Rufus can also answer questions about specific products, such as “Is this cosmetic tested on animals?” or “Is this jacket machine washable?” In **Figure 4-7**, shoppers asked Rufus questions including “What are the differences between trail and road running shoes?” and “What are good gifts for Valentine’s Day?”

Rufus uses Amazon’s proprietary machine learning models to power features like recommendations and product searches. The machine learning starts when a user asks a question. Rufus searches its knowledge base for relevant information to respond to the question. The search involves using techniques like keyword matching and **semantic searching** to find the most relevant information. A semantic search goes beyond matching keywords. Instead of just finding documents that contain the exact words in the question, semantic search aims to find documents that are relevant to the underlying meaning or intent of the question.

Another use of big data and machine learning is creating positive customer relationships in business. **Customer relationship management (CRM)** is a business strategy that involves using technology and processes to collect, store, and

Figure 4-7: Amazon’s Rufus is an expert shopping assistant that answers customer questions

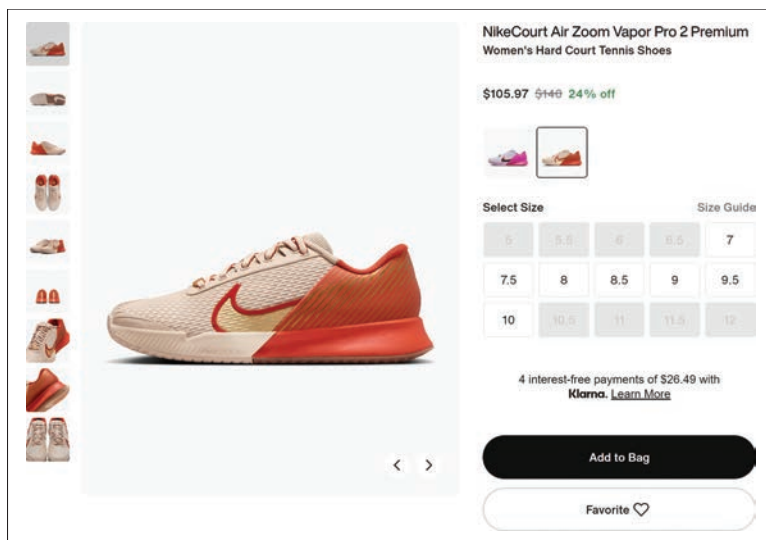


Amazon.com

Figure 4-8: The components of CRM create a positive relationship with customers



Figure 4-9: CRM creates a customer campaign to buy Nike products



analyze customer data to understand customer needs and preferences better (**Figure 4-8**).

Nike, a sportswear clothing retailer, uses a CRM system to track customer purchases, preferences, and interactions with the company. It can use the customer data to send personalized email campaigns, recommend products, and provide targeted customer support. Suppose you have shopped with Nike in the past and purchased a pair of Nike Air Zoom Vapor tennis shoes (**Figure 4-9**). Nike's CRM system would likely track this purchase and record the following information in their data center:

- Product details: The specific model of Nike Air Zoom Vapor tennis shoes you bought
- Purchase date: When you made the purchase
- Purchase location: Whether you bought the shoes online or in a physical store
- Customer information: Your name, address, email, and other relevant details

After the details of the Nike shoe information is saved to the servers in the data center, the machine learning process powers the CRM. The CRM suggests new Nike Air Zoom tennis shoes in other colors and newer models that you might be interested in based on your past purchase. The CRM sends you targeted emails about related Nike products matching your color scheme, new shoe releases, sales, and other exclusive offers like reward points. Additionally, the CRM can handle customer service issues. If your shoes need to be fixed or replaced, Nike can use your purchase history to quickly identify the product and provide appropriate assistance with a customer chatbot. In this way, the machine learning in its CRM system

helps Nike build a stronger relationship with you as a customer, provide personalized service, and increase brand loyalty.

4.1 | Two Rights & a Wrong

Read the following three statements and then identify the incorrect one.

1. Data centers provide the necessary infrastructure, including massive amounts of power, cooling, and security, to support the operation of computer servers.
2. The era of big data is usually associated with modern AI development and advanced with the Internet's widespread adoption in the 1990s.
3. Big data refers to datasets that meet multiple criteria including volume, velocity, and vortex.

□ Answers at end of chapter

4.2 Moving Forward with Large Language Models from Eliza to ChatGPT

Although data centers provide the infrastructure for processing machine learning algorithms, another tool called large language models (LLMs) has been developed to harness these capabilities for natural language understanding and generation. LLMs are trained by “reading” vast sources of content such as books, articles, computer code, websites, and social media posts and by “viewing” photographs, paintings, films, and artwork. The results generated by LLMs are comparable to those produced by a human who has a thorough understanding of language, grammar, imagery, and storytelling.

LLMs can identify common grammatical rules, sentence structures, and word usage patterns. For example, if you ask an LLM about the capital of Japan, the LLM might encounter the word *Japan* in various contexts throughout its training data:

- First context of Japan: Tokyo is the capital of Japan.
- Second context of Japan: Japanese food, like sushi and ramen, is very popular.
- Third context of Japan: The Japanese culture is known for its respect for tradition and its focus on harmony.

The LLM learns to associate *Japan* with the country, its capital of Tokyo, its food, and its culture. When asked “What is the capital of Japan?” the LLM can correctly answer Tokyo. LLMs can also learn from datasets of images paired with captions, helping them connect visual features and text. If you ask about the capital Tokyo, the LLM could also display an image (Figure 4-10). Image training helps LLMs learn the relationship between visual features and textual descriptions.

Figure 4-10: LLMs can associate Tokyo with related text and images



Chay_Tee/Shutterstock.com

Very Early LLMs

LLMs have evolved significantly since they were introduced. One of the earliest examples, Eliza, developed by Joseph Weizenbaum in 1966, could mimic human conversation using pattern matching and substitution. Picture this: a computer program that chats with you, like a therapist. Eliza could have a very simple conversation (Figure 4-11). For example, if you say, “I’m feeling sad,” Eliza might respond with “Why do you say you’re feeling sad?” Eliza did not really understand the meaning of the words that you spoke. It simply matched patterns and responded based on predefined rules. Though limited in handling complex questions, Eliza set the stage for advancements in natural language processing (NLP) and LLMs. The program SHRDLU, created in the 1970s, was the next step in the journey. It could comprehend entire sentences, discuss a fictional world, and even clarify any uncertainty in your questions. Consider it as a chatbot capable of maintaining a more coherent conversation.

The 2000s were a game-changer for more sophisticated language models. With more digital data and faster computers, scientists began working on a type of machine called a language model (LM). These LMs were like a dictionary for computers. They knew about words and how they fit together to form sentences. An LM had a large list of words

Figure 4-11: Eliza can hold a conversation in English

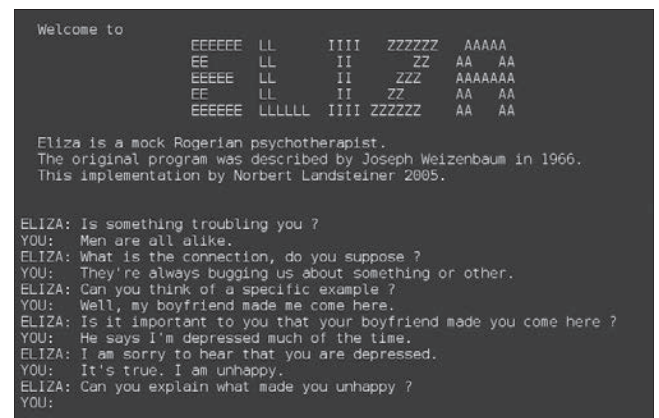
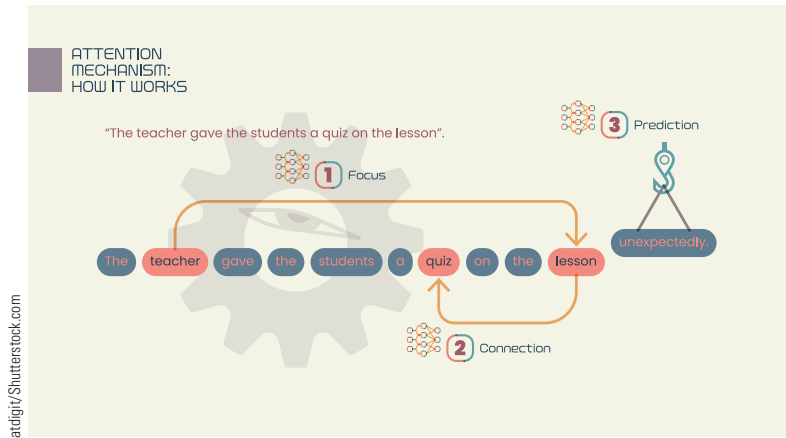


Figure 4-12: Transformer model



and knew some basic rules about how words fit together. For example, it knew that because a noun usually comes after an article like *the* or *a*, the word *house*, a noun, would come after *the* or *a*. LMs could recognize words and patterns, but they did not understand their meaning. It was like a parrot who is taught to speak words without knowing what they mean. Even though they were limited, these early LMs were still a big step forward.

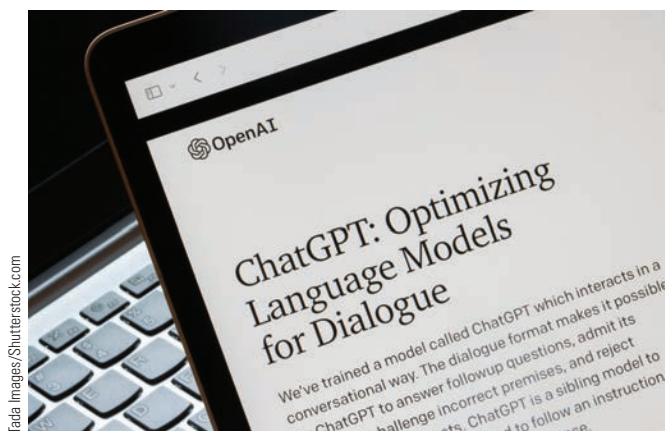
As computational power and access to large datasets grew, researchers began experimenting with larger and more complex language models. In 2017, researchers established **transformers** (no, not the robots). A transformer model can handle much more complex sentences than earlier LMs,

and it considers text before and after words to understand their context. When the transformer model reads the sentence, "The teacher gave the students a quiz on the lesson," (Figure 4-12), it follows these key steps:

1. **Encoder:** In the encoder step, the transformer reads the entire input text and breaks it down into smaller pieces. ("The teacher gave the students a quiz on the lesson.")
2. **Decoder:** In the decoder step, the transformer tries to generate the output text (like a translation or summary).
3. **Attention mechanism:** The attention mechanism is the key to how transformers work. The transformer model uses an attention mechanism to weigh the importance of different tokens in relation to each other (Figure 4-12). For example, the model might assign more weight to the words "teacher," "quiz," and "lesson" as they are likely more important for understanding the meaning of the sentence. This step captures the semantic meaning of the sentence, which can be used for various tasks, such as translation, summarization, or question answering.

ChatGPT-3 in the 2020s

Figure 4-13: ChatGPT-3 is one of the largest LLMs ever created, with billions of datasets



The 2010s and 2020s saw a rapid acceleration in LLM development, fueled by leaps in hardware processor speed in data centers, improved algorithms, and data availability. Built on the transformer model, more powerful LLMs were capable of generating human-quality text, translating languages, writing different kinds of creative content such as original poetry, and answering questions in an informative way. In 2020, the company OpenAI developed ChatGPT-3, one of the most advanced LLMs to date (Figure 4-13). ChatGPT-3 continues moving into newer, more advanced versions today, training on massive datasets to engage in informative and engaging conversations. ChatGPT-3's capabilities have inspired researchers to explore new applications for LLMs, such as drug discovery, code generation, educational tools, and business AI. This has led to a surge of innovation and creativity in the field.

4.2 | Two Rights & a Wrong

Read the following three statements and then identify the incorrect one.

1. LLMs can identify common grammatical rules, sentence structures, and word usage patterns.
2. One of the earliest examples of LLMs, Eliza, developed by Weizenbaum in 1966, could mimic human conversation using pattern matching.
3. When a transformer model processes a sentence, it completes encoder, decoder, and reporter steps.

□ Answers at end of chapter

4.3 ChatGPT Strengths and Weaknesses in Conversational AI Applications

ChatGPT is a type of conversational AI-based LLM, specifically trained to understand and then generate human-like text. How did all the previous innovations merge with newly developed AI technologies to lay the foundation for ChatGPT to become the fastest-growing consumer app in history shortly after its release in late 2022? By January 2023, it had attracted over 100 million active monthly users within just two months, surpassing the growth rates of other highly popular apps like TikTok and Instagram, which took much longer to reach similar milestones. This LLM has been trained on immense amounts of text data, allowing it to understand and respond to a wide range of prompts and questions.

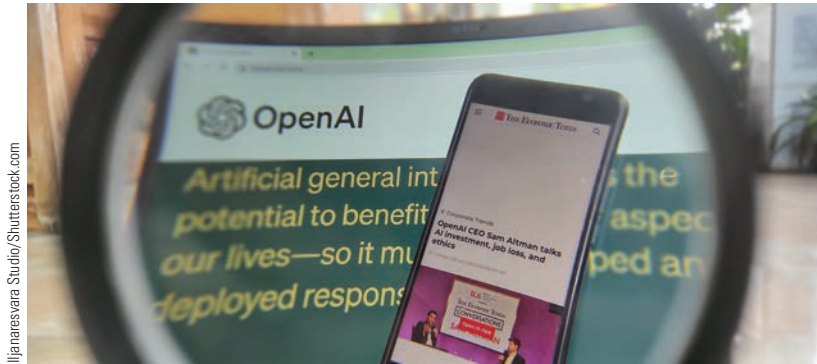
ChatGPT-3 Strengths

ChatGPT is built using transformers. The key foundation is the attention mechanism, which allows it to understand relationships between words in a sentence, regardless of their position. To protect proprietary training methods, Sam Altman, the CEO of OpenAI, the company that developed ChatGPT, does not share exactly which information was used to train their LLM. However, the dataset is massive, consisting of billions of words from various sources, including books, articles, code repositories, and more. The strengths of ChatGPT include its conversational responses in areas such as the following:

- **Customer service:** ChatGPT chatbots can address typical customer questions, offer assistance, and even solve customer problems.
- **Virtual assistants:** Personal assistants like Siri, Google Assistant, and Microsoft Voice Access use ChatGPT to understand and respond to voice commands.
- **Content creation:** ChatGPT can help with tasks such as writing articles, creating tables, generating creative content, holding conversations, and even translating languages.
- **Education:** ChatGPT can be used to create personalized tutoring experiences and answer student questions.

ChatGPT-3 Limitations

In its first public release, ChatGPT-3 amazed people with its strong language generation abilities, but it also demonstrated some limitations. GPT-3 lacked both an in-depth reasoning training and a deep semantic understanding in its responses. The factual accuracy lagged at times; the term “hallucination” basically became mainstream because the tool made up information. At times, GPT-3 even had difficulty with basic factual questions in fields such as science and elementary math.

Figure 4-14: ChatGPT-3 ethical issues and bias

Ethical bias in ChatGPT-3 and other AI models refers to the unintentional inclusion of biased or prejudiced language or behaviors in its responses. ChatGPT-3 sometimes generated responses that reflected discriminatory views. Because the model was trained on a dataset that included biased language, it reproduced the biases in the training data. When asked about ethical questions, ChatGPT provided answers based on the cultural norms from its learned data, potentially presenting culturally biased viewpoints (**Figure 4-14**). Even though OpenAI attempts to filter out offensive or biased material,

the sheer volume of training data makes it challenging to eliminate all harmful content.

GPT-5 and Beyond

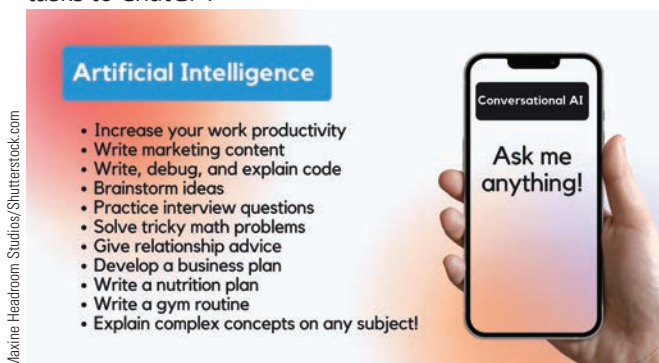
Going forward, ChatGPT and other AI tools have evolved in three major areas to improve their responses. First, the amount of training data has increased exponentially with each version by exposing ChatGPT to a wider range of text and visual data, including books, articles, code, and images, so it has been able to expand its knowledge base. By training on a more diverse dataset, ChatGPT can be less susceptible to biases in the original training data. Second, AI developers boosted the number of language parameters within the neural network model of LLM. These parameters are like synapses connecting neurons in your brain. Each neuron is like a tiny computer, and the synapses between them help the brain learn and understand information. With more connections, ChatGPT-5, the next generation of ChatGPT, could better understand the subtle differences in meaning among words and phrases. For example, it could tell the difference between “I’m pleased” and “I’m ecstatic.” The parameters are the instructions you would give a computer to learn chess moves, like “a queen can move any number of squares diagonally.”

In addition to improvements to more datasets and parameters, the newest versions of ChatGPT have added **supervised learning methods**. Supervised learning is a type of machine learning where a model is trained on a dataset of prompts and their corresponding desired results. This is like teaching a child to recognize animals by showing them pictures of dogs, cats, and frogs and labeling each one. Supervised learning can be used to train ChatGPT on specific tasks, such as summarizing a business interview or writing marketing content (**Figure 4-15**).

These improvements in AI responses have produced some unintended consequences. When LLMs did not know the answer to a question, they typically responded with “I don’t know.” Human evaluators downgraded these types of responses. The AI LLMs then “learned” to produce more convincing-sounding answers—though not necessarily more correct answers. The human evaluators would consider these more convincing answers as correct (although they were not)

and would not downgrade the LLMs. Not only did the answers trick the human evaluators but they also tricked the users who entered the prompt into thinking the answer was correct because it sounded convincing—though the answer was still wrong.

This is also a reminder that everything an LLM says may not be correct, even if it sounds convincing.

Figure 4-15: Supervised learning models teaching new tasks to ChatGPT

4.3 | Two Rights & a Wrong

Read the following three statements and then identify the incorrect one.

1. Copilot has become the fastest-growing consumer app in history.
2. ChatGPT is built using transformers, which understand the relationship of words in a phrase regardless of their order.
3. In its first public release, ChatGPT-3 impressed people with its strong language generation abilities, but it also demonstrated some limitations.

□ Answers at end of chapter

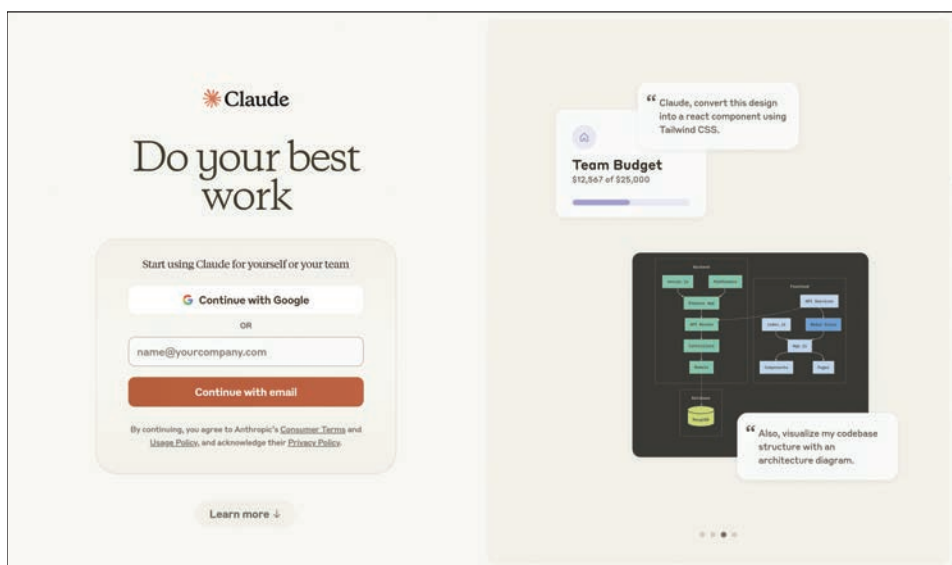
4.4 Other High-Functioning LLMs: Use Cases and Case Studies

In addition to ChatGPT, Gemini, Copilot, and a variety of other chatbots are emerging that are designed for specific tasks or industries. In today's rapidly evolving LLM landscape, AI is making significant strides in changing how business works. From healthcare to finance, education to entertainment, AI-powered applications are transforming the way people work and live.

Meet Claude

Claude AI (claude.ai), developed by a company named Anthropic, is a powerful LLM with a wide range of potential applications. Claude has a free version and a paid professional subscription service, both available on its website. Claude states that it is trained to be a helpful, honest, and harmless assistant with a conversational tone (**Figure 4-16**). It is designed to be a general-purpose LLM, capable of tasks such as generating text, translating languages, writing creative content, and answering questions in an informative way.

Figure 4-16: Claude LLM displays its conversational tone



Claude is known for prioritizing safety and efficiency, offering a unique approach to prompt handling that distinguishes it from other language models and makes it less likely to generate harmful or misleading content. Claude's distinct strengths and approaches to prompt handling include:

- Sourcing information publicly available via the Internet
- Obtaining information from datasets that Claude legally licenses from third-party businesses
- Achieving similar results to other LLMs while using less computing power, making it more cost effective
- Engaging in dynamic and informative conversations, providing relevant and comprehensive responses to a wide range of prompts and questions

Prompt Along 4-1

1. Open the website claude.ai and create a free account. You might need to complete a few welcome steps before Claude displays a prompt box.
2. In the Claude prompt box, type the following prompt: **I would like some career advice about going into an AI field. I am creative, love to learn, and would like an interesting career in an AI field.**
3. Click the arrow to submit the prompt. Claude responds with a conversational answer (**Figure 4-17**).

Figure 4-17: Claude can help you practice your interview skills

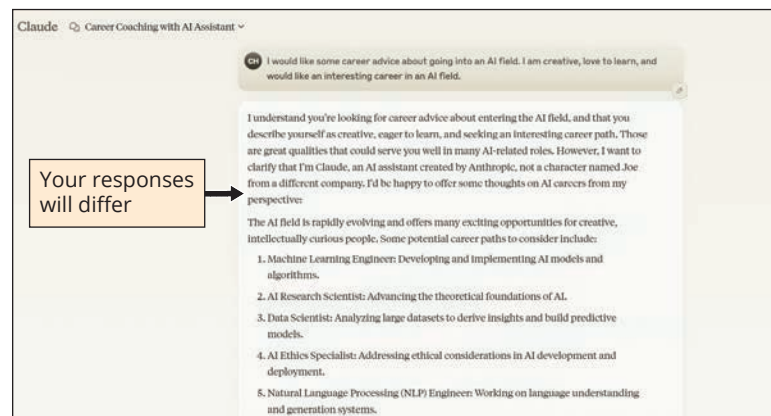
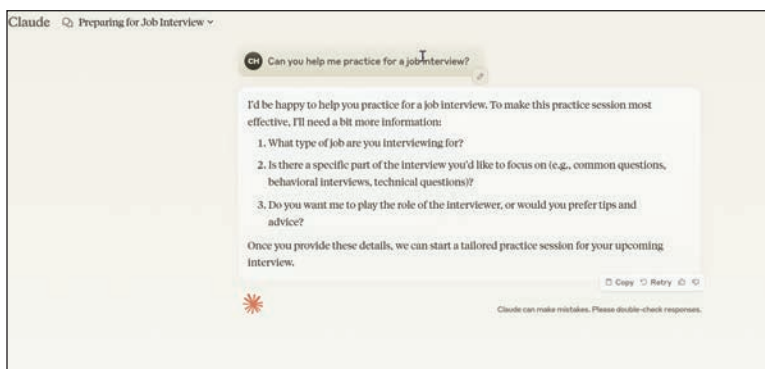


Figure 4-18: Amazon's new Alexa voice assistant uses Claude AI



Claude is enhancing Amazon's Alexa (Echo) smart device. Alexa is a virtual assistant (**Figure 4-18**), a voice-controlled device that uses artificial intelligence to understand and respond to your requests. You can ask Alexa to do various things, such as answer questions, find recipes, play music, set timers, and turn lights on and off. This means Alexa can provide more informative and detailed responses to your questions, making your interactions with your smart home devices more helpful and engaging. By integrating Claude into

Alexa, Amazon aims to elevate the voice assistant's conversational skills, interactivity, and task performance to unprecedented levels.

Meet Jasper

Jasper AI (jasper.ai) is an LLM designed to assist businesses in creating and scaling up their marketing content, including crafting product-based articles, engaging blog posts, detailed whitepapers, product descriptions, social media updates, and website content (Figure 4-19). Jasper is not free, but it does have a seven-day free trial. One of its best features is a wide range of preinstalled templates you can use to get started quickly creating company blogs, new product articles, ad copy, and business proposals. Many companies such as Wayfair, iHeartMedia, Feedfeed (Figure 4-20), and Ulta Beauty use Jasper for their marketing campaigns and brand voice.

Jasper AI can greatly improve a company's social media efforts with its diverse and effective tools. For instance, suppose a new fashion line called EcoChic is struggling with their brand's social media campaign. Jasper can help generate a variety of captivating captions, hashtags, and even visual posts. Based on a few keywords or a brief prompt, Jasper can quickly generate a plethora of options, saving time and sparking creativity. For example, if EcoChic wanted to promote a new product launch, Jasper could suggest campaign slogans such as "Conscious fashion, conscious future. Discover EcoChic's new line of sustainable clothing made from recycled materials" or "Ethical fashion, effortless style. Explore EcoChic's latest collection and make a positive impact." Jasper could also suggest relevant hashtags like #EcoChic, #sustainable-fashion, #ethicalfashion, #recycledmaterials, and #consciousliving. Jasper can assist in other social media tasks including writing Instagram captions, LinkedIn posts, and tweets. In addition, it can create a posting schedule or calendar (Figure 4-21) with the most popular times that the company's demographic may be online.

Meet Bloom

The LLM landscape includes a groundbreaking project that started in 2022 called **BLOOM (BigScience large open-science open-access multilingual language model)**, which was developed by a consortium of research institutions (Figure 4-22). BLOOM states its mission is to be more inclusive and diverse than other language models, with a strong ethical focus. BLOOM was born out of the BigScience project (led by Hugging Face), a collaborative initiative that brought together over 1,000 researchers from 60 countries and more than

Figure 4-19: Jasper is an LLM designed for business marketing

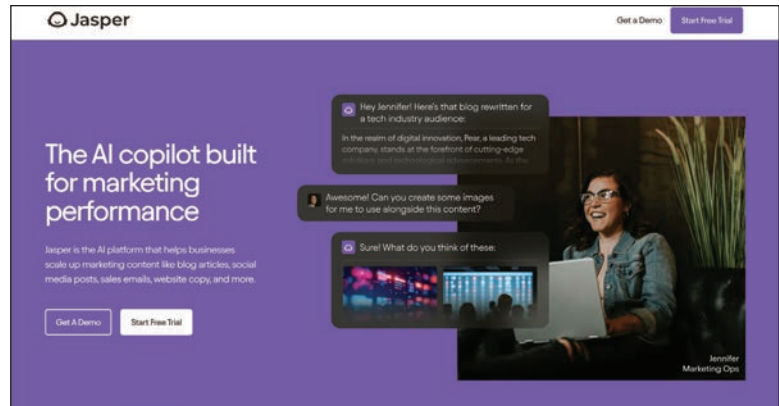


Figure 4-20: Feedfeed Instagram post based on a Jasper social media template

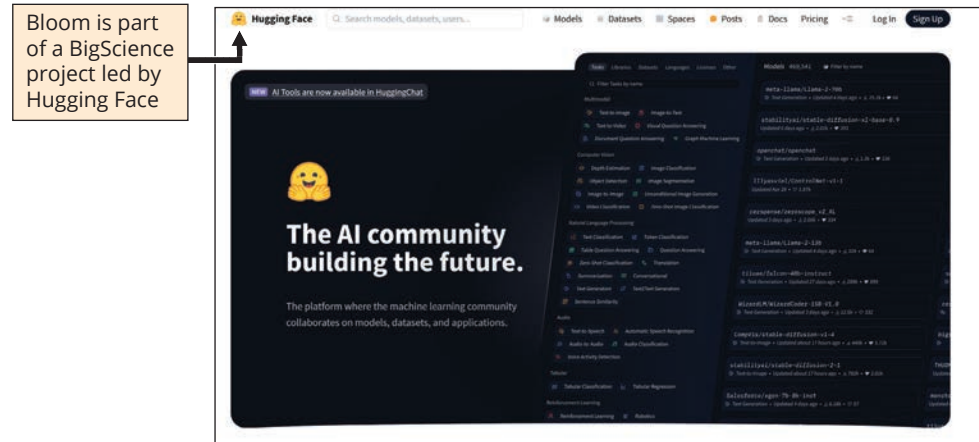


Figure 4-21: Jasper schedule for the optimal times to post to social media

I'm a social media manager and I need a posting schedule based on the ideal days and times to post on Instagram, Twitter, LinkedIn, and Facebook. Please create a posting schedule in table format.

Sure, here's a table format of the posting schedule based on the ideal days and times to post on Instagram, Twitter, LinkedIn, and Facebook:

Social Media Platform	Ideal Posting Days	Ideal Posting Times
Instagram	Monday, Tuesday, Wednesday, Friday	12pm-1pm, 2pm-3pm, 4pm-5pm
Twitter	Monday, Tuesday, Thursday	8am-9am, 10am-11am, 12pm-1pm

Figure 4-22: BLOOM is an open-source LLM for use, study, and modification

250 academic, research, and tech institutions. This diverse group focused on creating an LLM that is freely available to researchers, developers, and the public at large, distinguishing it from many other LLMs developed by private companies.

BLOOM (huggingface.co) was trained to understand and generate text in 46 languages, including widely spoken languages such as English, Spanish, and Mandarin, as well as languages with fewer speakers such as Catalan, Bengali, and Swahili. The inclusion of multiple languages was intended to make the model useful for a broader range of diverse communities.

BLOOM is unlike language models such as ChatGPT and Claude, where you open the site and start to enter prompts. BLOOM is primarily designed for developers due to its technical nature and the specialized knowledge required to use it effectively. Although it can be used to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way, it is not a straightforward tool for casual users. To get started with BLOOM, developers need to have a strong understanding of machine learning, natural language processing, and programming concepts. They must be able to do the following:

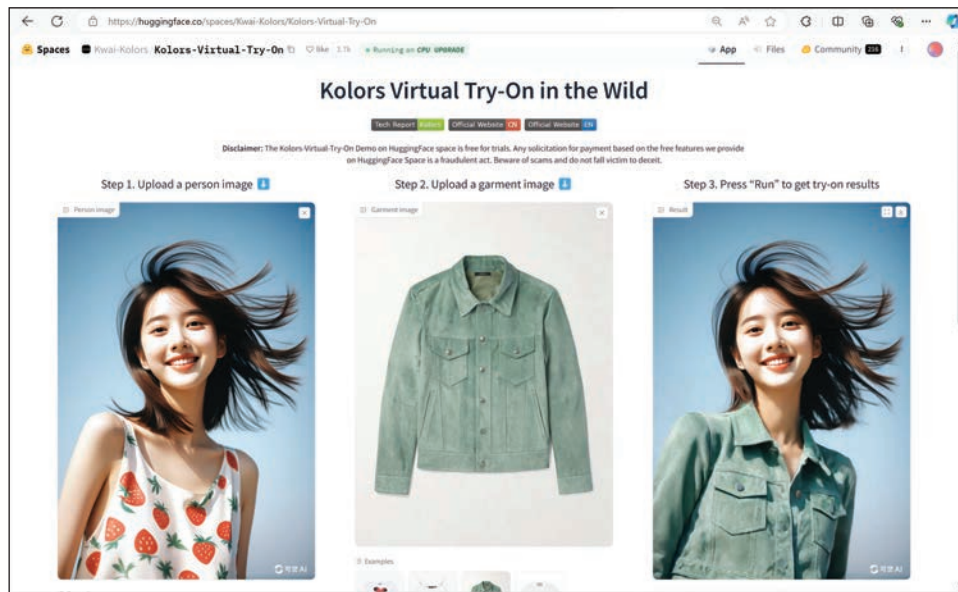
1. Download and install the model, which often involves technical setup and configuration.
2. Prepare and process data for training or fine-tuning the model.
3. Write code using programming languages like Python and frameworks like transformers to interact with the model and provide prompts.

You do not have to be a developer to experience BLOOM. **BLOOM spaces** (huggingface.co/spaces) are virtual environments developed by the community where users can experiment with interesting BLOOM-powered applications. These spaces serve as a hub for showcasing the diverse capabilities of BLOOM and providing a direct experience for users to explore different use cases such as transcribing audio in real time, storytelling, participating in language lessons, developing quiz masters for learning, and even using a Kolors Virtual Try-On app that allows you to upload a person's image or AI image, post an image of a garment of clothing, and then click Run to display the outfit on the person (**Figure 4-23**).

Meet Perplexity

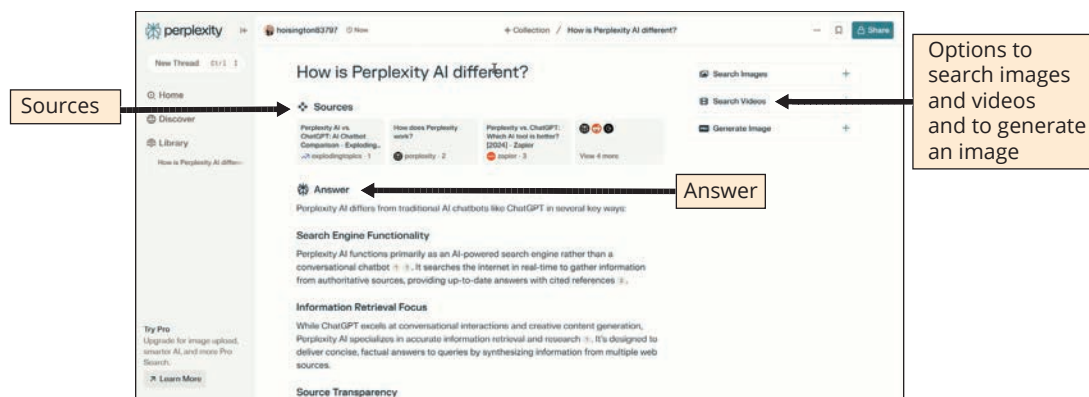
Another notable language model is **Perplexity AI** (perplexity.ai), which demonstrates proficiency in retrieving information from reliable sources and offering appropriate citations with an easy-to-use interface. Perplexity's greatest strength is its ability to access and provide up-to-date information. It pulls information from recent articles and indexes the web daily

Figure 4-23: BLOOM spaces, a virtual platform for BLOOM-powered apps like this virtual try on space



so you can ask about recent news and other typical search queries. When Perplexity was asked if its AI was different from other LLMs (**Figure 4-24**), the results first displayed the sources where it found its information, to be more transparent. By listing its references and providing footnotes for every piece of information, Perplexity can be a viable alternative to traditional search engines.

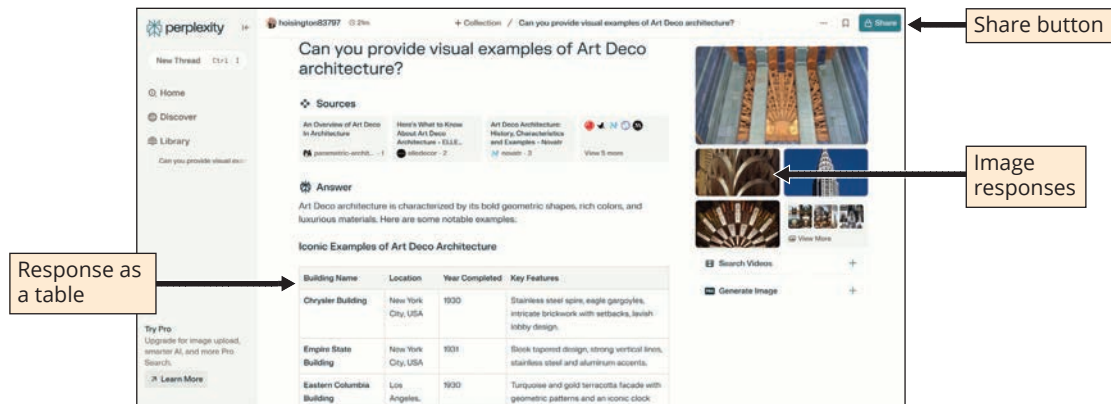
Figure 4-24: Perplexity is an LLM with access to the most recent news and information



Prompt Along 4-2

1. Open the website perplexity.ai and create a free account. You might need to click Continue until Perplexity displays a prompt box.
2. At Perplexity, enter the following prompt: **Can you provide visual examples of Art Deco architecture?**
3. Click the arrow to submit the prompt and display the response, which often includes links with examples and images (**Figure 4-25**).

Figure 4-25: Perplexity results can contain images or videos for sharing



Perplexity offers a visual search, responding to questions with images and videos. Users can view pictures in search results and play videos directly in the interface, without needing external sites like YouTube. The platform also supports image uploads for conducting enhanced searches, analyzing scenes, interpreting graphs, and generating social media captions. To share your Perplexity results, click the Share button and select Shareable. Copy the link of the website and share it with anyone you choose so that they can view your LLM results.

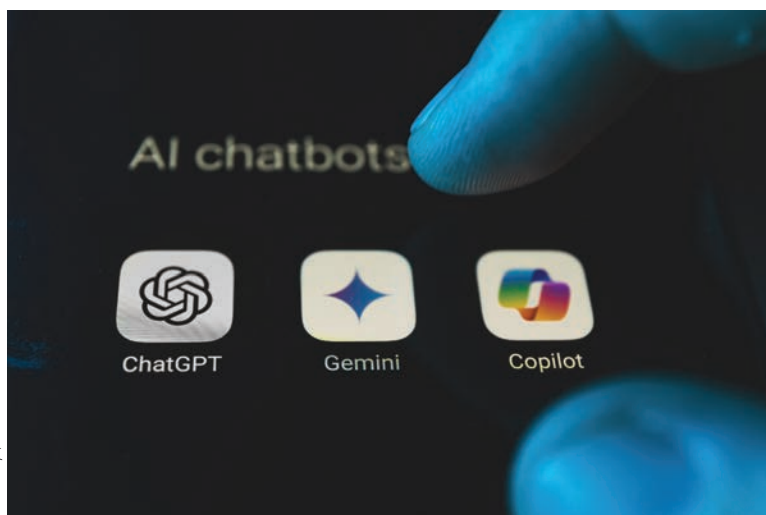
4.4 | Two Rights & a Wrong

Read the following three statements and then identify the incorrect one.

1. Jasper is a powerful AI writing assistant that leverages LLMs to help users create high-quality marketing content quickly and efficiently.
2. Claude is a special-purpose LLM for conducting scientific research.
3. BLOOM spaces provide direct experiences for users to explore use cases such as transcribing audio and participating in language lessons.

□ Answers at end of chapter

Figure 4-26: Mobile apps for ChatGPT, Gemini, and Copilot



4.5 Understanding AI on the Go: Mobile App Technology

AI has infiltrated almost every aspect of the workplace, from the LLMs that curate social media feeds to the virtual assistants that help users schedule their workday. One of the most significant daily operations of AI integration is in mobile apps. As smartphones have become indispensable tools in business and daily life, developers are seeking to enhance user experiences with AI. Most LLMs such as Open AI's ChatGPT, Google's Gemini, and Microsoft's Copilot (**Figure 4-26**) are available for installation in mobile app stores such as Android's Play Store and iPhone's App Store.

The apps mostly include the same functionality, but have more streamlined interfaces optimized for smaller screens, allowing users to access their features anywhere, anytime. This is particularly beneficial for those who need quick information or assistance while away from a larger computer.

Virtual Assistant: Your Digital Sidekick

Mobile AI has greatly improved NLP, allowing devices to better understand and respond to human language. Virtual assistants like Siri, Google Assistant, and Alexa exemplify this by setting reminders, sending messages, and offering information through voice commands. Additionally, NLP enhances translation apps for more accurate and context-aware translations. Mobile virtual assistants have become an essential tool for businesses across various industries. These AI-powered applications can streamline operations, enhance customer service, and drive efficiency. **Table 4-2** compares virtual assistant mobile usage across many industries:

Table 4-2: Mobile virtual assistant use cases

Business case	App example	Service
Healthcare	Babylon Health	When a patient uses the app, the virtual assistant schedules appointments, refills prescriptions, and accesses medical records. Babylon Health also offers virtual consultations with doctors.
IT support	TeamViewer	The virtual assistant can provide remote IT support.
General business	Slack	An employee can ask Slack for company HR policies, contact information, and scheduling.
Business travel	Triplt	This travel assistant can be used to book flights, hotels, and rental cars, and receive real-time updates on itineraries.
Sales and marketing	Google Assistant	This app can be used to qualify leads and gather customer information. It can provide personalized product recommendations based on customer preferences.
Banking	Capital One's Eno	The mobile virtual banking assistant is designed to check on your business bank account tasks (Figure 4-27).

Figure 4-27: Capital One's Eno banking mobile app

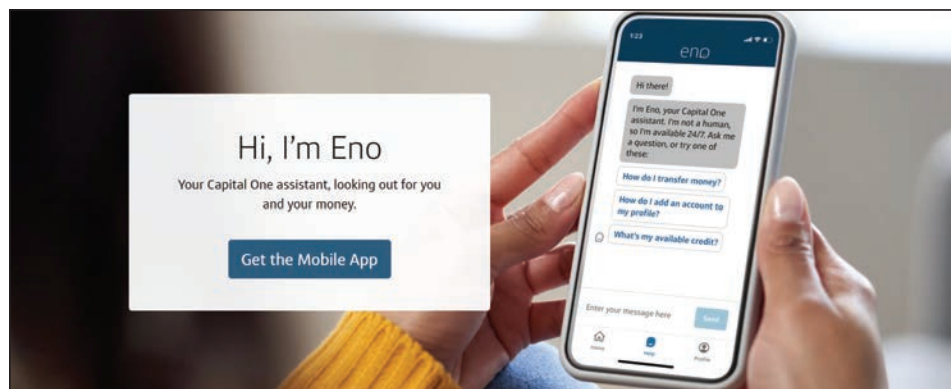


Figure 4-28: Google Lens mobile app for image recognition

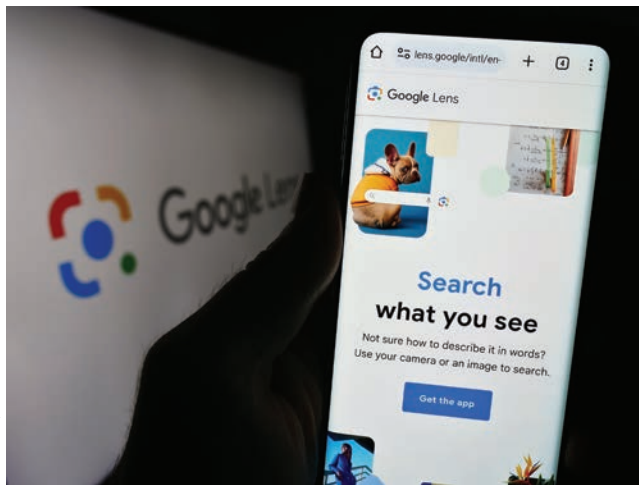


Image Recognition

Mobile AI excels in image recognition and classification, enabling devices to identify objects, scenes, and people. Camera apps use AI to tag photos with keywords for easy search and organization. AI analysis also helps identify health issues from skin scans, and it can detect fake news through image verification. Mobile apps like Google Lens (**Figure 4-28**) and Pinterest allow users to search for information based on images. By pointing their camera at an object or scene, users can find similar images, products, or relevant information. Google Lens can determine what type of plant is growing in your office or tell you where to purchase those work shoes you saw someone wearing on the subway.

4.5 | Two Rights & a Wrong

Read the following three statements and then identify the incorrect one.

1. Google Lens lets you use your camera or upload an image to find similar clothes, plants, and office décor.
2. The Triplt app is a travel organization app that helps you plan and manage your trips in one place.
3. Babylon Health is a mobile virtual assistant designed for fitness tracking.

□ Answers at end of chapter

Figure 4-29: Apple Intelligence Writing Tools



4.6 Using AI with the Apple Intelligence Platform

In fall 2024, Apple introduced the Apple Intelligence Platform, which is the personal intelligence system that puts powerful generative models at the core of iPhone, iPad, and Mac devices. Apple Intelligence Writing Tools are a suite of features designed to enhance writing efficiency and creativity on Apple devices. For example, as you type on an Apple device, the platform provides suggestions for words, phrases, and even entire sentences based on your writing style and the context of what you are writing (**Figure 4-29**). This helps to improve accuracy and speed up the composition process. It can also summarize a long group email thread or provide suggestions on improving clarity in your personal style of writing.

The Apple Intelligence Platform includes an array of technologies for businesses, including NLP, computer vision, machine learning models, and neural engines embedded in Apple's custom platform. Retailers can leverage **ARKit (Apple's augmented reality framework)** in combination with machine learning models to create immersive shopping

Figure 4-30: Apple's augmented reality framework for the retail industry



experiences, a technology that uses augmented reality to allow customers to visualize products in their own physical space before making a purchase (**Figure 4-30**). A customer can visualize what a piece of furniture will look like in a room or even what a certain color of paint would look like on their wall. Similarly, the intelligence platform's speech recognition and natural language understanding capabilities can power more sophisticated chatbots and virtual assistants, enabling businesses to provide constant customer support with more human-like interactions.

The Apple Intelligence Platform is committed to privacy and security. In an era where all users need to guard against data breaches and protect their privacy, Apple's approach to AI and machine learning are significant advantages for businesses prioritizing data protection. For example, by processing data locally on the device, Apple reduces the need to send sensitive information to cloud servers, significantly decreasing the attack surface for potential data breaches or interceptions. Keeping data on the device gives businesses more direct control over their information. This privacy protects both their employees and their customers.

This chapter has investigated the impact of big data and machine learning on various industries, highlighting their ability to process vast amounts of information and generate valuable insights in business. With early examples of LLMs like Eliza and advanced models like ChatGPT, you have examined how these technologies have revolutionized NLP and AI applications. The integration of AI into mobile apps and business strategies, as demonstrated by companies like Netflix, Amazon, and Nike, underscores the potential for AI to enhance user experiences, streamline operations, and drive innovation. Moving forward, developers and users must consider the ethical implications and ensure responsible use of AI to maximize its benefits while minimizing potential risks.

4.6 | Two Rights & a Wrong

Read the following three statements and then identify the incorrect one.

1. The Apple Intelligence Platform lacks features that protect privacy and security for businesses.
2. Apple includes an ARKit for customers to visualize the color of paint on their walls in augmented reality.
3. Apple Intelligence Writing Tools suggest words as you type.

□ Answers at end of chapter

Summary

- Big data is saved in datasets in large data centers. The era of big data and machine learning began with the widespread adoption of the Internet in the 1990s, leading to massive data generation through online activities. This data is now used for various purposes, including market research, scientific analysis, healthcare diagnoses, financial services, inventory management, and business decision-making.
- Large language models (LLMs) started with early examples like Eliza and evolved to advanced models like ChatGPT. LLMs have been trained on vast datasets to understand and generate human-like text, significantly enhancing natural language understanding and generation.
- ChatGPT is a conversational AI based on LLMs, specifically trained to understand and generate human-like text. ChatGPT became the fastest-growing consumer app in history shortly after its release in late 2022, attracting over 100 million active monthly users within just two months.
- Beyond ChatGPT, other chatbots have been designed for specific tasks or industries, showcasing their transformative impact on business operations. A new LLM called Claude AI, developed by Anthropic, is known for its safety and efficiency, as well as its integration into Amazon's Alexa tool to enhance conversational skills and task performance. Other LLMs include Jasper, BLOOM, and Perplexity.
- AI has become integral to new mobile applications, enhancing user experiences and providing quick access to information. Many LLMs, including ChatGPT, Gemini, and Copilot, are available as mobile phones apps, and offer streamlined interfaces optimized for smaller screens.
- Virtual assistants have significantly improved NLP, enabling devices to better understand and respond to human language. The use of virtual assistants like Siri, Google Assistant, and Alexa in various industries has streamlined operations, enhanced customer service, and driven efficiency.
- The Apple Intelligence Platform is a personal intelligence system integrated into iPhone, iPad, and Mac devices, designed to enhance writing efficiency and creativity.

Key Terms

ARKit (Apple's augmented reality framework)

BLOOM (BigScience large open-science open-access multilingual language model)

Bloom spaces

Claude

cloud data center

customer relationship management (CRM)

data center

dataset

large language model (LLM)

on-premises data center

Perplexity AI

semantic searching

structured data

supervised learning method

transformer

Review Questions

1. What is the relationship between big data and machine learning?
2. How does Netflix use big data and machine learning?
3. What are the five key criteria that classify data as big data?
4. What is a data center, and what purpose does it serve?
5. What are the two main types of data centers?
6. What is the role of customer relationship management (CRM) in business?
7. How does Nike use CRM to enhance customer relationships?
8. What is the significance of the transformer model in language models?
9. How did early language models like Eliza function?
10. What advancements did the transformer model bring to language models?
11. What are some limitations of ChatGPT-3?
12. What is the significance of supervised learning in training ChatGPT?
13. What are some examples of mobile virtual assistants and their use cases in different industries?
14. What is the Apple Intelligence Platform, and how does it enhance writing efficiency on Apple devices?

15. How do retailers benefit from using Apple's augmented reality framework, ARKit?

Hands-On Activities

AI Activity 4-1: Exploring Machine Learning (LO 4.1)

Create a single document containing the results for each of the following exercises.

1. Open YouTube and search for a company that is successfully using AI and big data. Using a tool such as PowerPoint or Google Slides, create five slides that identify how big data is being used at this company. Provide the YouTube link in the presentation.
2. Create a bulleted list of at least six bullets that serves as a timeline of each major development in the history of language models as shared in the chapter. The timeline should include significant milestones and the impact of each development in your own words.
3. Research and critically evaluate the potential benefits and drawbacks of LLMs on industry, society, and communication. Write an essay of at least 100 words that discusses both the positive and the negative aspects in your own words, supported by examples and research findings. Type a question related to your research into Perplexity at perplexity.ai and paste your result to compare your thoughts with the LLMs.
4. Open YouTube and search for a company that is successfully using an AI-powered CRM. Explore how these AI-powered CRM systems are used to enhance customer relationships. Share the YouTube link and write five bullet points of how this company is using the CRM.
5. Open YouTube and search for a video that shows how the Jasper LLM can assist in a marketing campaign for a company. Share the YouTube link and create a two-column table with six rows that shares the Jasper marketing feature mentioned in the video and describe how that feature assists a company.

AI Activity 4-2: Exploring Generative AI (LO 4.4)

Use the following generative AI models to perform the activities. Create a free account as needed.

1. Open Claude.ai to conduct a career job interview. Type the following prompts to guide you through a conversation with Claude.ai:
Please create a live interview to help me gain insights into my chosen career path of _____. (Tell Claude what field you are interested in within the blank.)
 - Start by telling Claude about your career interests or the specific job you'd like to explore. For example: "I'm interested in learning more about becoming a software engineer. Can you help me explore this career path?"
 - What are the basics of the career or job?
 - What does a typical day look like in this role?
 - What are the main responsibilities?
 - What skills and qualifications are usually required?
 - What degree or certifications are typically needed?
 - How long does it usually take to prepare for this career?

Enter each prompt and copy and paste your entire interview in a Word document.

2. Open Claude.ai, and then open the site docs.anthropic.com/en/prompt-library. Scroll down and try five of the prompts in the prompt library. Take a screenshot of each prompt with the results in Claude and place them in a Word document. Below each screenshot, write a short paragraph about what you think of the response.
3. Open Claude.ai. Create a prompt that asks Claude to share images and videos to help you understand eco-friendly architecture for high-rise buildings. Share screenshots of your results in a Word document.
4. Open Perplexity.ai. Look through the current news in business, such as a data breach at a company or a new AI tool. Create a prompt that would help you understand what is happening and develop your own business news story using Perplexity. Take a screenshot of your prompt and the results and any iterations in engineering your prompt and paste those results into a Word document.
5. Open Perplexity.ai. Enter the following prompt: Explain how Perplexity AI can be used to enhance business operations and decision-making. Focus on its key strengths in areas like research, data analysis, content creation, and staying updated on industry trends. Provide specific examples of how different roles within a company could leverage Perplexity's capabilities to gain competitive advantages. What were the sources that Perplexity shared in the response? Copy and paste the answer, then read through that answer and write three bullets in your own words about what Perplexity taught you.
6. Download the most recent free version of the IKEA store app. Open the app, navigate to the Rooms category and the Home office subcategory, and then locate an office desk that you could use in your home-based remote office. Once you select a desk, click the View in 3D button and then click the View in your space to display this desk in a sample room. If possible, choose to display the desk in a virtual office. Take a screenshot of this desk overlaid through augmented reality (AR) in your selected space. Place this image in the Word document and share how this would assist you in designing your home office.

Projects

Educators, please create a free account at padlet.com or use another video platform for students to use. Students, enter the following prompts to create a video discussion board for the following questions and tasks.

1. Research and record a two-minute video at your instructor's padlet.com link or an alternative video discussion board link. Download the free app called the Sherwin-Williams Color Expert onto your smartphone. Open the app and try the Color Recommendation Tool. Upload an image of your home or school and request recommendations of what colors might work. During your video, show some of the color recommendations that you liked on the app, holding your phone to display its screen in the video. How effective was the Sherwin Williams Color Recommendation tool in suggesting colors that matched your desired aesthetic or mood for the room? How easy was it to navigate the app and input information about the room's dimensions, lighting, and existing decor? Were there any limitations or drawbacks to using the app, such as difficulty in visualizing colors in different lighting conditions or a lack of options for certain color palettes? (LO 4.6)
2. Research and record a two-minute video at your instructor's padlet.com link or an alternative video discussion board link. Create a video that shares each of the LLMs mentioned in this chapter (ChatGPT, Claude, Jasper, BLOOM, Perplexity) with a discussion about the focus of each of these products. How might you use or not use each LLM? (LOs 4.3 and 4.4)

3. Research and record a two-minute video at your instructor's padlet.com link or an alternative video discussion board link. Investigate how Netflix uses big data and machine learning to enhance your user experience and inform you of movies that may interest you. Provide specific examples and explain the impact of these technologies on Netflix's business strategy. (LO 4.1)
4. Research and record a two-minute video at your instructor's padlet.com link or an alternative video discussion board link. Investigate how many data centers there are around the world. How will this number most likely grow? How do companies like Amazon Web Services (AWS) leverage data centers to provide reliable services? (LO 4.1)

Two Rights & a Wrong: Answers

4.1 The Rise of Big Data and Machine Learning

1. Data centers provide the necessary infrastructure, including massive amounts of power, cooling, and security, to support the operation of computer servers.
2. The era of big data is usually associated with modern AI development and advanced further with the Internet's widespread adoption in the 1990s.
3. Big data refers to datasets that meet multiple criteria including volume, velocity, and vortex.

Answer: The wrong statement is #3.

Explanation: Big data refers to datasets that meet volume, velocity, and variety in addition to veracity and value.

4.2 Moving Forward with Large Language Models from Eliza to ChatGPT

1. LLMs can identify common grammatical rules, sentence structures, and word usage patterns.
2. One of the earliest examples of LLMs, Eliza, developed by Weizenbaum in 1966, could mimic human conversation using pattern matching.
3. When a transformer model processes a sentence, it completes encoder, decoder, and reporter steps.

Answer: The wrong statement is #3.

Explanation: When a transformer model processes a sentence, it completes encoder, decoder, and attention mechanism steps.

4.3 ChatGPT Strengths and Weaknesses in Conversational AI Applications

1. Copilot has become the fastest-growing consumer app in history.
2. ChatGPT is built using transformers, which understand the relationship of words in a phrase regardless of their order.
3. In its first public release, ChatGPT-3 impressed people with its strong language generation abilities, but it also demonstrated some limitations.

Answer: The wrong statement is #1.

Explanation: Copilot is not the fastest-growing consumer app in history. ChatGPT became the fastest-growing consumer app in history shortly after its release in late 2022.

4.4 Other High-Functioning LLMs: Use Cases and Case Studies

1. Jasper is a powerful AI writing assistant that leverages LLMs to help users create high-quality marketing content quickly and efficiently.
2. Claude is a special-purpose LLM for conducting scientific research.
3. Bloom spaces provide direct experiences for users to explore use cases such as transcribing audio and participating in language lessons.

Answer: The wrong statement is #2.

Explanation: Claude is designed as a general-purpose LLM, capable of tasks such as generating text, translating languages, writing creative content, and answering questions in an informative way.

4.5 Understanding AI on the Go: Mobile App Technology

1. Google Lens lets you use your camera or an image to find similar clothes, plants, and office décor.
2. The TripIt app is a travel organization app that helps you plan and manage your trips in one place.
3. Babylon Health is a mobile virtual assistant designed for fitness tracking.

Answer: The wrong statement is #3.

Explanation: Babylon Health is a mobile virtual assistant for scheduling healthcare appointments, refilling prescriptions, accessing medical records, and consulting virtually with doctors.

4.6 Using AI with the Apple Intelligence Platform

1. The Apple Intelligence Platform lacks features that protect privacy and security for businesses.
2. Apple includes an ARKit for customers to visualize the color of paint on their walls in augmented reality.
3. Apple Intelligence Writing Tools suggest words as you type.

Answer: The wrong statement is #1

Explanation: The Apple Intelligence Platform has better privacy; by processing data locally on the device, Apple reduces the need to send sensitive information to cloud servers, significantly decreasing the attack surface for potential data breaches or interceptions.

Reference

1. https://www.youtube.com/watch?time_continue=15&v=bFYDrHYDns4&embeds_referring_euri=https%3A%2F%2Fwww.bing.com%2F&embeds_referring_origin=https%3A%2F%2Fwww.bing.com&source_ve_path=Mjg2NjY