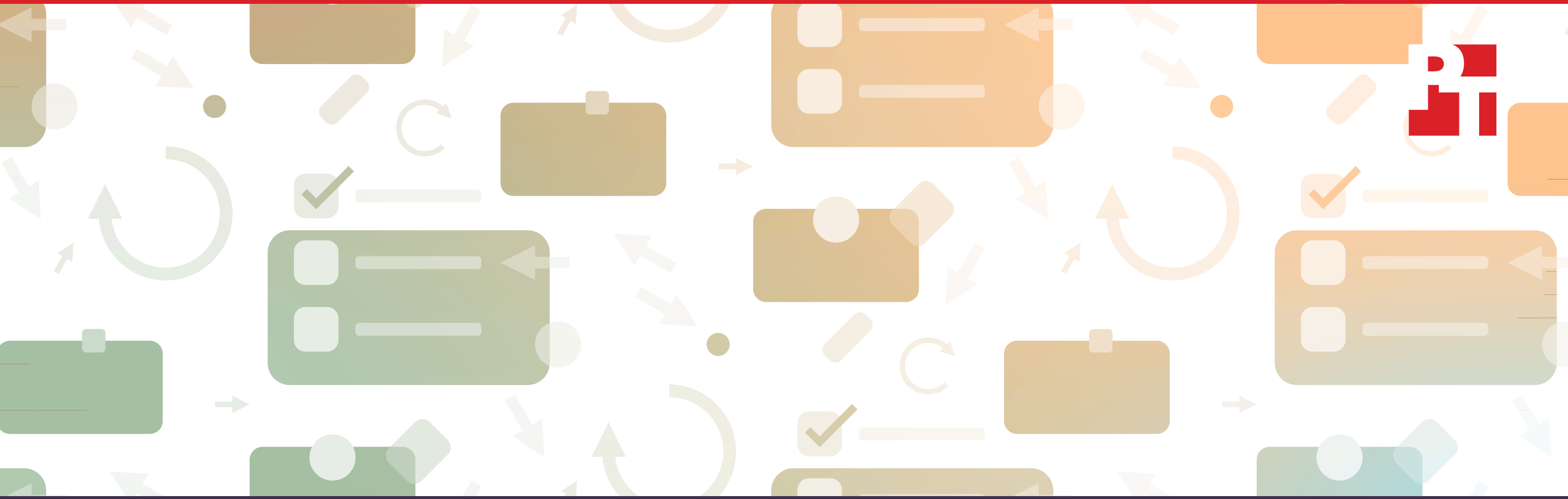
A Principled Technologies report: Hands-on testing. Real-world results.



Utilizing Azure Cosmos DB for intelligent AI-powered applications

How your organization can put artificial intelligence (AI) to work with Azure Cosmos DB

Introduction

A recent McKinsey & Company article called 2023 "generative AI's breakout year," noting that respondents to a recent survey "expect the new capabilities to transform their industries."[1] Organizations, both small and large, across every industry sector are looking to AI to revolutionize their approaches to solving business problems and seizing new opportunities. The majority of respondents to the McKinsey survey have tried generative AI (GenAI) tools at least once, and many regularly use them for work.[2]

Plenty of generative AI tools are publicly available, but some organizations are reluctant or unable to input their private data into a public tool. Nonetheless, training AI tools on your organization's data creates tremendous value, because those tools can then provide answers, insights, and even brand-new content tailored exclusively to your needs and your specific context. Imagine an AI assistant, also known as an AI copilot, for a project manager at a services business. Trained on that business's specific data, the copilot could help the project manager build a project plan and schedule, draft emails to clients, and offer alerts based on challenges from similar projects in the past. These capabilities can not only save time but also open up new insights and opportunities to improve service.

When you choose to build or populate AI tools with your own organization's data, you must decide what technology and infrastructure to use. Microsoft Azure offers a number of cloud tools and services that Microsoft designed for AI. Azure OpenAI Service, Azure AI Search, Azure Kubernetes Service (AKS), and Azure Cosmos DB—a database service that Microsoft calls "the database for the era of AI"[3]—offer companies a path to using GenAI with their own data, including potentially building AI copilots.

To investigate how Azure Cosmos DB could help support organizations' AI initiatives, we first researched how businesses in a variety of industries might use AI. We then built a proof-of-concept AI chat application that illustrates how organizations might use Azure Cosmos DB for their AI initiatives alongside their operational or transactional systems. In this report, we offer an overview of our proof of concept and several examples of areas where intelligent applications backed by Azure Cosmos DB might benefit a wide variety of organizations.

Our proof of concept

Decision-makers are examining how they can best utilize AI to speed their organizations' operations, gain efficiencies, and offer new capabilities. While the possibilities of AI are extremely broad, one type of model has received enormous attention over the past year: large language model (LLM) applications that use GenAI, such as ChatGPT. GenAI and LLMs, a type of artificial intelligence that takes in and produces text in a way that closely approximates human interaction, are changing the way we conduct business. Backed by high-performance technologies suited to their needs, GenAI applications, such as chat services and copilots, hold incredible potential. Many organizations want to harness the power of customized chat services using their own corporate data, aiming to provide better and more cost-effective customer support; to give their employees AI assistants or copilots, enabling them to access information more efficiently and work more productively; and to solve a host of other business problems. At the same time, they want to be able to use that data for transactional and other operational purposes. Azure Cosmos DB provides a data store that can support both of these needs.

To demonstrate how organizations could take a corporate database in Azure Cosmos DB and build an in-house chat application using other Azure services and the GPT-3.5-Turbo model from the Open AI organization, we created a proof-of-concept chat application. We based our application on one that Microsoft engineering developed using a bicycle shop website as a sample business. This Microsoft solution accelerator uses retrieval augmented generation (RAG) to help optimize to optimize responses based on more up-to-date data. Using its web-based chat interface, hypothetical customers of the bicycle shop can gather information on products by asking a sequence of connected questions that the solution uses to deliver better responses.[4] This method improves on the usual search function, which answers one question at a time and cannot link questions. Azure Cosmos DB stored both the underlying data set and the connected questions. We tuned this application substantially and brought in our own data set to represent the work of a real-world team creating such an application for their business. (In our case, the dataset was a large, publicly available collection of rental-property listings data derived from Airbnb.)

The solution comprised four primary components working in concert: Azure Cosmos DB for NoSQL, the web application with which the end user interacts to pose questions and receive answers, Azure AI Search, and Azure OpenAI using the GPT-3.5-Turbo and ADA text embedding models. Figure 1 shows at a high level how the data flows through the solution when a customer makes a request and gets a response.

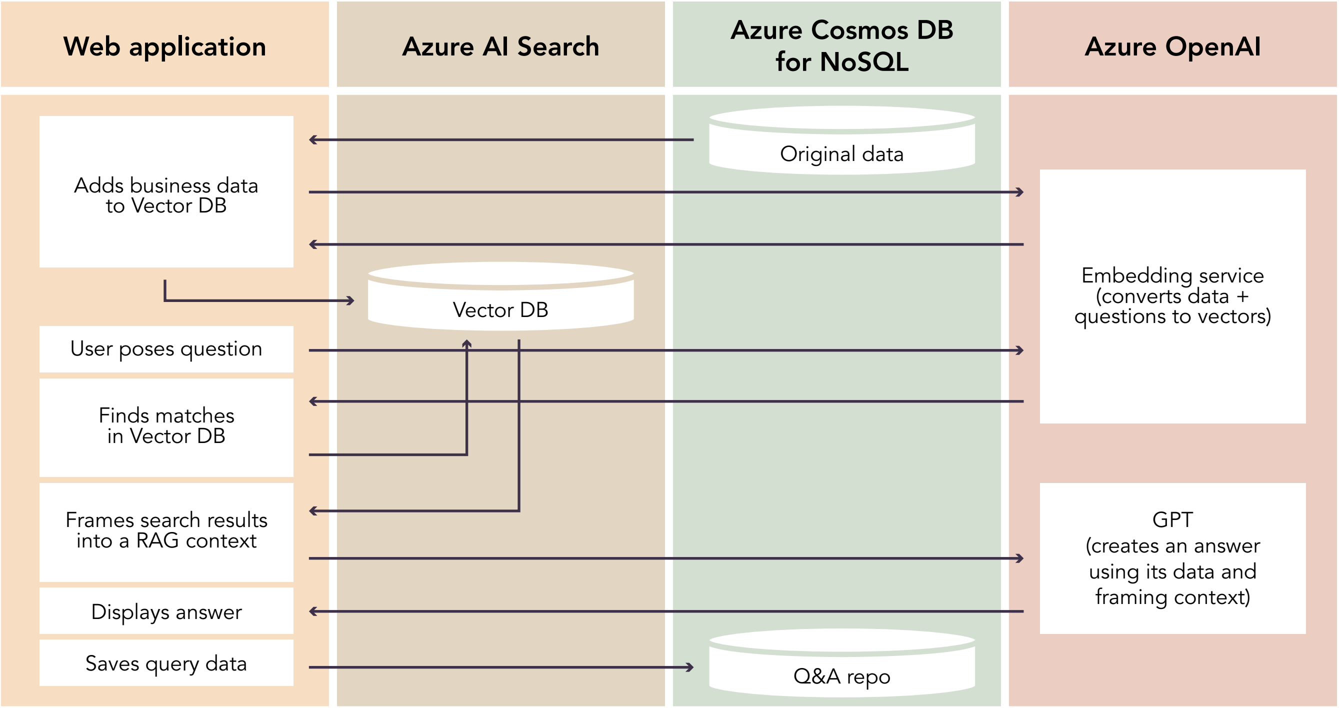

Web application
Azure AI Search
Azure Cosmos DB for NoSQL
Azure OpenAI
Original data
Adds business data to Vector DB
Embedding service (converts data + questions to vectors)
Vector DB
User poses question
Finds matches in Vector DB
Frames search results into a RAG context
GPT (creates an answer using its data and framing context)
Displays answer
Saves query data
Q&A repo

Figure 1: How the data flows through our Azure Cosmos DB-based Azure and GPT-3.5-Turbo solution when a customer makes a request and gets a response. Source: Principled Technologies.

You can read more about our proof of concept at https://facts.pt/2JHwrK6.

While our proof of concept was a chat application, the data set we chose was one that a transactional system might be simultaneously feeding with new listing information. Thus, you may choose to use these Azure Cosmos DB and these same Azure services to build a variety of AI-powered applications. Below, we describe how you might use each Azure service in an AI context.

About Azure Cosmos DB for NoSQL

Azure Cosmos DB for NoSQL is a fully managed NoSQL distributed database Microsoft has designed to enable low-latency responses and power transaction-driven workloads at scale. In contrast to traditional relational model databases, Azure Cosmos DB for NoSQL can efficiently store unstructured data. According to Microsoft, Azure Cosmos DB for NoSQL offers "single-digit millisecond response times, automatic and instant scalability, along with guaranteed speed at any scale. Business continuity is assured with SLA-backed availability and enterprise-grade security."[5] Azure Cosmos DB has a guaranteed 99th percentile read and write latency of less than 10 milliseconds.[6]

Azure Cosmos DB for NoSQL supports retrieval augmented generation for use in AI-powered applications built with Azure OpenAI models such as GPT-3.5 and GPT-4.[7] RAG is valuable because it helps to optimize LLMs for better responses based on more up-to-date data, without the high costs of retraining the model. For example, in a chat application in a retail environment, a shopper could search for a product using easy, conversational language, quickly receive a response, and then purchase the product.

While Azure Cosmos DB is a paid service, you can try it for free with a 30-day trial period using Try Azure Cosmos DB.[8] Alternatively, you can use the free tier of Azure Cosmos DB, which includes 1,000 request units per second and 25 GB of storage (with throughput and storage beyond those limits billed at the normal cost).[9] This free version is ideal for development and testing.

When should you use Azure Cosmos DB for your AI application?

If you anticipate building an operational AI application that has both chat and transactional functionality with “hot,” often-changing data, the Azure Cosmos DB database service is a strong choice. Microsoft designed Azure Cosmos DB for massive unstructured datasets, with features for accelerating performance and improving data protection through redundancy. These features include partitioning, autoscaling, and global distribution with multi-region writes and reads. Such features make Azure Cosmos DB suitable for both operational stores and chat applications or AI copilots, which need a reliable and fast backing store for their conversation history.

On the other hand, some applications rarely change older data and may only add new data; in these cases, your application could potentially use databases or storage types more suitable for “cold” storage, but still connect with an AI application. If you intend to pair an AI chat application to that solution, you may still choose to use Azure Cosmos DB to enable long-term storage for the chat application, as we did, as Azure Cosmos DB would provide a fast transaction-based unstructured data store for chat history.

Azure Cosmos DB is not the only Azure option for backing an AI chat service or other intelligent AI application. Utilizing Azure OpenAI together with Azure AI Search, vectorizing data in hybrid data sources, might also suit your needs. Alternatively, you could use Azure AI Bot service, which provides “an integrated development environment for bot building” designed for developers of all skill levels.[10]

About Azure AI Search

Azure AI Search is an AI-powered information retrieval platform. Developers use it to create effective search experiences and generative AI apps to apply LLMs to enterprise data. Organizations can use AI Search to implement search functionality for mobile and search applications and in conjunction with software-as-a-service apps. With store, index, and search vector embeddings for sentences, images, audio, graphs, and more, Azure AI Search lets users locate data semantically similar to their search queries, even when the terms they enter don’t match exactly. The customizable capabilities of Azure AI Search include key phrase extraction, language detection, optical character recognition (OCR), image analysis, translation, and role-based access control (RBAC).[11]

About Azure OpenAI Service

The Microsoft Azure OpenAI Service lets organizations access the OpenAI API through the Azure platform. Through the service, customers get to use OpenAI GPT-3.5 and GPT-4 models and enjoy the “security, reliability, compliance, data privacy and other enterprise-grade capabilities that are built into Microsoft Azure.”[12] In our testing, we used GPT-3.5-Turbo, which OpenAI calls its “most capable and cost effective model in the GPT-3.5 family.”[13] Microsoft has optimized GPT-3.5-Turbo for chat and it also works well for traditional completion tasks.[14]

According to Microsoft, “Customers of all sizes across industries are using Azure OpenAI Service to do more with less, improve experiences for end-users, and streamline operational efficiencies internally.”[15] The use cases to which organizations are applying the capabilities of Azure OpenAI Service include “customer support, customization, and gaining insights from data using search, data extraction, and classification.”[16]

About Vector Search & AI Assistant, the reference application we used as a starting point for our solution

Vector Search & AI Assistant is part of the Microsoft Official Build & Modernize AI Applications reference solutions library that Microsoft has developed to help users build AI-enabled applications and services in Azure. Organizations can use these solutions as a starting point for their own bespoke solutions, as we did at PT.[17]

According to Microsoft, the Vector Search & AI Assistant solution focuses on "a consumer retail 'Intelligent Agent' that allows users to ask questions (RAG Pattern) on vectorized product, customer and sales order data stored in the database."[18]

Learn more about these solutions at https://github.com/Azure/Build-Modern-AI-Apps#readme and access the Vector Search & AI Assistant solution at https://github.com/Azure/Vector-Search-AI-Assistant/tree/cognitive-search-vector.

Example real-world use cases of intelligent AI applications with Azure Cosmos DB

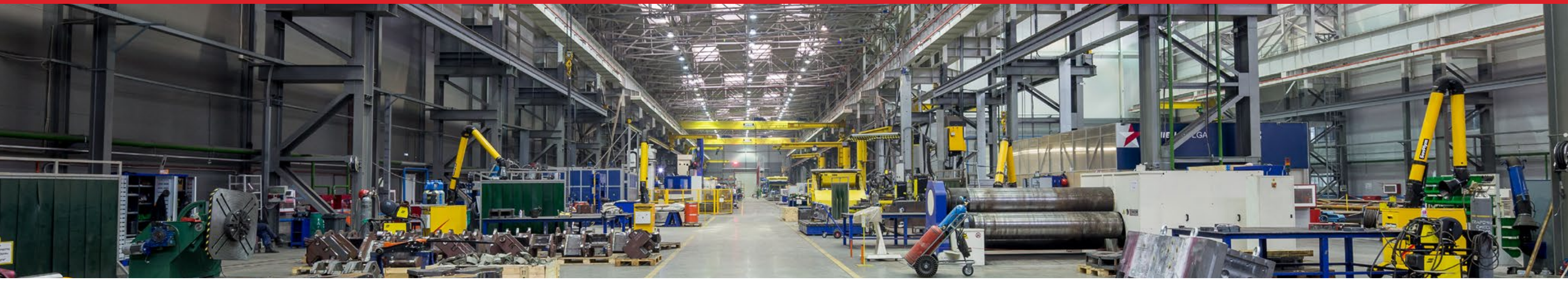
It's easy to see the utility of an application that allows you to search through your troves of data and find answers, but the value of intelligent AI applications backed by Azure Cosmos DB has the potential to far surpass that of a search engine. While our proof of concept focused on the chat functions, you can build intelligent, AI-enabled applications that assist in transactions in a copilot style, as well as providing a more natural way to search for enterprise data. Regardless of your industry, AI can help grow your organization's capabilities and solve business problems. In this section, we cover potential chat application use cases and other AI opportunities in retail, manufacturing, healthcare, legal work, and finance. In all these industries—and more—you can build on Azure Cosmos DB and the other Azure services we've described to realize these types of benefits.

Retail

Retail operations can benefit from AI chat or AI copilot applications in a wide variety of areas. In addition to letting companies provide friendly, realistic customer support 24/7 while reducing support costs, chat applications can send targeted marketing messages, give sales reps access to product information in an AI copilot context during a call, integrate with loyalty programs, boost sales by presenting complementary products, and improve the customer experience by sending automated follow-up or onboarding messages after sales.[19] AI applications that enable quick transactions can also assist in making the purchasing process easier, faster, and more seamless for shoppers, potentially helping companies boost their revenue.

Companies can use AI applications to help detect fraud by identifying suspicious activity, such as customers making large purchases with different credit cards; analyze sales data to predict demand and help manage inventory levels; and help with the returns and exchanges process, around which customers often seek guidance.[20]

AI chat applications also support so-called "conversational commerce," which, in contrast to traditional linear search approaches, can improve conversion rates and average basket size by presenting customers with clusters of products, such as all the ingredients a recipe requires or all the articles of clothing and accessories that make up an outfit.[21]



Manufacturing

In manufacturing, AI chat applications have the potential to improve operations, handle customer service efficiently, provide an interactive platform, market content through online channels, improve organizational efficiency, provide easy access to knowledge and databases, and even perform some HR-related tasks.[22]

AI chat applications are also solving many supply chain challenges. The quick responses they offer improve both sales engagement and supplier engagement. They can help in the operational side of the business by placing orders for raw materials automatically and arranging logistics, such as navigating the location of products and raw materials in warehouses and can manage stock-keeping units (SKUs) to assist with inventory.[23]

AI plays an important role in additive manufacturing, more widely known as 3D printing, where you create objects by adding one layer after another. Manufacturers can use AI to optimize how they dispense and apply materials and to identify and fix errors in real-time.[24] Blacksmith is an AI-based tool that compares product designs with finished products and aligns them more closely by automating fine-tuning of the construction process. Among the potential users of this type of technology are "footwear giants Adidas and Reebok, which are now using 3D printing technology to create complex lattice structures for more comfortable and performance-enhancing running shoes."[25]

Healthcare

AI has enormous potential in medical analysis and disease diagnosis, with multiple machine learning models, such as 3D U-Net, specifically designed for medical AI tasks. But medical and healthcare groups are also putting AI to work in everyday patient interactions. Walgreens, for example, is using an AI solution backed by Microsoft Azure—including Azure Cosmos DB—to offer real-time insights to its pharmacists, enabling them to understand more about their patients.[26]

Intelligent AI chat applications can assist with many aspects of communication between healthcare providers and patients, including interviewing patients after surgery, providing suggestions and information about medications and managing their side effects and refill schedules, and carrying out mental health screenings.[27] Chat applications can also help with scheduling appointments, assist with insurance coverage and claims, and assess symptoms.[28]

One aspect of chat applications that many view as negative, their impersonal nature, can turn into an advantage in healthcare; when patients perceive questions coming from a non-human and non-judgmental source, they can "feel more comfortable sharing certain medical information such as checking for STDs, mental health, sexual abuse, and more."[29]

Legal

AI chat applications have the potential to help legal firms meet client expectations for timely communication.[30] They can also offload many routine tasks from legal professionals, helping with research, client communication, and lead generation.[31] Because drafting contracts frequently consists of changing relatively few details on a standard document, this is a task well-suited to using bots, automation that "can lead to big savings on time and cost for human lawyers, as it needs little to no intervention."[32] AI-driven chat applications can also help automate and simplify operational processes such as tracking billable hours and reviewing and summarizing documents.[33]

AI also has the potential to vastly accelerate "one of the most time-consuming tasks in litigation: extracting structure, meaning, and salient information from an enormous set of documents produced during discovery."[34] AI can also quickly produce initial drafts of motions to file with a court, "citing the relevant case law, advancing arguments, and rebutting (as well as anticipating) arguments advanced by opposing counsel."[35] While these initial drafts require human review before they become usable, incorporating AI into the process can save enormous amounts of time.

Finance

In the finance industry, institutions are using AI chat applications in customer service, fraud detection, loan origination, wealth management, compliance, financial planning, customer onboarding, and risk management. They are also utilizing GPT models in their know-your-customer (KYC) and anti-money laundering (AML) processes, which banks implement to mitigate the risk of financial crime and comply with regulations. These models can analyze large volumes of customer data and identify potential compliance issues—all while the underlying data is part of standard operational workflows. They can also "verify customer identity, check customers against sanction lists, and flag suspicious transactions."[36]

Opportunities for AI innovation in finance go far beyond chat applications. One example is the new PwC Next Generation Audit initiative, which utilizes Azure Cosmos DB and Azure OpenAI Service, among other Azure services.[37] This initiative aims to help PwC "enhance audit quality and value while streamlining data acquisition," giving employees more time to focus on high-risk problems and helping to improve user experience for PwC clients.[38]

Conclusion

It's clear that AI is bringing new opportunities to organizations across industries as they work to create more intelligent, AI-enabled applications. What is not always clear, though, is where and how to start with your own AI implementation on your own operational data. In this report, we've highlighted several possible use cases where Azure Cosmos DB, among other Azure services, may be able to help. To demonstrate how well some of these services work with Azure Cosmos DB, we found it straightforward to build a proof-of-concept intelligent chat application utilizing data we brought to the table. For your AI initiatives, consider exploring Azure Cosmos DB and the Azure ecosystem.

[1] "The state of AI in 2023: Generative AI's breakout year," accessed December 19, 2023, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year.

[2] "The state of AI in 2023: Generative AI's breakout year."

[3] "Azure Cosmos DB," accessed December 19, 2023, https://azure.microsoft.com/en-us/products/cosmos-db.

[4] GitHub, "Azure/Vector Search AI Assistant," accessed December 1, 2023, https://github.com/Azure/Vector-Search-AI-Assistant/tree/cognitive-search-vector.

[5] Microsoft, "Azure Cosmos DB – Unified AI database," accessed December 1, 2023, https://learn.microsoft.com/en-us/azure/cosmos-db/introduction.

[6] "Consistency levels in Azure Cosmos DB," accessed December 19, 2023, https://learn.microsoft.com/en-us/azure/cosmos-db/consistency-levels.

[7] Microsoft, "Azure Cosmos DB – Unified AI database," accessed December 1, 2023, https://learn.microsoft.com/en-us/azure/cosmos-db/introduction.

[8] "Try Azure Cosmos DB free," accessed December 19, 2023, https://learn.microsoft.com/en-us/azure/cosmos-db/try-free?tabs=nosql.

[9] "Azure Cosmos DB free tier," accessed December 19, 2023, https://learn.microsoft.com/en-us/azure/cosmos-db/free-tier.

[10] "Azure AI Bot Service," accessed December 19, 2023, https://azure.microsoft.com/en-us/products/ai-services/ai-bot-service.

[11] Microsoft, "Azure AI Search," accessed December 1, 2023, https://azure.microsoft.com/en-us/products/ai-services/cognitive-search/.

[12] Microsoft, "New Azure OpenAI Service combines access to powerful GPT-3 language models with Azure's enterprise capabilities," accessed December 1, 2023, https://news.microsoft.com/source/features/ai/new-azure-openai-service/.

[13] OpenAI, "Models," accessed December 1, 2023, https://platform.openai.com/docs/models/gpt-3-5.

[14] Microsoft, "Azure OpenAI Service models," accessed December 8, 2023, https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models.

[15] Eric Boyd, "General availability of Azure OpenAI Service expands access to large, advanced AI models with added enterprise benefits," accessed December 1, 2023, https://azure.microsoft.com/en-us/blog/general-availability-of-azure-openai-service-expands-access-to-large-advanced-ai-models-with-added-enterprise-benefits/.

[16] Eric Boyd, "General availability of Azure OpenAI Service expands access to large, advanced AI models with added enterprise benefits."

[17] GitHub, "Azure/Build-Modern-AI-Apps," accessed November 30, 2023, https://github.com/Azure/Build-Modern-AI-Apps#readme.

[18] GitHub, "Azure/Build-Modern-AI-Apps."

[19] Bernard Marr, "Revolutionizing Retail: How ChatGPT Is Changing The Shopping Experience," accessed December 7, 2023, https://www.forbes.com/sites/bernardmarr/2023/03/21/revolutionizing-retail-how-chatgpt-is-changing-the-shopping-experience/?sh=3a4d685d2540.

[20] Bernard Marr, "Revolutionizing Retail: How ChatGPT Is Changing the Shopping Experience."

[21] Sudip Mazumder, Rakesh Ravuri, and Sara Alloy, "How Retailers Can Increase Profits With Generative AI," accessed December 7, 2023, https://www.publicissapient.com/insights/generative-artificial-intelligence-chatgpt.

[22] Botpress, "Chatbots for manufacturing industry | Benefits & Use Cases," accessed December 7, 2023, https://botpress.com/blog/chatbots-for-manufacturing-industry.

[23] Rashmi Ranjan Panigrahi, Avinash K. Shrivastava, Karishma M. Qureshi, Bhavesh G. Mewada, Saleh Yahya Alghamdi, Naif Almakayeel, Ali Saeed Almuflih, and Mohamed Rafik N. Qureshi, "AI Chatbot Adoption in SMEs for Sustainable Manufacturing Supply Chain Performance: A Mediational Research in an Emerging Country," accessed December 7, 2023, https://doi.org/10.3390/su151813743.

[24] Bernard Marr, “Artificial Intelligence In Manufacturing: Four Use Cases You Need To Know In 2023,” accessed December 19, 2023, https://www.forbes.com/sites/bernardmarr/2023/07/07/artificial-intelligence-in-manufacturing-four-use-cases-you-need-to-know-in-2023/?sh=350fe08b3bd8.

[25] Bernard Marr, “Artificial Intelligence In Manufacturing: Four Use Cases You Need To Know In 2023.”

[26] “Walgreens empowers pharmacists with an intelligent prescription data platform on Azure,” accessed December 19, 2023, https://customers.microsoft.com/en-us/story/1411448755996187154-walgreens-health-provider-azure.

[27] Janice Dombrowski, “Chatbot Examples: How 8 Industries Are Thriving With Chatbots,” accessed December 8, 2023, http://www.streamcreative.com/blog/chatbot-examples.

[28] Inbenta, “Benefits of Chatbots in Healthcare: 9 Use Cases of Healthcare Chatbots,” accessed December 8, 2023, https://www.inbenta.com/articles/benefits-of-chatbots-in-healthcare-9-use-cases-of-healthcare-chatbots/.

[29] Inbenta, “Benefits of Chatbots in Healthcare: 9 Use Cases of Healthcare Chatbots.”

[30] Airdroid, “6 Best Chatbots for Lawyers & What They Can Do,” accessed December 8, 2023, https://www.airdroid.com/ai-insights/chatbot-for-lawyers/.

[31] Yauhen Zaremba, “Legal Chatbots: How Bots Can Bring in More Clients For Law Firms,” accessed December 8, 2023, https://landbot.io/blog/chatbots-legal-sector.

[32] Yauhen Zaremba, “Legal Chatbots: How Bots Can Bring in More Clients For Law Firms.”

[33] Yauhen Zaremba, “Legal Chatbots: How Bots Can Bring in More Clients For Law Firms.”

[34] John Villasenor, “How AI will revolutionize the practice of law,” accessed December 19, 2023, https://www.brookings.edu/articles/how-ai-will-revolutionize-the-practice-of-law/.

[35] John Villasenor, “How AI will revolutionize the practice of law.”

[36] Bernard Marr, “Top 10 Use Cases For ChatGPT In The Banking Industry,” accessed December 8, 2023, https://www.forbes.com/sites/bernardmarr/2023/03/08/top-10-use-cases-for-chatgpt-in-the-banking-industry/?sh=1464cdc72fbf.

[37] “PwC delivers the human-led tech-powered future of auditing with Azure OpenAI Service,” accessed December 19, 2023, https://customers.microsoft.com/en-us/story/1702043495902005748-pwc-azure-openai-service-united-kingdom.

[38] “PwC delivers the human-led tech-powered future of auditing with Azure OpenAI Service.”

This project was commissioned by Microsoft.

Principled Technologies®

Facts matter.®

Principled Technologies is a registered trademark of Principled Technologies, Inc.
All other product names are the trademarks of their respective owners.

DISCLAIMER OF WARRANTIES; LIMITATION OF LIABILITY:
Principled Technologies, Inc. has made reasonable efforts to ensure the accuracy and validity of its testing, however, Principled Technologies, Inc. specifically disclaims any warranty, expressed or implied, relating to the test results and analysis, their accuracy, completeness or quality, including any implied warranty of fitness for any particular purpose. All persons or entities relying on the results of any testing do so at their own risk, and agree that Principled Technologies, Inc., its employees and its subcontractors shall have no liability whatsoever from any claim of loss or damage on account of any alleged error or defect in any testing procedure or result.

In no event shall Principled Technologies, Inc. be liable for indirect, special, incidental, or consequential damages in connection with its testing, even if advised of the possibility of such damages. In no event shall Principled Technologies, Inc.’s liability, including for direct damages, exceed the amounts paid in connection with Principled Technologies, Inc.’s testing. Customer’s sole and exclusive remedies are as set forth herein.