

# Three Proven Patterns for Azure OpenAI

A step-by-step guide to building  
scalable, AI-powered applications

# Three Proven Patterns for Azure OpenAI

3 /

Achieve enterprise-grade AI with  
Azure OpenAI in Foundry Models

19 /

How Azure OpenAI drives customer  
success

21 /

Next steps

8 /

Kickstart your journey with Azure  
OpenAI

21 /

Azure OpenAI for innovative and  
responsible AI solutions

# Achieve enterprise-grade AI with Azure OpenAI in Foundry Models

AI is all around and has been for years, from recommendations on e-commerce sites to posts on social media feeds and facial recognition used to unlock devices. In the midst of this AI revolution, the biggest breakthrough to burst on to the scene in 2022 was commercially viable generative AI (GenAI).

GenAI refers to AI that produces text, images, videos, code, sound, or other similar types of content, and its potential applications are virtually limitless. Writers use it for blogs, copywriting, and to overcome writer's block. Engineers use it to write, fix, and understand code. Business leaders use it to source new ideas, get constructive feedback, and identify opportunities for investment and expansion. And helping users on this transformative journey is Azure OpenAI.

**Azure OpenAI in Foundry Models**, built in collaboration with OpenAI, is Microsoft's premier cloud-based GenAI

tool. It combines GPT-4's multimodal capabilities with Azure's enterprise-grade security, compliance, and governance. Easily integrating with Azure's cloud and data solutions, it allows businesses to incorporate GenAI into both internal and customer-facing operations.

But the evolution doesn't stop there. The next wave of GenAI lies in **multimodality**—an advanced form of deep learning that uses multiple data types, such as text, audio, and images together to build more versatile and powerful AI models, and it's already here. Microsoft Azure lets users easily manage and integrate various data sources with multimodal GenAI models, enabling them to create more dynamic, responsive, and secure AI applications. Azure OpenAI helps organizations harness their data, innovate faster, and ultimately deliver richer AI-powered experiences to their employees as well as their customers.

# Drive GenAI innovation with Azure OpenAI

GenAI is revolutionizing industries by enabling diverse applications across domains. To fully harness Azure OpenAI's potential, it's crucial to understand its capabilities that position it as a leader in applied AI.

<b>Hyperscale AI</b>	Microsoft offers provisioned throughput capacity (PTU) across a wide range of global regions, enabling customers to create, deploy, and host production-grade, customer-facing AI applications.
<b>Variety of AI models</b>	Azure OpenAI provides same-day model availability, with 1600+ models from well-known providers such as Meta, Mistral, and Cohere.
<b>Performant AI</b>	Azure OpenAI offers an industry-leading 99.99% SLA on PTUs and 99.9% SLA for pay-as-you-go deployments.
<b>Optimized AI</b>	Azure OpenAI allows customers to use their data with retrieval-augmented generation (RAG), reducing costs by offering specialized SLMs and an intuitive interface for model deployment, management, and cost control.
<b>Private AI</b>	Microsoft doesn't use user inputs or outputs to train Azure OpenAI models, nor does any data flow back to OpenAI or third parties. Microsoft does not use this data to improve any products.
<b>Secure AI</b>	Azure OpenAI offers enterprise-grade features, such as data residency, data exfiltration protection, private networking, managed identity, and integration with Microsoft Defender for Cloud.
<b>Industry-leading AI features</b>	Microsoft adds value through content filtering, safety system templates, custom blacklists, and more, compared to OpenAI's Moderation API or similar offerings.

By using Azure AI Foundry solutions, businesses can build innovative, enterprise-grade applications tailored to their unique needs—whether automating customer interactions, extracting insights, or creating personalized experiences—accelerating digital transformation.

# Understand Azure OpenAI use cases

Let's explore the real-world use cases Azure OpenAI enables.

Top use cases for Azure OpenAI				
Business needs	Increase productivity	Build creative content	Automate processes	Improve customer experience
Business use case	Internal virtual assistant	Digital asset management	Workflow management/RPA	Intelligent contact center
	Developer efficiency	Personalized content generation	Document processing	Accessible services
	Document creation and analysis	Product design and development	Fraud, security, and threat detection	Personalized customer experience
	Business analytics	Digital art (including branded content)	Digital inspection and comparison	
	Learning	Marketing, advertising, and sales content generation	Supply chain optimization Compliance	
What can GenAI do?	Generate new revenue streams	Modernize internal processes	Deliver differentiated customer experiences	

Figure 1: Azure OpenAI use cases

## Increase productivity

- **Internal virtual assistant:** Automate routine tasks within the company, providing instant support for employees' inquiries and reducing manual effort.
- **Developer efficiency:** Enable code generation, debugging, and suggestions using AI tools to speed up the software development lifecycle.
- **Document creation and analysis:** Automate drafting of documents and extract insights from text-based data using AI to save time and improve accuracy.
- **Business analytics:** Use AI technology to quickly process and analyze business data, identifying emerging trends and patterns that support more informed decision making.
- **Learning:** Provide personalized learning and training resources for employees, utilizing AI to continuously adapt to individual needs and improve skill development.

## Build creative content

- **Marketing, advertising, and sales content generation:** Use AI to create high-quality content for campaigns, including product descriptions, social media posts, and more.
- **Digital asset management:** Efficiently manage, organize, and search for digital assets with AI to enhance creative workflows and simplify finding resources.
- **Personalized content generation:** Create content tailored to individual customer profiles, leveraging AI to craft messages and media that resonate with specific audiences.
- **Product design and development:** Employ AI tools to aid in the design process, predicting trends and providing design suggestions, speeding up product development.
- **Digital art (including branded content):** Utilize AI to create digital artworks, graphics, and branded content quickly, supporting brand aesthetics and marketing efforts.

## Automate processes

- **Workflow management/RPA:** Streamline workflows by using AI-powered bots to handle repetitive tasks, improving process efficiency.
- **Document processing:** Use AI for document recognition, data extraction, and automated data entry, speeding up document handling.
- **Fraud, security, and threat detection:** Utilize AI to monitor and detect unusual activities that may signal fraud or security breaches.
- **Digital inspection and comparison:** Use AI to inspect and compare digital assets or physical products through image processing.
- **Supply chain optimization:** Improve supply chain operations by forecasting demand, optimizing inventory, and reducing bottlenecks.
- **Compliance:** Automate compliance checks and ensure adherence to regulatory standards through AI-based monitoring and reporting.

## Improve customer experience

- **Personalized customer experience:** Use AI to analyze customer behavior and preferences, tailoring offerings and communication for a more personalized interaction.
- **Intelligent contact center:** Implement AI-driven customer service, such as chatbots and virtual agents, to provide instant responses and resolve customer queries efficiently.
- **Accessible services:** Ensure that all customers have access to your services, with AI tools that can translate languages, provide text-to-speech, and more, catering to diverse user needs.

With an understanding of Azure's AI capabilities and their practical applications established, it's time to begin the implementation journey with Azure OpenAI.

# Kickstart your journey with Azure OpenAI

Azure AI Foundry empowers users to harness the power of GenAI, offering innovation through a wide range of cutting-edge models tailored to various use cases. Users can select the Azure OpenAI model that fits their needs: for highly detailed conversations, GPT-4o (0513) offers advanced capabilities, while GPT-4o mini (0718) provides a more efficient, budget-conscious option. For general-purpose chatbots, GPT-4 (0613) is recommended.

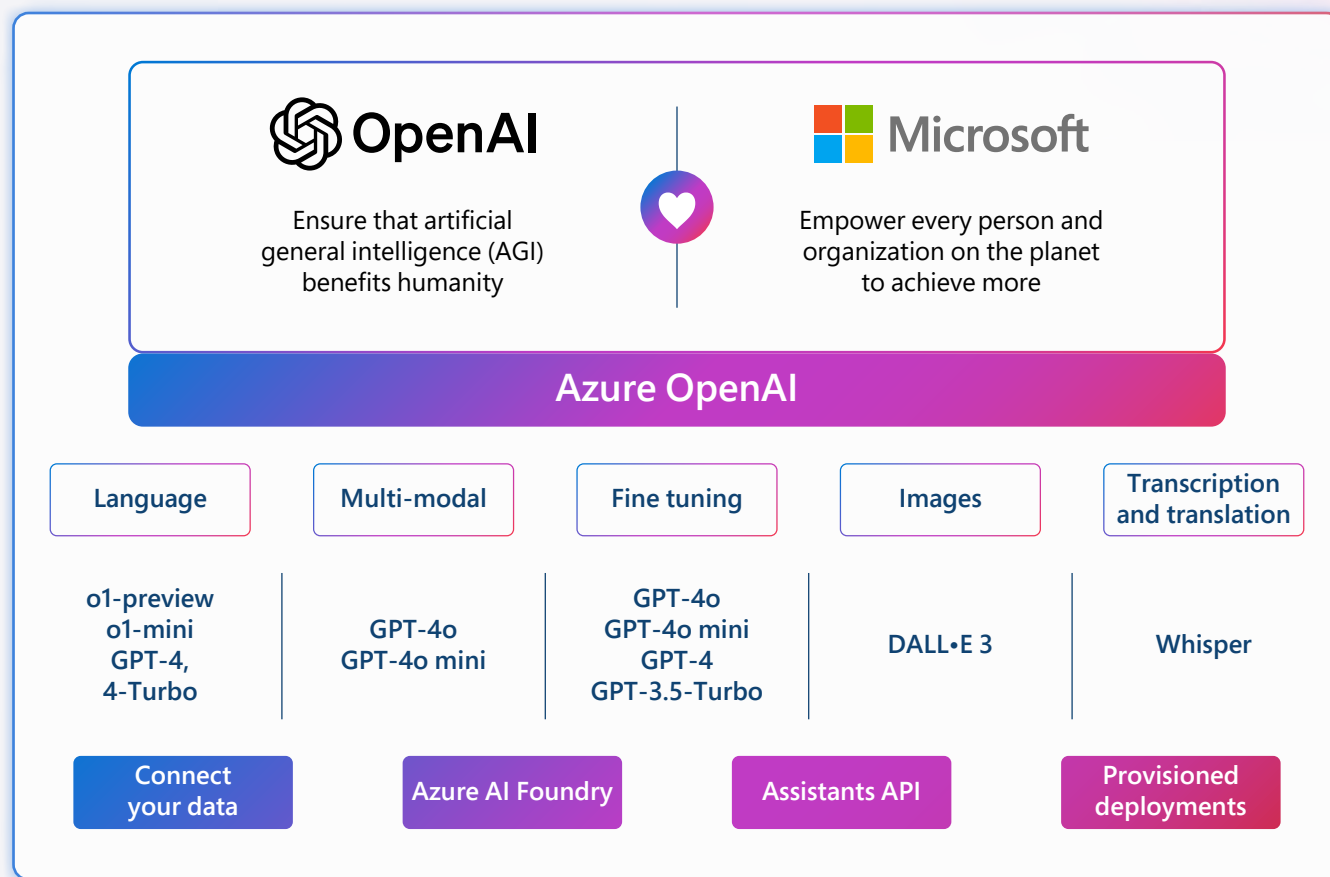


Figure 2: Microsoft + OpenAI collaboration



This flexibility ensures that users can deploy the most suitable model for their project.

By using RAG, users can create high-performing experiences with their own data, ensuring more contextual and personalized interactions. Built-in data privacy features ensure confidence and proactive safety, while streamlined tools empower developers with the agility to operationalize AI solutions quickly and efficiently across diverse applications.

With Azure AI Foundry, users can confidently choose from over 1,600 models, supported by comprehensive benchmarking and lifecycle evaluations. While the patterns in this book focus on specific models, Azure's flexibility lets users explore other offerings. The unified API makes it easy to swap models, ensuring adaptability and future-proofing investments as AI evolves.

Find the right model to build your custom AI solution Show filters

Announcements

**Meta Llama 3.2 models are here!**

Llama 3.2 11B Vision Instruct and 90B Vision Instruct are here for your image reasoning use cases.

[View models](#) [Read blog](#)

**News from Cohere!**

Cohere's collection now includes Command R 08-2024 and Command R+ 08-2024.

[View models](#)

**Experience the o1 models**

The o1 series feature an enhanced reasoning abilities to solve science and coding problems.

[Try limited access](#) [Read blog](#)

**ALLaM-2-7B: latest Arabic LLM**

ALLaM-2-7B is here! A robust 7B LLM model crafted to boost Arabic language technology.

[View models](#) [Read blog](#)

[All filters](#) [Collections](#) [Industry](#) [Deployment options](#) [Inference tasks](#) [Fine-tuning tasks](#) [Licenses](#)

Search Models 1794

<b>gpt-4o-realtime-preview</b> Audio generation	<b>openai-whisper-large-v3</b> Speech recognition	<b>openai-whisper-large</b> Speech recognition	<b>gpt-4</b> Chat completion	<b>gpt-3.5-turbo</b> Chat completion
<b>o1-preview</b> Chat completion	<b>o1-mini</b> Chat completion	<b>gpt-4o-mini</b> Chat completion	<b>gpt-4o</b> Chat completion	<b>gpt-4-32k</b> Chat completion
<b>gpt-3.5-turbo-instruct</b> Chat completion	<b>gpt-3.5-turbo-16k</b> Chat completion	<b>dall-e-3</b> Text to image	<b>dall-e-2</b> Text to image	<b>whisper</b> Speech recognition
<b>tts-hd</b> Text to speech	<b>tts</b> Text to speech	<b>text-embedding-3-small</b> Embeddings	<b>text-embedding-3-large</b> Embeddings	<b>Phi-3-mini-4k-instruct</b> Chat completion
<b>Phi-3-medium-4k-instruct</b> Chat completion	<b>Phi-3-mini-128k-instruct</b> Chat completion	<b>Phi-3-medium-128k-instruct</b> Chat completion	<b>Phi-3-small-8k-instruct</b> Chat completion	<b>Phi-3-small-128k-instruct</b> Chat completion
<b>Phi-3.5-vision-instruct</b> Chat completion	<b>Phi-3.5-mini-instruct</b> Chat completion	<b>Phi-3.5-MoE-instruct</b> Chat completion	<b>Phi-3-vision-128k-instruct</b> Chat completion	<b>Meta-Llama-3.1-8B-instruct</b> Chat completion
<b>Meta-Llama-3.1-8B</b> Text generation	<b>Meta-Llama-3.1-70B-Instruct</b> Chat completion	<b>Meta-Llama-3.1-70B</b> Text generation	<b>Meta-Llama-3-8B-instruct</b> Chat completion	<b>Meta-Llama-3-8B</b> Text generation

## Pattern 1: Build a GenAI chatbot

This example focuses on setting up an Azure OpenAI GPT model to create an engaging chatbot. The model is configured to use custom data for tailored interactions, using various customizations and deployment techniques within Azure AI Foundry. To set up this model, you will need access to Azure AI Foundry and an Azure subscription.

Follow the steps below to create and deploy an engaging chatbot using Azure AI Foundry:

1. Go to <https://ai.azure.com> and log in with your Entra ID username and password.
2. Select + **New Project** in the top-right corner to create a new project.
3. Enter a unique project name. Choose a hub to host the project.
4. In Azure AI Foundry, navigate to **Deployments** in the left panel to deploy a chat model.



### Customer Story

Ally Financial (Ally) wanted to enhance the customer experience its call-center associates provided while producing detailed documentation for each call.

With the conversational-AI capabilities of Azure OpenAI, Ally Financial **cut associates' post-call effort by 30 percent—with the expectation of a 50 percent reduction in the near future.**



Read more

5. Select the Azure OpenAI model that fits your use case. Then, click on the **Confirm** button.
6. On the **Deploy model** page, enter a name for the deployment and select **Deploy**.
7. Once deployed, go to the **Details** tab and select **Open in playground** to begin testing your chatbot.

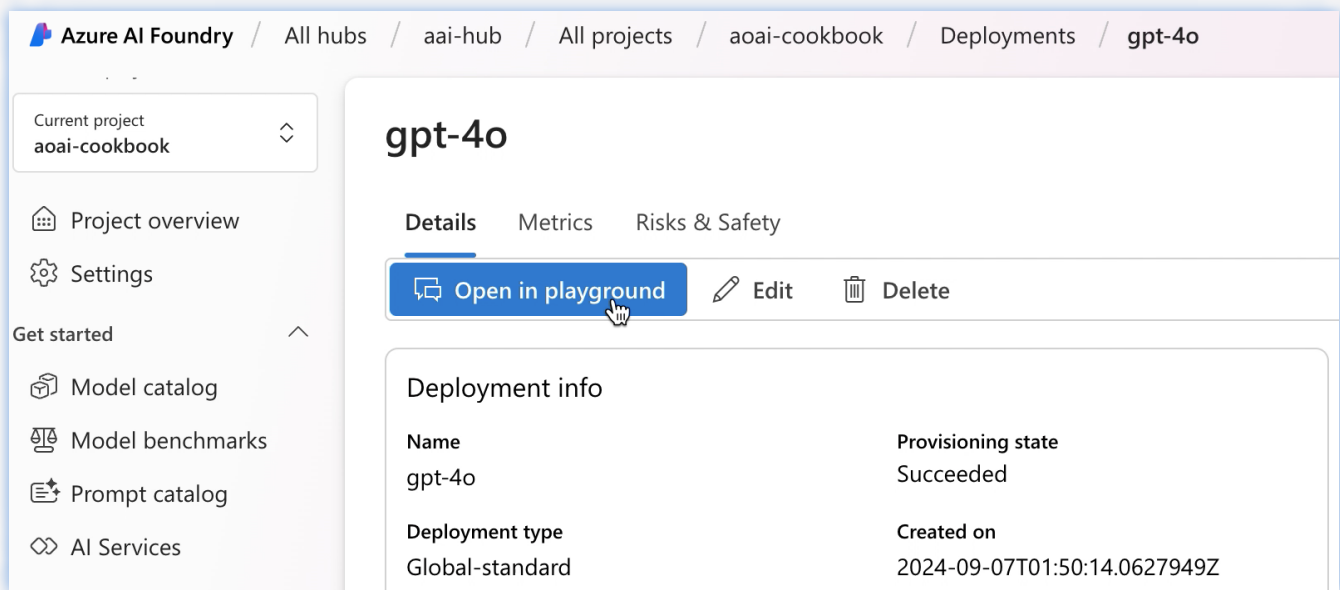


Figure 3: Open in playground

8. Enter a **System message** to guide how the chatbot interacts with users. Define the tone, boundaries, and topics. To learn more about the system message framework on Azure, visit <https://learn.microsoft.com/azure/ai-services/openai/concepts/system-message>.
9. Start a conversation by asking a question: **What is Azure OpenAI?**
10. In **Chat playground**, select **Deploy/...as a web app** after testing your chatbot in the playground.
11. Provide the necessary deployment details: web app name, Azure subscription, resource group, and pricing plan. Deploy the web app and wait for the process to complete.

12. Once deployed, launch the app and ask questions.

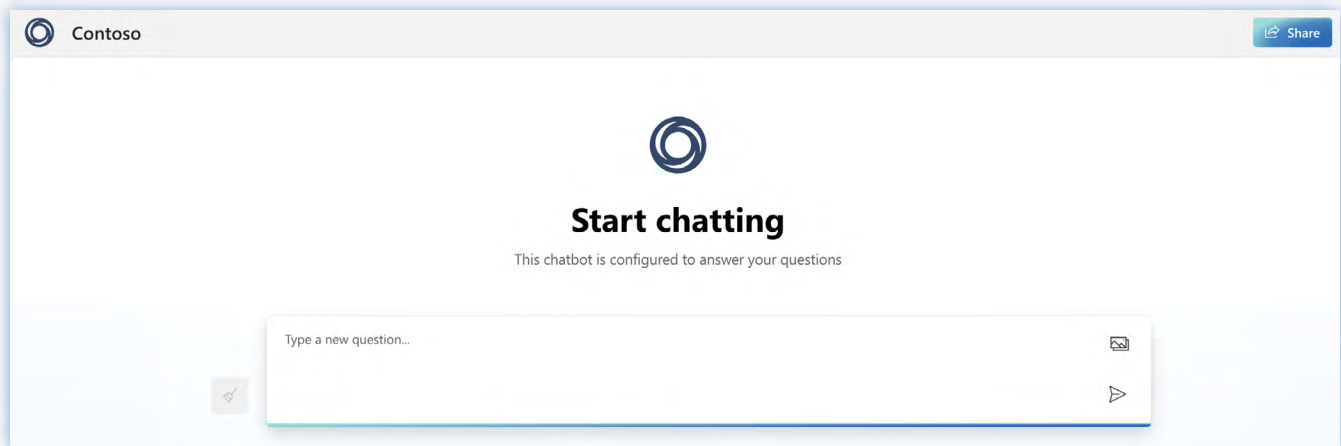


Figure 4: The web app

Building a GenAI chatbot in Azure AI Foundry is an efficient way to innovate quickly using a broad selection of cutting-edge models. The process starts by creating a project to organize your work, setting up resources, and ensuring security configurations. You can then deploy models like GPT-4 or GPT-3.5, and even explore multimodal capabilities with DALL-E 3 or GPT-4 Turbo with Vision for richer, more interactive experiences. This flexibility allows users to evaluate and choose the best model for their specific use case.

Beyond model selection, Azure AI Foundry ensures responsible AI practices with built-in safety measures. Azure AI Content Safety helps mitigate risks by monitoring prompt inputs and GenAI outputs and applying filtering and measurement tools. Microsoft's Customer Copyright Commitment also provides peace of mind by protecting against copyright infringement, ensuring that both you and your customers can innovate securely. Finally, once your model is deployed, you can test it in the chat playground. Here, you provide system messages to guide the chatbot's behavior and refine it based on the responses, ensuring it aligns with your project goals.

## Pattern 2: Build a personalization engine

Building a recommendation engine using Azure AI Foundry and configuring a personalized recommendation model enables the engine to analyze user data to generate personalized suggestions, such as content or products, and adapt to user input in real time. This helps businesses gain deeper insights into user behavior, allowing them to predict preferences and enhance user engagement by delivering more relevant experiences. To build such an engine, you will need permissions to create projects and deploy Azure OpenAI models, and an Azure subscription with sufficient resources for deploying new models.

To create a recommendation engine in Azure AI Foundry, follow the steps below:

1. Go to <https://ai.azure.com> and log in using your Azure credentials.
2. Navigate to **Deployments** in the left pane and select + **Deploy Model** from the top-left corner of the screen.

### LIONBRIDGE

#### Customer Story

Lionbridge Technologies, LLC (Lionbridge) sought to revolutionize the localization industry with personalized, high-speed content delivery. Using the conversational AI and large language models (LLMs) in Azure OpenAI, Lionbridge improved its ability to deliver tailored content in multiple languages, **cutting project turnaround times by up to 30%.**

This advancement allowed Lionbridge to streamline workflows, creating a hyper-personalized experience for its global clients.



Read more

3. Select the **GPT-4 (0613)** model for its versatility and token efficiency, which is ideal for personalization and recommendation use cases. Click on **Confirm** to deploy it.
4. On the **Deploy model** page, enter a name for the deployment and select **Deploy**.
5. In Azure AI Foundry, go to the chat playground.
6. In the chat playground, select **Add your data**, and then **Add a new data source**.
7. On the **Upload file** page, you can drag and drop your file or browse for a file to upload. For this example, use a sample data file from Microsoft at [https://github.com/Azure-Samples/azure-search-openai-demo/blob/main/data/employee\\_handbook.pdf](https://github.com/Azure-Samples/azure-search-openai-demo/blob/main/data/employee_handbook.pdf). Once the file is uploaded and processed, the OpenAI model can use it to personalize the output in response to your prompts. This allows the model to generate contextually relevant answers based on the content of the uploaded file. Select **Next** after the file is uploaded.
8. Create a connection to your AI Search service. Index your data (for example, employee-book-info for employee handbook information). Select **Next** and wait for the process to be completed.
9. Ask the chatbot a question based on your uploaded data. Verify the response references your data. For example, you can ask questions such as **What's our mission?** You will see the response generated by the language model based on your input. Select a reference in the **References** section to display the citations used to generate this response.

Azure AI Foundry empowers users to unlock highly accurate and contextual experiences through RAG, tailored to their data. With a full-stack approach, this platform ensures that GenAI solutions deliver up-to-date, relevant information by seamlessly integrating Azure Cosmos DB as a vector database and Azure AI Search as a cutting-edge retrieval system. You will gain real-time access to data and surface the most pertinent responses for any GenAI scenario. Whether you're transforming workflows or enhancing applications, Azure AI services enable you to infuse AI into various aspects of app conversations, insights, speech, translation, and more—all without needing deep AI expertise. With RAG, Azure AI Foundry offers a powerful foundation to deliver differentiated, contextual experiences that drive meaningful outcomes.

## Pattern 3: Build your own agent

To build a custom agent, you must first create an Azure OpenAI resource. To do so, follow these steps:

1. Log in to the Azure portal (<https://portal.azure.com>).
2. Select **Create a resource**, the plus button, located in the top-right corner of your screen under Azure services.
3. Search for **Azure OpenAI** in the search bar.
4. Select **Azure OpenAI** in the dropdown.
5. Select **Azure OpenAI**; it will be the first item on the next screen.
6. Select **Create**. This will bring up the Azure OpenAI resource creation page.



### Customer Story

PwC is exploring how the generative AI capabilities from Azure OpenAI running on Azure global infrastructure can support a vision that includes: **increased audit efficiency and transparency; reduction of unnecessary audit burden on clients; and enhanced data acquisition and utilization.** PwC has a focus on relevant data regulations and compliance protocols, as well as robust data security supported by Microsoft.



Read more

7. Fill out the **Basics** tab by selecting your Azure subscription, creating a new resource group, selecting an Azure region, selecting **Standard S0** for your pricing tier, and naming your Azure OpenAI resource. Your resource name can only include alphanumeric characters and hyphens, and it cannot start or end with a hyphen.
8. Click **Next** to advance to the **Network** tab.
9. On the **Network** tab, select **All networks, including the internet, can access this resource** for networking and select **Next** to advance to the **Tags** tab.
10. On the **Tags** tab, add any tags you wish, or simply leave it blank.
11. Click **Next** to advance to the **Review + submit** tab.
12. Click **Create**.

Once you have created an Azure OpenAI resource, the next step is to deploy a model. The best way to do this is through **Azure AI Foundry**, Azure's premier development tool for all things AI. To access it, follow the link to <https://ai.azure.com/>, click **Azure OpenAI** on the left-hand side of your screen, and select the Azure OpenAI resource you just created.

Once there, follow these steps to deploy a model:

1. Click **Deployments** on the left-hand side of your screen.
  2. Click **Deploy Model** next to the plus button.
  3. Click **Deploy base model**.
  4. Select **gpt-4o** and then **Confirm**. **GPT-4o** is the most dynamic OpenAI model available and is particularly suitable for content innovation.
- While GPT-4o is the latest model at the time of writing, more advanced models may be out by the time you read this.
5. Leave all settings as is and click **Deploy**.

Now, you have everything you need to easily create your agent. To continue, follow these steps:

1. Click **Chat** on the left-hand side of your screen.
2. Select **gpt-4o** under **Deployment**.



3. Click **Add your data** in the setup panel on the left-hand side of your screen, and then click + **Add a data source**.
  4. Select **Upload files** from the dropdown menu.
  5. Select your **Azure Blob storage resource** and your **Azure AI Search resource** from the dropdown menus.
- To chat with your own data, you need to have data loaded into an Azure Blob storage resource. You will also need to have an Azure AI Search resource. Please create these resources if you haven't already.
6. Enable cross-origin resource sharing (CORS) by clicking **Turn on**.
  7. Create a name for your search index under **Enter the index name**.
  8. Next, you need to write a system message. **System messages** are used to prime your model and provide it with context and instructions on how it should interact with end users. Your system message determines the primary purpose and personality of your agent.

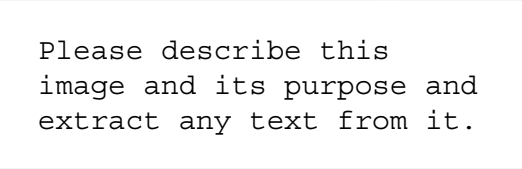
9. Click **System message** in the setup panel on the left-hand side of your screen, and, in the **Instructions and context** box, write the following:

```
You are an agent for  
a large enterprise  
organization. Your role is  
to generate a professional  
and polished first draft  
of content based on the  
information provided by  
the user. The content  
must reflect the high  
standards and formal tone  
expected in enterprise  
communications.
```

10. Click **Apply Changes** in the upper left-hand corner of the screen.
11. Try the following prompt to test out your agent:

```
Write an e-mail asking  
a vendor for an update on  
the most recent purchase  
order I sent. They are  
3 days late.
```

12. This agent is also multimodal. Try uploading an image with text on it by clicking the paperclip button and asking the chatbot:



Please describe this image and its purpose and extract any text from it.

You can also ask it a question about a data source that you have connected to. For example, if you've uploaded instructions into your Azure Blob storage resource, try asking the chatbot to reproduce those instructions.

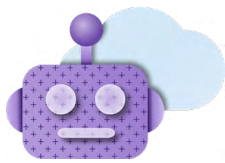
13. Now that you've created an agent, the next step is to deploy it to a web app that other employees can use. Click **Deploy** in the center of your screen to access a dropdown, and click **As a web app**.

14. Fill out the web app form by giving it a name, subscription, resource group, and location.
15. Select **Free (F1)** for the pricing plan.
16. Leave **Enable chat history in the web app** unchecked, as this will incur costs to your account.
17. Click **Deploy**.

After waiting for ten minutes, a web version of your agent will be available on your Azure account, all with a few clicks of a button! You can easily configure other users to have access and provide them with links. This will enable them to use this agent to draft content of all types, ensuring that everyone in your enterprise maintains a professional, courteous tone.

# How Azure OpenAI drives customer success

Azure OpenAI offers developers and businesses a powerful platform that blends cutting-edge AI models with robust enterprise-grade capabilities, allowing organizations to innovate while maintaining full control over their data, security, and trust. Here are some key features that illustrate how Azure OpenAI supports these objectives:



## Trustworthy AI:

Azure OpenAI is built with enterprise needs in mind, ensuring a trusted environment where sensitive data is protected by advanced encryption, privacy safeguards, and a comprehensive security framework. It helps organizations by providing:

- Data privacy
- Content safety
- Compliance



## Data integrity and insights:

One of Azure OpenAI's key advantages is its ability to enhance AI applications with secure, real-time access to high-quality data sources. This ensures that every interaction and output is contextually relevant and trustworthy, offering:

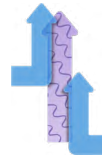
- Grounding AI in enterprise data
- RAG



### Custom models and multimodal capabilities:

Azure OpenAI provides flexibility to tailor models according to unique business needs. With the ability to create and fine-tune models, businesses can build innovative solutions with ease, using:

- RAG and fine-tuning
- Multimodal interactions



### Scalability and ease of use:

Azure OpenAI is designed to simplify the deployment of AI models at scale, making it accessible even for organizations that are just starting their AI journey. Out-of-the-box integrations and an AI-ready infrastructure enable:

- Seamless deployment
- Developer agility

Azure OpenAI helps developers unlock the potential of AI across a wide range of applications. Whether it's natural language processing, computer vision, or multimodal experiences, developers can use pre-trained models to build intelligent solutions quickly. By incorporating powerful models such as GPT-4 for text analysis and DALL-E for image generation, developers can transform workflows without the need for deep AI expertise.

# Azure OpenAI for innovative and responsible AI solutions

Azure OpenAI empowers organizations to innovate rapidly by utilizing a broad range of cutting-edge models, such as GPT-4 and GPT-3.5, and even multimodal models, such as GPT-4 Turbo with Vision or DALL-E 3. These models help streamline processes, enhance customer interactions, and deliver personalized user experiences.

With Azure AI Foundry, businesses can not only deploy powerful AI solutions but also ensure responsible AI practices with built-in safety measures such as Azure AI Content Safety and Microsoft's Customer Copyright Commitment, which are critical for deploying AI in enterprises. To explore Azure OpenAI further, the full e-book, *Azure OpenAI Cookbook*, offers practical, hands-on examples and detailed information that enables developers to maximize the platform's potential.



## Next steps

1. Explore Azure OpenAI within [Azure AI Foundry](#)
2. Discover industry-leading coding and language models tailored to your needs with [Azure OpenAI](#)
3. Click [here](#) to read the full e-book, *Azure OpenAI Cookbook*