GPT-4o Realtime API for speech and audio (Preview)

Article • 02/07/2025

> **Note**
>
> This feature is currently in public preview. This preview is provided without a service-level agreement, and we don't recommend it for production workloads. Certain features might not be supported or might have constrained capabilities. For more information, see Supplemental Terms of Use for Microsoft Azure Previews .

Azure OpenAI GPT-4o Realtime API for speech and audio is part of the GPT-4o model family that supports low-latency, "speech in, speech out" conversational interactions. The GPT-4o audio `realtime` API is designed to handle real-time, low-latency conversational interactions, making it a great fit for use cases involving live interactions between a user and a model, such as customer support agents, voice assistants, and real-time translators.

Most users of the Realtime API need to deliver and receive audio from an end-user in real time, including applications that use WebRTC or a telephony system. The Realtime API isn't designed to connect directly to end user devices and relies on client integrations to terminate end user audio streams.

Supported models

The GPT 4o real-time models are available for global deployments.

- `gpt-4o-realtime-preview` (version `2024-12-17`)
- `gpt-4o-mini-realtime-preview` (version `2024-12-17`)
- `gpt-4o-realtime-preview` (version `2024-10-01`)

See the models and versions documentation for more information.

API support

Support for the Realtime API was first added in API version `2024-10-01-preview`. Use the latest `2024-12-17` model version.

Prerequisites

- An Azure subscription. Create one for free .
- Python 3.8 or later version . We recommend using Python 3.10 or later, but having at least Python 3.8 is required. If you don't have a suitable version of Python installed, you can follow the instructions in the VS Code Python Tutorial for the easiest way of installing Python on your operating system.
- An Azure OpenAI resource created in one of the supported regions. For more information about region availability, see the models and versions documentation.
- Then, you need to deploy a `gpt-4o-mini-realtime-preview` model with your Azure OpenAI resource. For more information, see Create a resource and deploy a model with Azure OpenAI.

Microsoft Entra ID prerequisites

For the recommended keyless authentication with Microsoft Entra ID, you need to:

- Install the Azure CLI used for keyless authentication with Microsoft Entra ID.
- Assign the `Cognitive Services User` role to your user account. You can assign roles in the Azure portal under **Access control (IAM)** > **Add role assignment**.

Deploy a model for real-time audio

To deploy the `gpt-4o-mini-realtime-preview` model in the Azure AI Foundry portal:

1. Go to the Azure OpenAI Service page in Azure AI Foundry portal. Make sure you're signed in with the Azure subscription that has your Azure OpenAI Service resource (with or without model deployments.)
2. Select the **Real-time audio** playground from under **Playgrounds** in the left pane.
3. Select **+ Create new deployment** > **From base models** to open the deployment window.
4. Search for and select the `gpt-4o-mini-realtime-preview` model and then select **Deploy to selected resource**.
5. In the deployment wizard, select the `2024-12-17` model version.
6. Follow the wizard to finish deploying the model.

Now that you have a deployment of the `gpt-4o-mini-realtime-preview` model, you can interact with it in real time in the Azure AI Foundry portal **Real-time audio** playground or