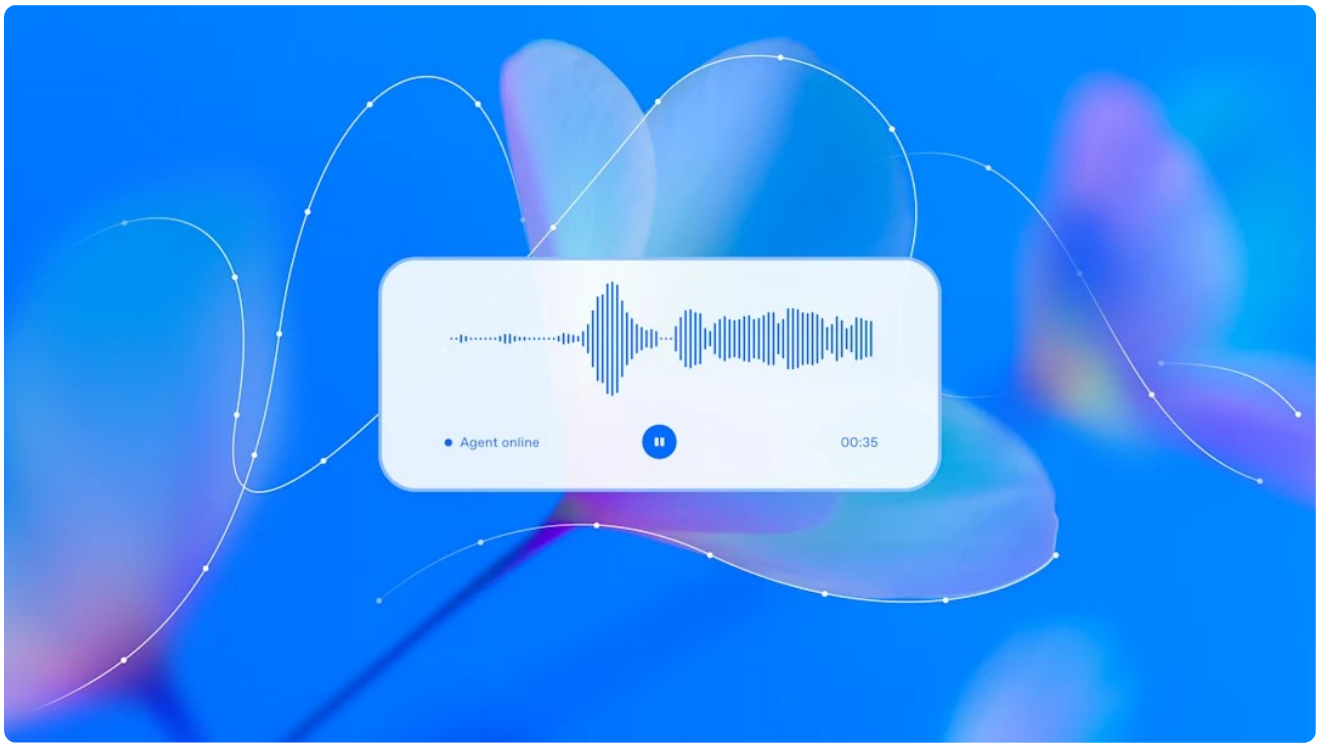
August 28, 2025 Product Release

Introducing gpt-realtime and Realtime API updates for production voice agents

We’re releasing a more advanced speech-to-speech model and new API capabilities including MCP server support, image input, and SIP phone calling support.



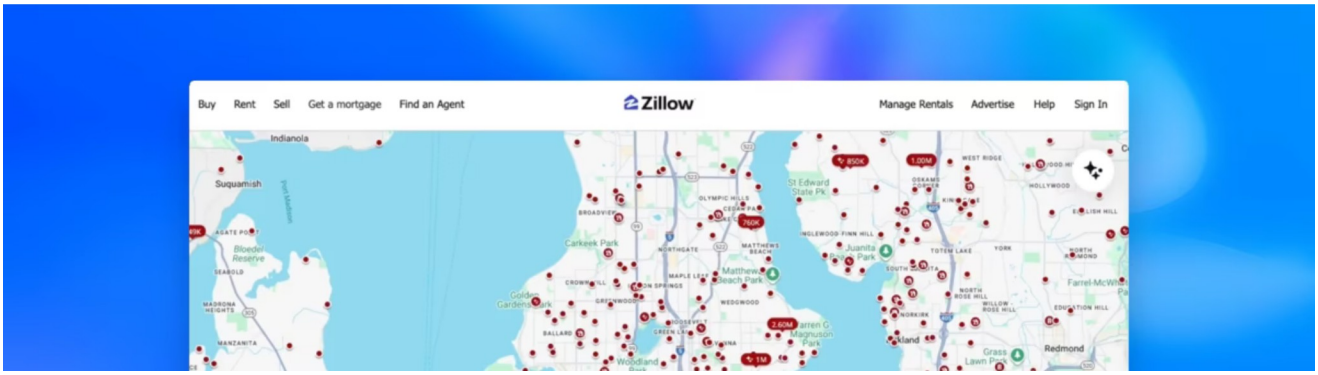
Listen to article 8:11 Share

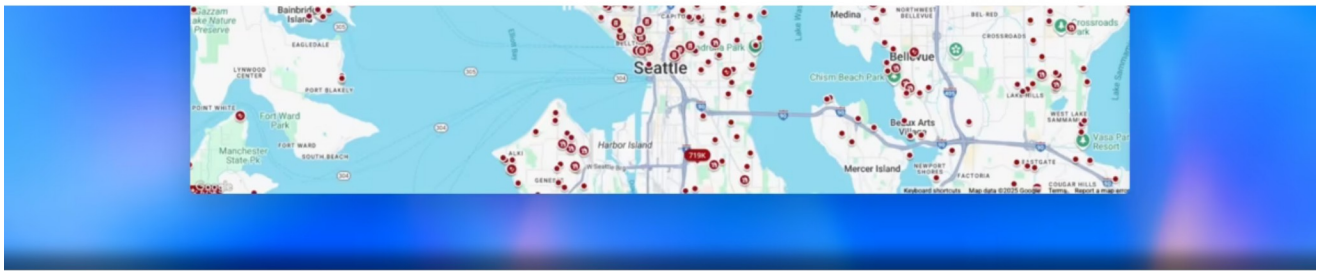
developers and enterprises to build reliable, production-ready voice agents. The API now supports remote MCP servers, image inputs, and phone calling through Session Initiation Protocol (SIP), making voice agents more capable through access to additional tools and context.

We're also releasing our most advanced speech-to-speech model yet—`gpt-realtime`. The new model shows improvements in following complex instructions, calling tools with precision, and producing speech that sounds more natural and expressive. It's better at interpreting system messages and developer prompts—whether that's reading disclaimer scripts word-for-word on a support call, repeating back alphanumerics, or switching seamlessly between languages mid-sentence. We're also releasing two new voices, Cedar and Marin, which are available exclusively in the Realtime API starting today.

Since we first introduced the Realtime API in public beta last October, thousands of developers have built with the API and helped shape the improvements we're releasing today—optimized for reliability, low latency, and high quality to successfully deploy voice agents in production. Unlike traditional pipelines that chain together multiple models across speech-to-text and text-to-speech, the Realtime API processes and generates audio directly through a single model and API. This reduces latency, preserves nuance in speech, and produces more natural, expressive responses.

Zillow | T-Mobile | StubHub | Oscar Health | Lemonade

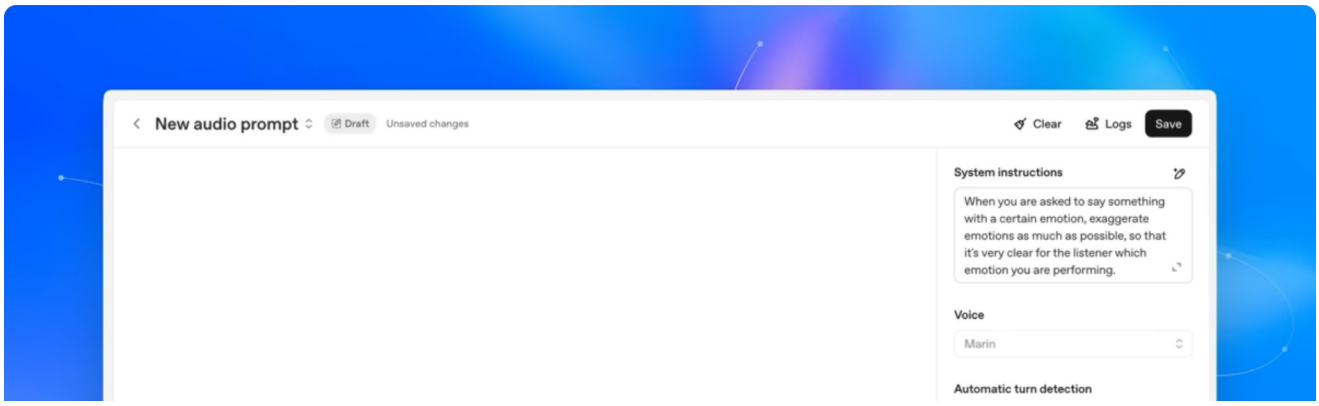


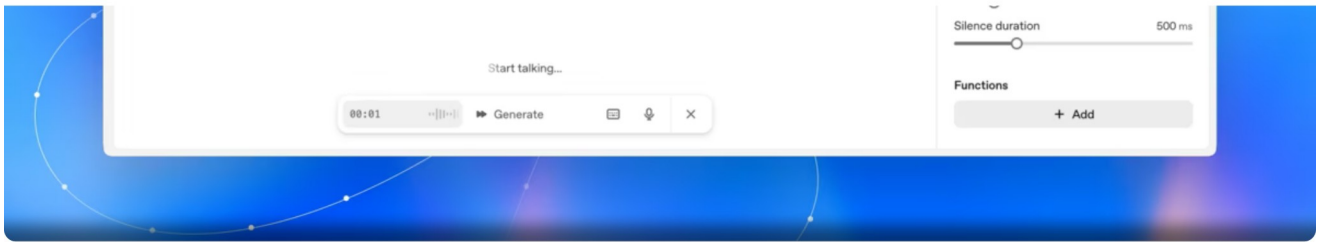


“The new speech-to-speech model in OpenAI's Realtime API shows stronger reasoning and more natural speech—allowing it to handle complex, multi-step requests like narrowing listings by lifestyle needs or guiding affordability discussions with tools like our BuyAbility score. This could make searching for a home on Zillow or exploring financing options feel as natural as a conversation with a friend, helping simplify decisions like buying, selling, and renting a home.”
– Josh Weisberg, Head of AI at **Zillow**

Introducing gpt-realtime

The new speech-to-speech model—`gpt-realtime`—is our most advanced, production-ready voice model. We trained the model in close collaboration with customers to excel at real-world tasks like customer support, personal assistance, and education—aligning the model to how developers build and deploy voice agents. The model shows improvements across audio quality, intelligence, instruction following, and function calling.





Audio quality

Natural-sounding conversation is critical for deploying voice agents in the real world. Models need to speak with the intonation, emotion, and pace of a human to create an enjoyable experience and encourage continuous conversation with users. We trained `gpt-realtime` to produce higher-quality speech that sounds more natural and can follow fine-grained instructions, such as “speak quickly and professionally” or “speak empathetically in a French accent.”

We’re releasing two new voices in the API, Marin and Cedar, with the most significant improvements to natural-sounding speech. We’re also updating our existing eight voices to benefit from these improvements.

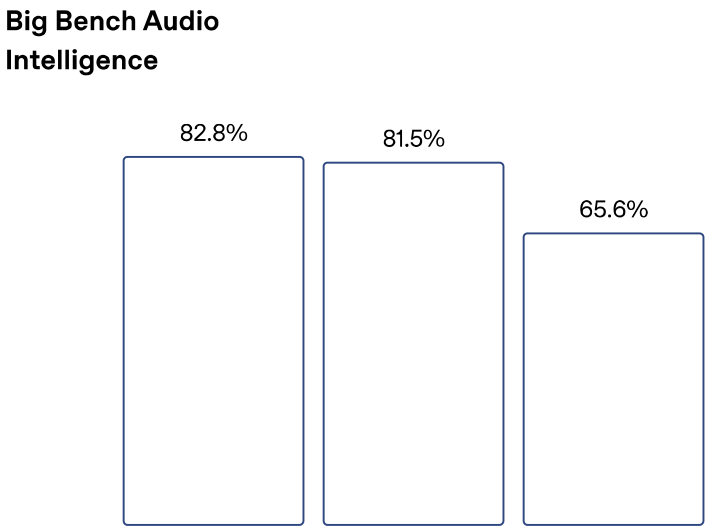
Voice sample - Marin

Voice sample - Cedar

Intelligence and comprehension

`gpt-realtime` shows higher intelligence and can comprehend native audio with greater accuracy. The model can capture non-verbal cues (like laughs), switch languages mid-sentence, and adapt tone (“snappy and professional” vs. “kind and empathetic”). According to internal evaluations, the model also shows more accurate

Bench Audio eval measuring reasoning capabilities, `gpt-realtime` scores 82.8% accuracy—beating our previous model from December 2024, which scores 65.6%.

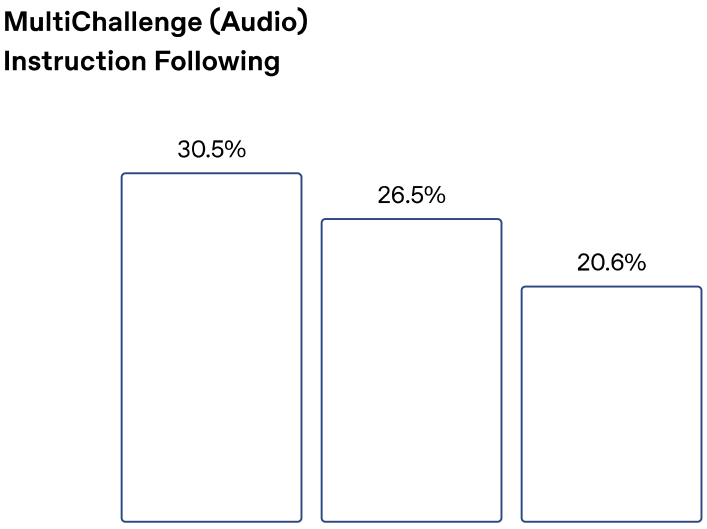


The Big Bench Audio benchmark is an evaluation dataset for assessing the reasoning capabilities of language models that support audio input. This dataset adapts questions from Big Bench Hard—chosen for its rigorous testing of advanced reasoning—into the audio domain.

Instruction following

When building a speech-to-speech application, developers give a set of instructions to the model on how to behave, including how to speak, what to say in a certain situation, and what to do or not do. We've focused our improvements on the adherence to these instructions, so that even minor directions carry more signal for the model. On the

December 2024, which scores 20.6%.



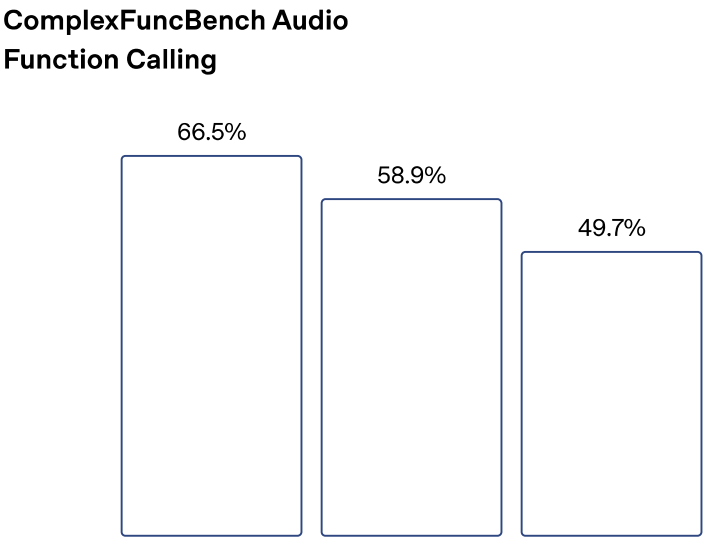
MultiChallenge evaluates how well LLMs handle multi-turn conversations with humans. It focuses on four categories of realistic challenges that current frontier models struggle with. These challenges require models to combine instruction-following, context management, and in-context reasoning simultaneously. We converted an audio-friendly subset of the test questions from text-to-speech to create an audio version of this evaluation.

Function calling

To build a capable voice agent with a speech-to-speech model, the model needs to be able to call the right tools at the right time to be useful in production. We've improved function calling on three axes: calling relevant functions, calling functions at the appropriate time, and calling functions with appropriate arguments (resulting in higher

performance, `gpt-realtime` scores 66.5%, while our previous model from December 2024 scores 49.7%.

We've also made improvements to asynchronous function calling. Long-running function calls will no longer disrupt the flow of a session—the model can continue a fluid conversation while waiting on results. This feature is available natively in `gpt-realtime`, so developers do not need to update their code.



ComplexFuncBench measures how well models handle challenging function calling tasks. It evaluates performance across scenarios like multi-step calls, reasoning about constraints or implicit parameters, handling very long inputs. We converted the original text prompts into speech to build this evaluation for our model.

Remote MCP server support

You can enable MCP support in a Realtime API session by passing the URL of a remote MCP server into the session configuration. Once connected, the API automatically handles the tool calls for you, so there's no need to wire up integrations manually.

This setup makes it easy to extend your agent with new capabilities—just point the session to a different MCP server, and those tools become available right away. To learn more about configuring MCP with Realtime, check out this guide.

JavaScript

```
// POST /v1/realtime/client_secrets
{
  "session": {
    "type": "realtime",
    "tools": [
      {
        "type": "mcp",
        "server_label": "stripe",
        "server_url": "https://mcp.stripe.com",
        "authorization": "{access_token}",
        "require_approval": "never"
      }
    ]
  }
}

```

Image input

With image inputs now supported in `gpt-realtime`, you can add images, photos, and screenshots alongside audio or text to a Realtime API session. Now the model can

Instead of treating an image like a live video stream, the system treats it more like adding a picture into the conversation. Your app can decide which images to share with the model and when to share them. This way, you stay in control of what the model sees and when it responds.

Check out our docs to get started with image input.

JavaScript

```
{
    "type": "conversation.item.create",
    "previous_item_id": null,
    "item": {
        "type": "message",
        "role": "user",
        "content": [
            {
                "type": "input_image",
                "image_url": "{some_base6
            }
        ]
    }
}

```

Additional capabilities

We've added several other features to make the Realtime API easier to integrate and more flexible for production use.

- **Session Initiation Protocol (SIP) support:** Connect your apps to the public phone network, PBX systems, desk phones, and other SIP endpoints with direct support in the Realtime API. Read about it in docs.

developer messages, tools, variables, and example user/assistant messages—across Realtime API sessions, like in the Responses API. Learn more in docs.

Safety & privacy

The Realtime API incorporates multiple layers of safeguards and mitigations to help prevent misuse. You can learn more about our safety approach and system card details in the beta announcement blog. We employ active classifiers over Realtime API sessions, meaning certain conversations can be halted if they are detected as violating our harmful content guidelines. Developers can also easily add their own additional safety guardrails using the Agents SDK.

Our usage policies prohibit repurposing or distributing outputs from our services for spam, deception, or other harmful purposes. Developers must also make it clear to end users when they're interacting with AI, unless it's already obvious from the context. The Realtime API uses preset voices to help prevent malicious actors from impersonating others.

The Realtime API fully supports EU Data Residency for EU-based applications and is covered by our enterprise privacy commitments.

Pricing & availability

The generally available Realtime API and new `gpt-realtime` model are available to all developers starting today. We're reducing prices for `gpt-realtime` by 20% compared to `gpt-4o-realtime-preview`—$32 / 1M audio input tokens ($0.40 for cached input tokens) and $64 / 1M audio output tokens (see detailed pricing). We've also added fine-grained control for conversation context to let developers set intelligent token limits and truncate multiple turns at a time, significantly reducing cost for long sessions.

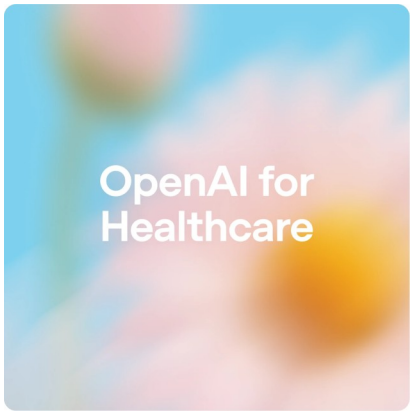
Playground, and view our Realtime API prompting guide.

Livestream replay

2025

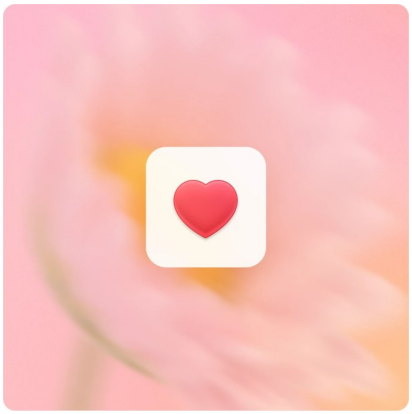
Author

OpenAI



Introducing OpenAI for Healthcare

Product Jan 8, 2026



Introducing ChatGPT Health

Product Jan 7, 2026

Introducing GPT-5.2-Codex

Product Dec 18, 2025

Ask ChatGPT

Our Research

Research Index

Research Overview

Research Residency

OpenAI for Science

Latest Advancements

GPT-5

OpenAI o3

OpenAI o4-mini

GPT-4o

ChatGPT

Explore ChatGPT

Business

Enterprise

Education

Pricing

Download

Sora

Sora Overview

Features

For Business

Business Overview

Solutions

Contact Sales

Company

About Us

Our Charter

Foundation

Careers

Brand

Terms & Policies

Terms of Use

Privacy Policy

Other Policies