OpenAI

GPT-Live

System Card

2026-07-08

1 Introduction

GPT-Live-1 and GPT-Live-1 mini are a new generation of voice models designed to make conversations with AI feel more natural and intelligent.

These models — enabled by research advances — are full-duplex, meaning they can listen and respond continuously instead of waiting for a clearly defined turn to end. This means they can follow pauses, interruptions, and changes in pace, and decide in the moment whether to respond or keep listening.

GPT-Live-1 will be the default voice model for paid users, while GPT-Live-1 mini will be the default model for free users.

The most important things to know about our safety work for this launch are that:

1. We trained these new models to respond safely, using the same infrastructure we rely on for training our flagship models. They can also delegate more complex work to our other models, and when they do, the resulting work will reflect the safety training of the underlying model that is doing that work. The evaluations in this card describe how the new GPT-Live-1 models perform with delegation, matching the deployment context.
2. These models have system-level safety integrations that are designed to be on par with the existing safety stack for text models, while adopting some new safeguards specifically for the new voice modality: inputs and generated outputs are checked as the conversation unfolds; when potentially unsafe content is detected, the system can steer or interrupt the response, play a spoken safety message, provide support resources in text, or, in higher-risk cases, end the voice conversation.
3. As part of our safety work for this launch, we built new evaluations that focus specifically on the distinctive ways that people use voice models, as opposed to text-based chats, and on observations from real-world use of the existing Advanced Voice Mode. We report those results below.
4. The same monitoring, review, and enforcement infrastructure that we use for our text models also applies to the GPT-Live models, enabling us to measure prevalence, detect abuse, and enforce our safety policies.

2 Model Data and Training

Like OpenAI's other models, GPT-Live-1 and GPT-Live-1 mini were trained on diverse datasets, including information that is publicly available on the

internet, information that we partner with third parties to access, and information that our users or human trainers and researchers provide or generate. Our data processing pipeline includes rigorous filtering to maintain data quality and mitigate potential risks. We use advanced data filtering processes to reduce personal information from training data. We also employ safety classifiers to help prevent or reduce the use of harmful or sensitive content, including explicit materials such as sexual content involving a minor.

Note that comparison values from previously launched models are from the latest versions of those models, so may vary slightly from values published at launch for those models.[1]

3 Model Safety

3.1 Voice-Native Evaluations for Disallowed Content

3.1.1 Voice-Native Evaluations: Production Prompts

In these newly developed evaluations, we use real audio examples from users who have chosen to share their voice interactions to help improve our models. Before these examples are used, they are processed through our privacy and eligibility safeguards, including checks for user permissions and deletion/opt-out status, filtering of ineligible data, and steps to reduce personal information through PII scrubbing and de-identification. We then transcribe the audio, generate the model's response, and evaluate that response for safety.

We compare the new GPT-Live models to their respective predecessors, the models that power Advanced Voice Mode (AVM).

These evaluations are not prevalence weighted, meaning they do not reflect rates of safety performance we see in real usage. Instead, these evaluations were meant to be difficult. For each category below, the evaluation is built

[1] GPT-Live-1 and GPT-Live-1 mini are intended to be used in accordance with OpenAI's Usage Policies, Service Terms, and Terms of Use. These policies apply universally to OpenAI services and are designed to ensure safe and responsible usage of AI technology. You can review OpenAI's Usage Policies at openai.com/policies/usage-policies/ .

If you need assistance with respect to GPT-Live-1 and GPT-Live-1 mini, you can find further information on OpenAI's website (openai.com), or you can contact OpenAI Support by opening the chat bubble icon displayed at the bottom-right of help.openai.com.
A list of the languages that ChatGPT currently supports can be found here.

around cases in which the existing AVM models were not yet giving ideal responses.

Table 1: Voice-Native Evaluations: Production Prompts

	AVM	GPT-Live-1	AVM mini	GPT-Live-1 mini
Sexual	0.96	0.97	0.97	0.95
Illicit behavior	0.74	0.97	0.60	0.94
Mental health	0.90	0.90	0.78	0.84
Personal data	0.96	0.96	0.95	0.95
Emotional reliance	0.88	0.82	0.78	0.78
Self-harm	0.89	0.96	0.81	0.92

We observe that the GPT-Live models generally provide equal or better safety performance across these adversarially selected prompts than AVM models. GPT-Live-1 shows a slight regression on emotional reliance from 0.88 to 0.82, and GPT-Live-1 mini shows a slight regression on sexual content from 0.97 to 0.95. Note that neither of these are statistically significant.

3.1.2 Voice-Native Evaluations: Synthetic Prompts

The below evaluations are similar to the ones above, except that for these we synthetically generate audio prompts to target challenging edge cases. The text for these prompts is generated from safety policies and related guidance, targeting specific safety categories, policy boundaries, and difficult cases. This approach allows us to deliberately cover a broad range of scenarios, including rare or hard-to-sample safety-relevant situations. We then convert the prompts into speech and use them as audio inputs, allowing us to assess whether models apply the intended safety behavior when safety-relevant content is presented in spoken form. These evaluations focus more intensively on the areas where we believe our existing models are least likely to give ideal responses. As a result, the numbers in the table below are likewise not a guide to safety performance across all production traffic.

Table 2: Voice-Native Evaluations: Synthetic Prompts

	AVM	GPT-Live-1	AVM mini	GPT-Live-1 mini
Sexual	0.76	0.97	0.85	0.95
Illicit behavior	0.63	0.97	0.63	0.97
Mental health	0.57	0.84	0.47	0.81
Personal data	0.90	0.97	0.79	0.97
Emotional reliance	0.72	0.91	0.72	0.89
Self-harm	0.72	0.98	0.71	0.96
Hate	0.87	1.00	0.82	0.98
Gore	0.61	0.97	0.80	0.96

We observe that the GPT-Live models uniformly provide equal or better safety performance across these synthetic, adversarially selected prompts than AVM models.

The synthetic evaluation is designed to assess different risk surfaces than the production evaluation, so performance may vary between them. The synthetic set tests targeted, policy-grounded adversarial scenarios, including rare cases that may be underrepresented in production data. Strong performance on this set indicates that safety training is transferring for clear, intentionally constructed risks, but may not translate to production behavior. The production set reflects real-world user behavior and often includes more ambiguous or borderline context, longer interaction histories, and persistent attempts to steer the model toward unsafe outputs. Failures in the production set can be subtler and less severe, but still important. We therefore view the two benchmarks as complementary.

4 Red Teaming

OpenAI worked with a team of internal and external red teamers across languages to stress test the models' safety training with no system level mitigations. We started very early in the process to baseline performance with no model safety training, and ultimately completed two additional rounds of testing as we continued to improve safeguards. Red teaming spanned multiple categories including child-coded voice, impersonation, speaker identification, sensitive train identification, self-harm, emotional reliance, scams and manipulation, and audio-specific perturbations. Early findings were used to prioritize risk areas to focus mitigation efforts on (e.g. sexual content, emotional reliance, and self harm) while validating some other areas that were policy-compliant by default (e.g. voice cloning and impersonations). Consequent follow-up rounds validated that the mitigations we built to address the identified issues are robust.

5 Preparedness Framework

The Safety Advisory Group reviewed this launch and determined that neither GPT-Live-1 nor GPT-Live-1 mini, when operating without delegation, could plausibly be considered High in any of our Preparedness Framework's Tracked Categories – Biological and Chemical Risk, AI Self-Improvement, or Cybersecurity.

- For the Biological and Chemical domain, where GPT-Live models can use our highly capable flagship models, the GPT-Live experience inherits the safeguards of those underlying models. In addition, we've built automated monitors that may interrupt and end the call when potentially harmful conversations are detected, or degrade user experience for repeated abuse. We also take actor level enforcement actions where necessary.
- For the Cybersecurity domain, delegated work will likewise receive the safeguards associated with the model to which work is delegated. In addition, cybersecurity risk from the GPT-Live models themselves is highly constrained at launch because these models lack broad access to tools independently of the models to which they delegate, and do not have code execution capability. We will reassess the cybersecurity safeguards posture before enabling additional tools.
- AI Self-Improvement capability evals were not run, as GPT-Live-1 and GPT-Live-1 mini are less capable than GPT-5.5 Thinking across several intelligence evaluations.