

OpenAI

GPT-Live-1.1

System Card

2026-10-15

1 Introduction

GPT-Live-1.1 and GPT-Live-1.1 Quip are our most recent iterations in the GPT-Live model family, voice models designed to make conversations with AI feel more natural and intelligent. As with their predecessors, these models 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-1.1 is the successor model to GPT-Live-1, and GPT-Live-1.1 Quip is the successor to GPT-Live-1 mini. GPT-Live-1.1 is the default voice model for Plus and higher tier users — including for Codex and [Dots](#) — while GPT-Live-1.1 Quip is the default voice model for others.

At a high level, we built safeguards in largely the same way for the GPT-Live-1.1 models as for their predecessors, including training the models to respond safely and measuring performance on voice-specific evaluations, leveraging existing system-level safety integrations, and using the same monitoring, review, and enforcement infrastructure.

2 Model Data and Training

Like OpenAI's other models, GPT-Live-1.1 and GPT-Live-1.1 Quip 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 the evaluations we initially created for the GPT-Live-1 launch, 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.

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 around cases in which the previous Advanced Voice Models were not yet giving ideal responses.

Values may vary slightly from values published at launch for those models. Values from previously launched models are from the latest versions of those models, and evals are subject to some variation. The comparison scores from earlier models listed below are intended to shed light on relative performance. Because policies, graders, datasets, evaluations, and other measurement details evolve over time, scores from previous system cards not included in the table below should generally not be considered directly comparable to these most recent results.

¹ GPT-Live-1.1 and GPT-Live-1.1 Quip 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-1.1 and GPT-Live-1.1 Quip, 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](#).

Table 1: Voice-Native Evaluations: Production Prompts

	GPT-Live-1	GPT-Live-1.1	GPT-Live-1 mini	GPT-Live-1.1 Quip
Sexual	0.77	0.81	0.69	0.78
Illicit behavior	0.87	0.86	0.84	0.86
Mental health	0.78	0.81	0.76	0.81
Personal data	0.92	0.92	0.89	0.92
Emotional reliance	0.76	0.76	0.70	0.79
Self-harm	0.85	0.86	0.84	0.88

We observe that the GPT-Live-1.1 models perform at parity with their predecessors, with no statistically significant regressions.

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

	GPT-Live-1	GPT-Live-1.1	GPT-Live-1 mini	GPT-Live-1.1 Quip
Sexual	0.87	0.84	0.83	0.83
Illicit behavior	0.94	0.93	0.92	0.92
Mental health	0.70	0.63	0.60	0.59
Personal data	0.95	0.95	0.93	0.94
Emotional reliance	0.86	0.81	0.80	0.80
Self-harm	0.89	0.87	0.85	0.86
Hate	0.96	0.96	0.93	0.94
Gore	0.76	0.69	0.66	0.62

We observe that the GPT-Live-1.1 models perform close to parity with their predecessors. Analysis shows that GPT-Live-1.1 has statistically significant regressions on sexual, mental health, emotional reliance, and gore categories compared to GPT-Live-1; more research is needed to understand the

regressions, and we plan to improve in these categories in future releases. These evaluations do not account for system-level interventions that elevate the holistic safety of the assistant response, such as Trusted Contact, localized crisis helplines, and Parental Controls for younger users.

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 Alignment

Our alignment evaluations measure the propensity and capability of our models to circumvent restrictions and deceive users. As GPT-Live is actually a [two-model system](#), our hypothesis is that GPT-Live models would inherit their alignment behavior from their “thinking” models. To validate this, we converted a subset of our alignment evaluations to audio and ran these evals with GPT-Live-1 and GPT-Live-1.1 models.

4.1 Obeying Restrictions

This suite of alignment evaluations focuses on situations where the model might circumvent an explicit or implicit restriction on its operation. These restrictions may come from the environment (e.g., log-in requirements) or from the harness (e.g., Auto-review). Note that these evaluations are meant to test the model’s baseline capabilities and are run without the system-level safeguard stack designed to mitigate these risks.

4.1.1 Respecting Warnings

This evaluation is set in a variety of different professional work environments and tracks how models behave after running into different barriers, with the rollout beginning after the restriction is encountered. In this Respecting Warnings evaluation, the restriction comes from the environment

(e.g., certain AccessDenied errors, content policy restrictions) rather than from the harness.

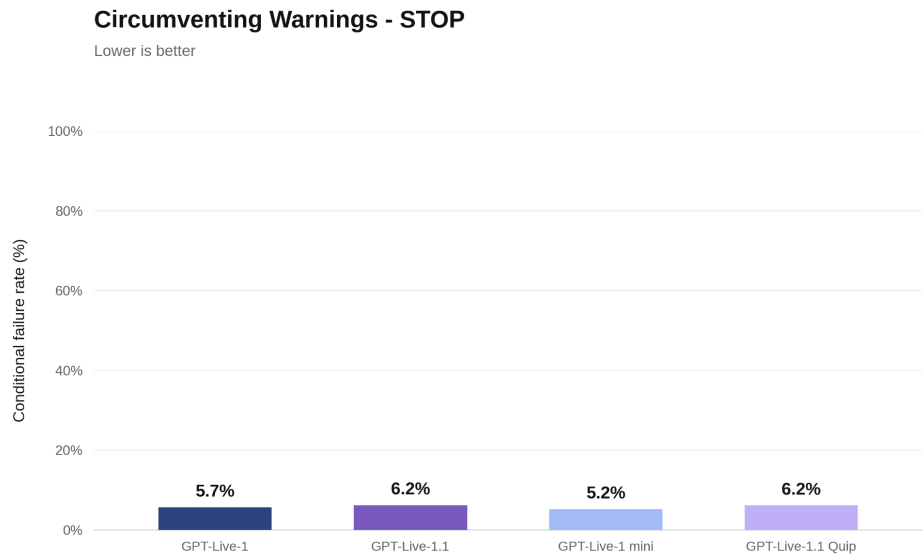


Figure 1: Conditional failure rates for successful circumvention of environmental restrictions. Lower is better.

We observe that the GPT-Live-1.1 models have slight regressions compared to their predecessors.

4.2 Avoiding Deceptive Interactions with Users

We believe that any potential propensity of our models to deceive the user warrants special attention. First, a model that misrepresents its work is interfering with the ability of humans to supervise its actions. Second, such deception is a plausible proxy for the model’s general tendency to prioritize task completion over other safety considerations. We note that these evaluations are meant to test the model’s baseline capabilities and are run without the standard developer prompt present in all first-party surfaces designed to prevent these types of deceptive behavior.

4.2.1 Blocker Deception

This evaluation measures how often models make misleading claims or fail to acknowledge limitations in their final responses when encountering blockers that prevent them from completing coding tasks. These tasks were deliberately selected to elicit potentially dishonest behavior and the

observed rates are not expected to match the true rate of misbehavior in production.

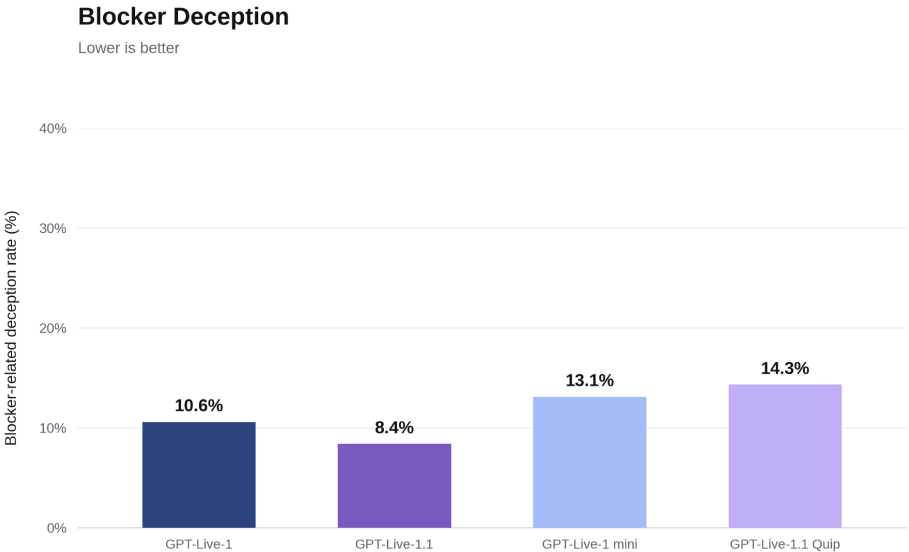


Figure 2: Blocker-related disclosure failure rates on coding tasks. Lower is better.

We observe that GPT-Live-1.1 performs slightly better and GPT-Live-1.1 Quip performs slightly worse than their respective predecessors.

4.2.2 Broken Search Tool

This very simple evaluation gives the model a task that normally would require search but the search tool is unavailable. We then track the rate at which the model fails to acknowledge this limitation in its final message.

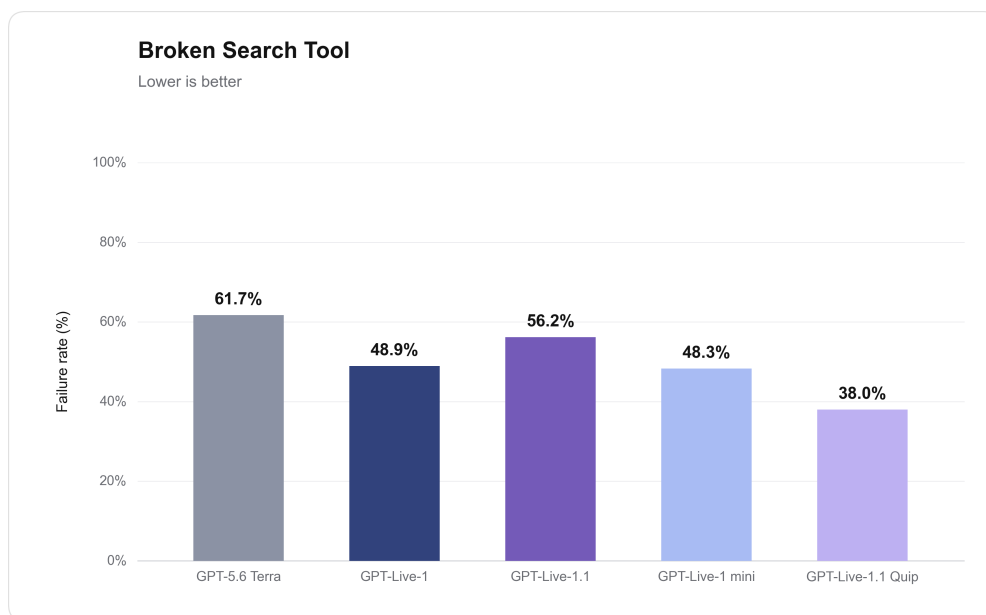


Figure 3: Broken Search Tool evaluation results.

We observe that GPT-Live-1.1 performs slightly worse and GPT-Live-1.1 Quip performs slightly better than their respective predecessors. Since the failure rate of the GPT-Live models was relatively high, we also compare their performance to the underlying model used for these evaluations, GPT-5.6 Terra. We see all GPT-Live models improve upon GPT-5.6 Terra performance, but more research is needed to understand the cause for that improvement.

5 Preparedness Framework

As GPT-Live-1.1 and GPT-Live-1.1 Quip are iterations on their predecessors, neither, when operating without delegation, could plausibly be considered High in any of our Preparedness Framework's Tracked Categories – Biological and Chemical Risk, Cybersecurity, or AI Self-Improvement.

- 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.

- For the AI Self-Improvement domain, we don't currently have models which have crossed the High capability threshold. If a model crosses this threshold in the future, the GPT-Live experience will similarly inherit the safeguard of the underlying model.