OpenAI

Seven tips for prompting voice agents with the Realtime API

#1. Be precise. Avoid conflicts.

The new Realtime model is very good at instruction following. However, that also means small wording changes or unclear instructions can shift behavior a lot. Inspect and iterate on your System Prompt. Iterate on wording and fix instruction contradictions.

Example of a real improvement:
Changing “inaudible” → “unintelligible” in a system prompt significantly improved how the model handled noisy inputs.

Try this:
After edits, have an LLM review your prompt for ambiguity or conflicts and propose tighter phrasing.

#2. Bullets over paragraphs.

Realtime models have shown to follow short bullet points better than long paragraphs.

Harder to follow:

```
When you can’t clearly hear the user, don’t proceed. If there’s background noise or you only caught part of the sentence, pause and ask them politely to repeat themselves in their preferred language, and make sure you keep the conversation in the same language as the user.
```

Easier to follow:

```
- Only respond to clear audio or text
- If audio is unclear/partial/noisy/silent, ask for clarification in {preferred_language}
- Continue in the same language as the user if intelligible
```

#3. Handle unclear audio.

The Realtime model is good at following instructions on how to handle unclear audio. Spell out what to do when audio isn’t usable.

```
## Unclear audio

- Always respond in the same language the user is speaking in, if intelligible.
- Only respond to clear audio or text.
- If the user's audio is not clear (e.g. ambiguous input/background noise/silent/unintelligible) or if you did not fully hear or understand the user, ask for clarification using {preferred_language} phrases.
```

#4. Pin language when needed.

By default, mirroring the user's language works well. If you see unwanted language switching, *add a dedicated "Language" section*. Make sure it doesn't conflict with other rules.

```
## Language
- The conversation will be only in English.
- Do not respond in any other language even if the user asks.
- If the user speaks another language, politely explain that support is limited to English.
```

#5. Sample phrases and flow snippets.

The model learns style from examples. Give short, varied samples for common conversation moments.

```
## Conversation flow — Greeting
Goal: Set tone and invite the reason for calling.

How to respond:
- Identify as ACME Internet Support.
- Keep it brief; invite the caller's goal.

Sample phrases (vary, don't always reuse):
- "Thanks for calling ACME Internet—how can I help today?"
- "You've reached ACME Support. What's going on with your
service?"
- "Hi there—tell me what you'd like help with."

Exit when: Caller states an initial goal or symptom.
```

#6. Avoid robotic repetition.

If responses sound repetitive or robotic, include an explicit variety instruction. This can sometimes happen when using sample phrases.

```
## Variety
- Do not repeat the same sentence twice. Vary your responses
so it doesn't sound robotic.
```

#7. Use capitalized text for emphasis.

Like many LLMs, using capitalization for important rules can help the model to understand and follow. It is also helpful to convert non-text rules (such as numerical conditions) into text before capitalization.

Instead of:

```
## Rules
- If [func.return_value] > 0, respond 1 to the user.
```

Use:

```
## Rules
- If [func.return_value] IS BIGGER THAN 0, RESPOND 1 TO THE
USER.
```

For more details, check out our prompting guide. Test your prompts in the Realtime playground.