

# Audio MultiChallenge: A Multi-Turn Evaluation of Spoken Dialogue Systems on Natural Human Interaction

Advait Gosai\*, Tyler Vuong\*, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, Yunzhong He

Scale AI

## Abstract

End-to-end (E2E) spoken dialogue systems are replacing cascaded pipelines for voice-based human-AI interaction. Existing benchmarks primarily evaluate these systems on synthetic speech and single-turn tasks, leaving multi-turn conversational ability underexplored. We introduce **Audio MultiChallenge**, an open-source benchmark to evaluate these systems under natural multi-turn interaction patterns. Building on the text-based MultiChallenge framework, which evaluates *Inference Memory*, *Instruction Retention*, and *Self Coherence*, we introduce a new axis *Voice Editing* that tests robustness to mid-utterance speech repairs and backtracking. We augment each axis to the audio modality, such as introducing Audio-Cue challenges for *Inference Memory* that require recalling ambient sounds and paralinguistic signals beyond semantic content. We curate 452 conversations from 47 speakers with 1,712 instance-specific rubrics through a hybrid pipeline that exposes model failures at scale while preserving natural disfluencies found in unscripted human speech. Our evaluation reveals that even frontier models struggle on our benchmark, with our highest-performing model achieving a 54.65% pass rate. Error analysis shows that models are not sufficiently robust to human speech when tracking instructions, edits, and audio cues, highlighting the need for improved audio-native multi-turn interaction capabilities.

## 1 Introduction

Spoken dialogue systems (Jokinen and McTear, 2009) are evolving from cascaded pipelines comprising Automatic Speech Recognition (ASR) (Radford et al., 2023), Large Language Models (LLMs) (Brown et al., 2020), and Text-to-Speech (TTS) systems (Wang et al., 2017) to end-to-end (E2E) architectures that map input speech directly to responses without intermediate transcription

(Nguyen et al., 2023; Arora et al., 2025b). Speech-to-speech (S2S) models and broader Audio Language models (Audio LMs) process raw or minimally processed acoustic features and generate spoken or textual outputs, often operating over shared latent representations or learned discrete speech units (Nguyen et al., 2023; Zhang et al., 2023; Borsos et al., 2023; Défossez et al., 2024, 2023). This integration can reduce response latency and mitigate error propagation, while better preserving paralinguistic and acoustic information (Cui et al., 2025; Ji et al., 2024).

As E2E systems move toward real-world user-facing interaction, their evaluation must reflect the demands of extended, spontaneous dialogue rather than isolated prompt-response pairs or short, synthetic TTS inputs typical of existing benchmarks such as Big Bench Audio (Hill-Smith et al., 2024; Suzgun et al., 2023). Evidence from text-only LLMs shows that strong single-turn performance does not ensure robust multi-turn behavior (Deshpande et al., 2025; Laban et al., 2026; Bertsch et al., 2025). These limitations are amplified in the audio modality. Audio is temporally continuous and often encoded as high-rate token streams, complicating long-horizon state tracking compared to text (Borsos et al., 2023; Cui et al., 2025). As a result, while current Audio LM benchmarks report strong performance on short or synthetic inputs, robustness in realistic multi-turn settings remains largely unexplored. Recent multi-turn evaluations are limited to brief dialogues (Du et al., 2025), leaving it unclear whether systems can sustain self-consistency, follow user constraints, exploit audio cues over long contexts, or recover from mid-utterance corrections and barge-ins.

To rigorously evaluate these dimensions, we introduce **Audio MultiChallenge**<sup>1</sup>, an extension of

\* Equal contribution.

<sup>1</sup>Dataset is publicly available at: <https://huggingface.co/datasets/ScaleAI/audiomc>

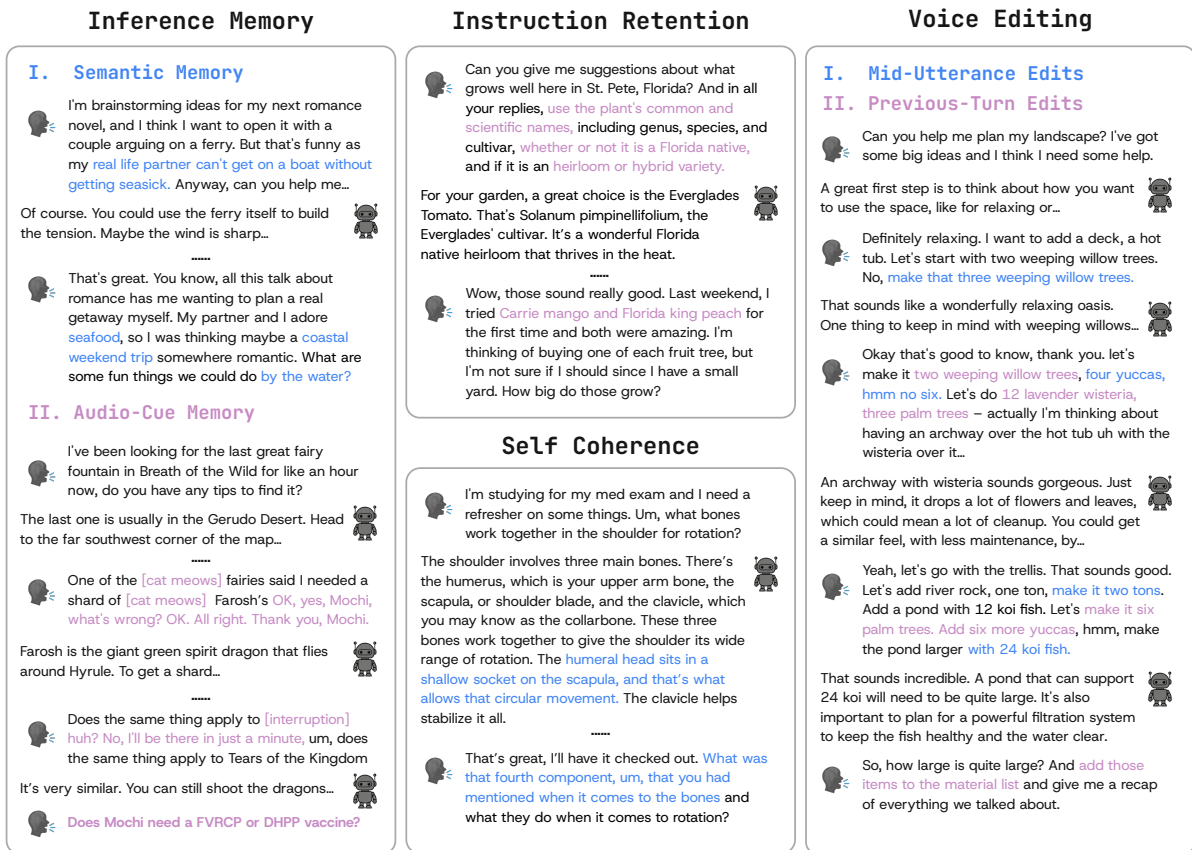


Figure 1: Samples from Audio MultiChallenge. Highlighted text shows key interaction patterns per axis.

MultiChallenge (Deshpande et al., 2025) to the audio modality with the following key contributions:

- We open-source **452** multi-turn conversations from **47** unique speakers, summing up to **15** hours of human audio with no post-processing to retain ambient acoustic variability.
- We introduce a novel *Voice Editing* axis that evaluates robustness to mid-utterance speech repairs and backtracking, capturing natural human behaviors such as self-corrections that are common in spontaneous dialogue. We further add *Audio-Cue Inference Memory* tasks that require recalling ambient sounds or paralinguistic signals rather than only semantic content, and adapt the remaining *Instruction Retention* and *Self Coherence* axes to reflect the complexities of voice interactions.
- Our data-generation pipeline combines multi-agent synthetic component with a human-in-the-loop process. Rather than providing human contributors with verbatim scripts, we provide high-level *blueprints* and interaction goals derived from a scalable agentic loop de-

signed to elicit diverse failure modes. These blueprints encourage improvisational interactions that surface realistic failures, making our dataset sufficiently adversarial while remaining entirely unscripted and preserving paralinguistics and natural human disfluencies.

- We replace coarse binary success metrics with **1,712** atomic, instance-specific *rubrics* across the dataset that decompose each ideal response into critical sub-requirements. Our LLM-as-a-judge method using these rubrics achieves a **0.937** Macro F1 with human graders, exceeding previous work.

Our analysis of Audio MultiChallenge reveals substantial gaps in current E2E capabilities across frontier models. Even the highest-performing model, Gemini 3 Pro Preview (Thinking), achieves only 54.65% pass rate, with other model families, especially open-source ones lagging considerably behind. We find that Self Coherence degrades consistently as cumulative user audio duration increases, dropping from 33.3% on shorter tasks to 20.0% on tasks with 3–5 minutes of audio. Models also struggle to recall paralinguistic and ambient

signals from previous turns, scoring 36.5% lower on Audio-Cue Inference Memory than on semantic memory tasks. Voice Editing proves to be the most challenging axis, where behaviors such as mid-utterance self-corrections and prior-turn edits frequently trigger failures that are largely absent in cleaner text baselines. Together, our results suggest the need for audio-native training and evaluation that target realistic multi-turn spoken interaction.

## 2 Audio MultiChallenge

Audio MultiChallenge evaluates the multi-turn capabilities of E2E spoken dialogue systems along four axes: **Inference Memory**, **Instruction Retention**, **Self Coherence**, and **Voice Editing**. We explicitly distinguish from text-based MultiChallenge to address the unique requirements of spoken interaction. For instance, directly applying text benchmarks via TTS is insufficient for spoken interaction, as many text-centric tasks (e.g., code writing or text formatting) do not naturally translate to voice and synthetic inputs fail to capture human paralinguistics. Audio MultiChallenge instead uses unscripted human-recorded conversations that reflect the variability of spontaneous speech, including disfluencies, backtracking and interruptions. Accordingly, all four axes are adapted to evaluate both semantic and acoustic reasoning under realistic audio conditions.

### 2.1 Axes

**Inference Memory** Inference Memory evaluates a model’s capacity to recall and synthesize user-specified details from previous turns that are implicitly required to satisfy their final request. Unlike direct queries, these tasks demand that the model proactively reallocate attention to prior context to inform its reasoning. For example, a standard Semantic test might require remembering a stated nut allergy when generating a dessert recipe. We extend this definition to include *Audio-Cue Gated Inference Memory* where the critical context to be retained is not explicitly spoken, but is embedded in the acoustic environment or paralinguistic signals. An Audio-Cue gated test might feature the background sound of heavy rain in an initial turn, implicitly constraining a later request for an outfit recommendation. This evaluates the model’s ability to retain and integrate both semantic and acoustic information over a multi-turn conversation.

<b>Conversations</b>	452
Inference Memory	132
Instruction Retention	120
Self Coherence	83
Voice Editing	117
<b>Turn Count</b>	3-8
<b>Rubrics</b>	1,712
<b>Unique Speakers</b>	47
<b>Audio Duration</b>	14.99 h
<b>Sampling Rate</b>	48 kHz
<b>Language</b>	English

Table 1: Dataset Statistics

**Instruction Retention** This axis evaluates whether an Audio LM can maintain and correctly apply precise, evolving instructions over dialogue histories, a failure mode that is frequently observed in speech interactions (Lu et al., 2025a,b). In Audio Multichallenge, we broaden the instruction space beyond simple stylistic or structural constraints, to *stacked* and *conditional* rule sets. These include multi-constraint directives and trigger-based behaviors, for example changing the response format only after a specific phrase is spoken. Conversation topics here include realistic voice scenarios such as debate and roleplay.

**Self Coherence** Self Coherence evaluates the model’s ability to maintain internal consistency regarding factual assertions, opinions, and established personas throughout a dialogue. A common failure mode is a model contradicting its prior correct responses to align with a user who questions them. For Audio MultiChallenge, we refine this category by excluding contrived, adversarial prompts in favor of realistic conversation scenarios. While adaptive reasoning is desirable in light of new evidence, we target unwarranted contradictions where the model invalidates its own previous statements without justification.

**Voice Editing** Voice Editing is a novel axis introduced specifically to mimic speech interactions with a personal voice assistant. We define this axis as the model’s ability to recognize and apply immediate, spontaneous speech repairs, such as mid-utterance self-corrections or implemented planned edits that span multiple turns. This axis targets a unique challenge for audio-native models. While

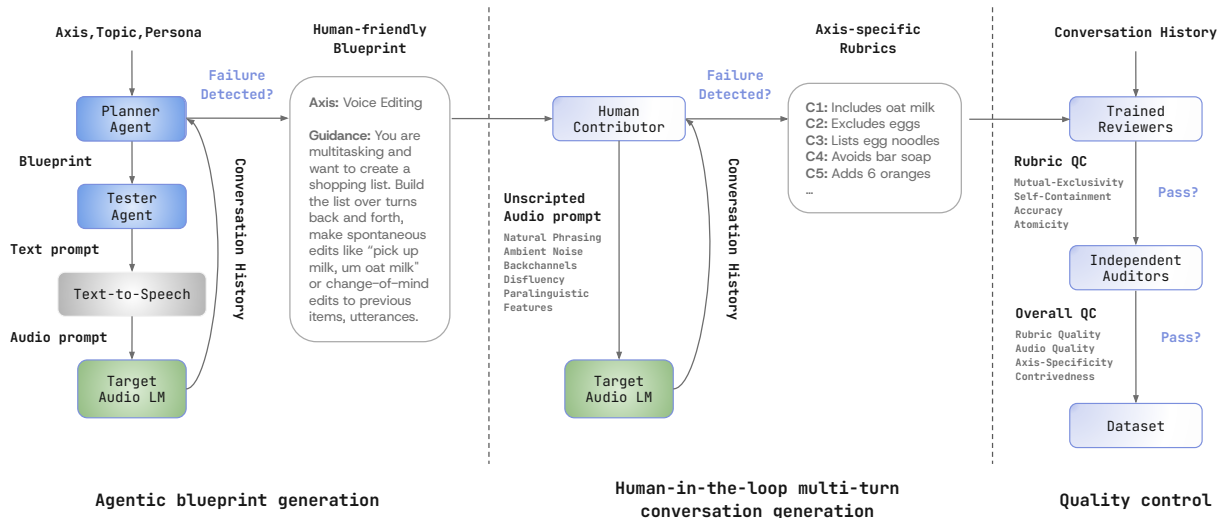


Figure 2: Audio MultiChallenge data curation utilizes agentic blueprint generation followed by human multi-turn conversations and quality control. Topic seeds, blueprints and data samples are provided in the Appendix.

LLMs are trained on refined prompts where errors are backspaced or redrafted prior to submission, speech is a continuous stream where the editing process is audible. Consequently, models must dynamically filter out retracted content (e.g., "Let's make it four yuccas, hmm no six.") and resolve implicit contradictions that arise within consecutive turns. Evaluation here focuses on the model's robustness to these natural disfluencies and its capacity to discern the user's final intent amidst the messiness of spoken interaction.

## 2.2 Dataset Curation

We curate Audio MultiChallenge using a two-stage hybrid approach that produces *naturalistic*, *diverse*, and *challenging* tasks. Specifically, we combine (i) an automated multi-agent pipeline that explores Audio LM failure modes at scale with (ii) a human-in-the-loop stage that translates these failures into realistic conversations and evaluation rubrics.

**Agentic Audio LM Failure Generation** We implement an automated multi-agent framework that adapts the synthetic generation methodology of [Deshpande et al. \(2025\)](#) to audio, producing conversations designed to elicit axis-specific failures. As illustrated in Figure 2, the process begins by sampling a target evaluation axis  $A$ , a seed topic  $T$ , and a persona  $P$  from curated taxonomies. A Planner Agent proposes an initial conversation strategy aimed at triggering a failure under the selected axis, which a Tester Agent instantiates as user prompts.

These prompts are rendered as speech via TTS and provided to a Target Audio LM. The Planner Agent evaluates the model's responses and iteratively increases task complexity until a failure is observed. Once a failure occurs, the Planner analyzes the dialogue history and distills the underlying failure mechanism into a concise, human-readable blueprint  $B_{\text{final}}$ , which serves as strategic guidance for human contributors.

We use `gpt-4o-mini-tts` for TTS, randomly sample GPT 4o Audio Preview and Gemini 2.5 Pro as Target Agents, and use OpenAI o3 as the Planner. Sample blueprints are provided in Appendix A.8.

### Human-In-The-Loop Conversation Generation

Using blueprints as seeds, human contributors record unscripted spoken interactions with an Audio LM (either GPT 4o Audio Preview and Gemini 2.5 Pro). Contributors are given the blueprint and a detailed definition of the assigned axis but are encouraged to freely reinterpret, extend, or pivot the scenario to expose model weaknesses. This process yields spontaneous speech containing natural disfluencies, self-corrections, and topic shifts.

For Audio-Cue Inference Memory, we modify blueprints to make audio-cue retention the primary task, covering environmental sounds, recording locations and paralinguistic features such as tone, emotion and emphasis. Annotators are instructed to select blueprints that match their available props or location for realistic capture.

While recording, contributors interact with the

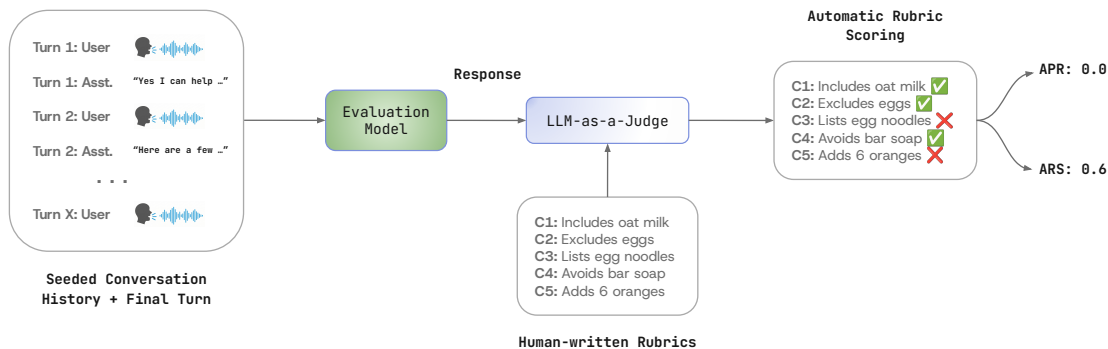


Figure 3: Fixed-Context Evaluation breakdown. Each model receives identical conversation history and is evaluated on its response to the final user turn.

model in a turn-based manner, recording an audio prompt at each turn and reviewing the model’s response. After each exchange, they verify that the dialogue remains coherent, challenging, and aligned with the target axis, avoiding premature failures unrelated to the intended evaluation such as model hallucinations. The conversation terminates once an objective, axis-specific failure is observed.

**Rubrics for Model Evaluation** For each collected conversation, contributors define a set of atomic, binary evaluation criteria corresponding exclusively to the final user turn and the assigned axis. Rather than providing a single reference response, contributors decompose the ideal behavior into multiple discrete rubric items, each assessing a specific requirement. Criteria that reflect only general conversational quality or abstract axis definitions are explicitly excluded. This structured rubric design enables precise LLM-based judgment by allowing each requirement to be evaluated independently.

### 3 Evaluation

To ensure a standardized evaluation, we adopt a fixed-context protocol, as depicted in Figure 3. For each test example, we pre-seed the full conversation history (both user and assistant turns) and evaluate only the model’s response to the final turn using our LLM-as-a-judge approach. User turns are provided in context as base64-encoded audio files, while assistant turns are provided as raw text.

This contrasts with sequential evaluation methods used in prior work, where a model generates responses turn-by-turn to pre-recorded user prompts (Du et al., 2025). We avoid the sequential approach because it can suffer from conversation drift. If a model’s response in the first turn diverges even

minorly from the reference transcript, subsequent pre-recorded user follow-ups and their associated rubrics may become semantically invalid. Our approach ensures that every model receives identical context, enabling a fair, reproducible comparison.

**Evaluation Metrics** We utilize the Average Pass Rate (APR) metric (Guo et al., 2025) to evaluate model performance. A model’s response is considered successful if and only if it satisfies all rubrics; otherwise, it is marked as a failure. Instead of restricting our data to one rubric criteria as seen in Deshpande et al. (2025), we allow for multiple binary weighted atomic rubrics for increased LLM judge reliability. Additionally, we report the Average Rubric Score (ARS) (Guo et al., 2025) for each model which represents the average percentage of satisfied rubrics per task. We formally define our metrics in Appendix A.4.

### 3.1 Results

We present scores on Audio MultiChallenge across a variety of E2E architectures in Table 2, along with their performance per-axis. We observe that Gemini 3 Pro Preview (Thinking) achieves the highest overall score of 54.65%, followed by its counterparts Gemini 2.5 Pro (Thinking) and 2.5 Flash (Thinking). In comparison, the remainder of proprietary and open-source models perform modestly, with GPT 4o Audio Preview and Voxtral Small 24B leading, scoring a 25.44% and 26.33% APR respectively, closely followed by GPT Realtime with an APR of 23.45% (output modality set to Text)<sup>2</sup>. Open-source S2S architectures in particular

<sup>2</sup>GPT Realtime and Mimo Audio models previously report scores on a TTS set of text-based MultiChallenge (referring to it as MultiChallenge Audio) which we distinguish from.

Model	Output Modality	Overall		Per-Axis APR (%)			
		APR (%)	ARS (%)	IM	IR	SC	VE
Gemini 3 Pro Preview (Thinking)	Text	54.65	82.54	57.58	65.283	40.96	49.57
Gemini 2.5 Pro (Thinking) (Comanici et al., 2025)	Text	46.90	78.91	44.70	56.67	40.96	43.59
Gemini 2.5 Flash (Thinking)	Text	40.04	74.59	30.30	37.50	54.22	43.59
Gemini 2.5 Flash	Text	26.11	65.42	19.70	29.17	31.33	26.50
GPT Realtime	Text	23.45	65.95	21.97	26.67	22.89	22.22
GPT Realtime	Audio	20.35	63.03	19.70	20.00	26.51	17.09
GPT 4o Audio Preview (OpenAI et al., 2024)	Text	25.44	67.58	23.48	27.50	31.33	21.37
GPT 4o Audio Preview	Audio	23.23	64.25	21.21	30.83	24.10	17.09
GPT 4o Mini Audio Preview	Text	14.82	57.01	15.15	16.67	20.48	8.55
GPT 4o Mini Audio Preview	Audio	13.05	54.91	9.85	12.50	21.69	11.11
Qwen 3 Omni 30B A3B Instruct (Xu et al., 2025b)	Audio	24.34	63.43	20.45	27.50	30.12	21.37
Qwen 2.5 Omni 7B (Xu et al., 2025a)	Text	11.95	45.39	12.88	13.33	18.07	5.13
Voxtral Small 24B (Liu et al., 2025)	Text	26.33	66.59	20.45	25.00	27.71	33.33
Mimo Audio 7B Instruct (Thinking) (Xiaomi, 2025)	Text	19.69	55.78	19.70	22.50	27.71	11.11
Mimo Audio 7B Instruct	Text	18.58	53.86	17.42	20.00	27.71	11.97
Phi 4 MM Instruct (Microsoft et al., 2025)	Text	15.49	46.96	16.67	16.67	28.92	3.42
Gemma 3n E4B IT (Google, 2025)	Text	15.49	43.11	14.39	22.50	26.51	1.71
Kimi Audio 7B Instruct (KimiTeam et al., 2025)	Text	13.72	48.36	14.39	16.67	20.48	5.13
Kimi Audio 7B Instruct	Audio	10.40	22.08	15.91	6.67	13.25	5.98
LFM2 Audio 1.5B (Amini et al., 2025)	Audio	9.29	19.10	15.15	7.50	15.66	0.00

Table 2: Overall and per-Axis scores on Audio MultiChallenge per model and output modality. APR serves as final scores. IM: Inference Memory, IR: Instruction Retention, SC: Self Coherence, VE: Voice Editing. Green highlights indicate best performance with output modality as Text; blue highlights indicate best Audio output performance.

struggle the most, with no model surpassing Qwen 3 Omni which achieves a modest 24.34% APR. Analyzing per-axis scores, our newly adapted Voice Editing and Inference Memory prove to be the most challenging tasks for frontier models, with them scoring 17.99% and 21.55% on average, respectively. We highlight the following key findings and provide case studies of sample model failures in Appendix A.10.

### Audio-Cue Inference Memory lags substantially

We classify Audio-Cue Inference Memory tasks as ones where the information the model must remember from a previous turn is contained only in the audio signal rather than in the words spoken, as described in Section 2.1. Our examples span both *paralinguistic* cues such as the speaker’s tone and *ambient* cues such as background sound. Figure 4 shows a significant gap in remembering these audio cues in our multi-turn setup with ARS falling by 36.5% relatively compared to semantic memory on average across models. This unique subset of our data highlights the limitations of current E2E spoken dialogue systems when it comes to audio-native state tracking and recall, which is fundamental to realtime and multi-turn voice conversations.

### Conditional Instructions cause model failures

Frontier models score 25.08% APR and 46.11% ARS on average on Instruction Retention (IR) tasks. To analyze the main sources of error, we automatically cluster each IR rubric item into two instruction types: *Fixed*, which are unconditional, persistent constraints meant to hold throughout the conversation (e.g., response style or roleplay prompts), and *Conditional*, which are context-dependent instructions that only apply in specific future-turn scenarios. Figure 9 shows that each model, regardless of their overall performance, demonstrates worse performance on conditional IR. This motivates treating conditional instruction following as a distinct, harder sub-problem to optimize for.

### Self Coherence declines on long-context tasks

Self Coherence degrades linearly as cumulative user audio duration increases. Figure 5b shows that APR on Self Coherence drops from 33.3% for tasks with 0–60 seconds of cumulative user audio to 20.0% for tasks with 3–5 minutes, and further to 13.3% for tasks exceeding 5 minutes (though this final bin contains only 3 samples) when averaged across all models. We analyze duration effects on other axes in Section 3.2.

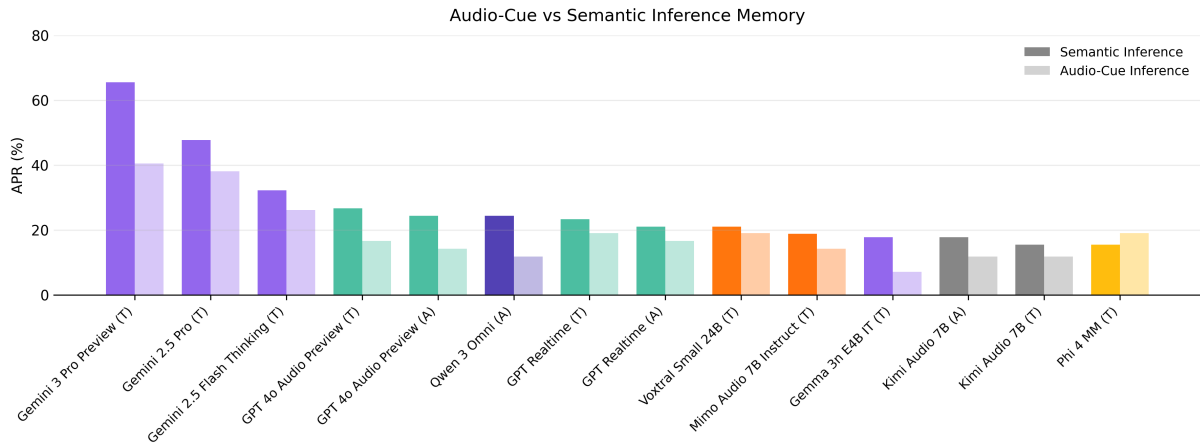


Figure 4: APR (%) on Audio-Cue and Semantic IM tasks. (T) indicates text output, (A) indicates audio output.

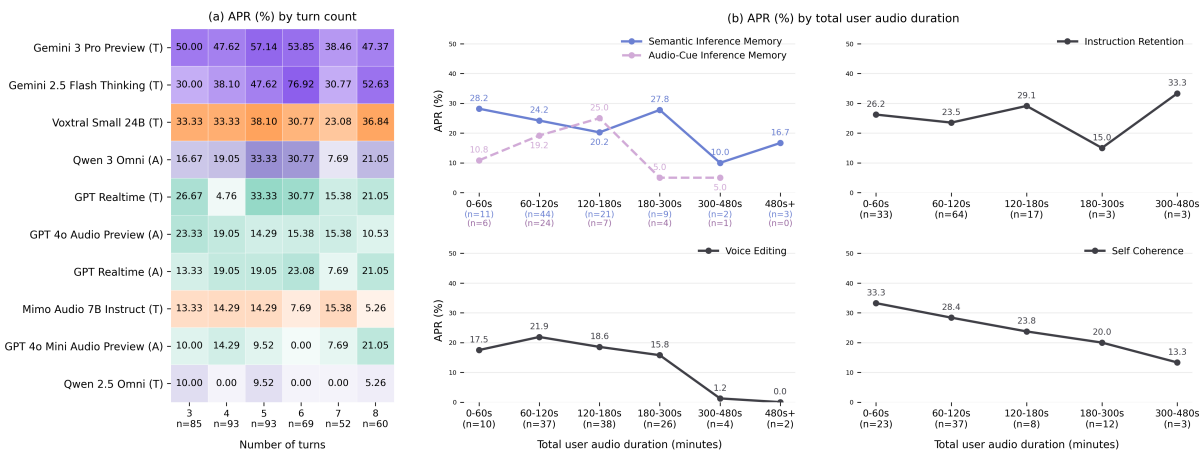


Figure 5: APR (%) across (a) number of turns, (b) binned total user audio duration (in seconds).

**Voice Editing is most challenging** Voice Editing remains the most challenging axis, with an APR of 17.99%. Most failures on Voice Editing occur when users introduce multiple *mid-utterance* and *prior-turn* edits throughout a conversation, with the final prompt requesting a summary or revision that integrates all specified changes (e.g. listing a final changed itinerary). Successfully completing these tasks requires state tracking and multi-constraint integration at the semantic level, as well as robustness to mid-utterance backchannels and disfluencies at the acoustic level, which weaker models struggle with, as shown in Figure 10. In contrast, Gemini models track both edited and unedited items better, and with nearly equal accuracy.

### 3.2 Analysis

**Modality Gap** We observe a "modality gap" when evaluating models that support both text and audio output streams. Under a Text only output configuration, these models score 19.36% on average,

however, when the same models are configured to output Audio (or joint Audio and Text depending on their architecture), their performance degrades to 16.76%. These findings are independently reported in other recent works (Xiaomi, 2025), highlighting the need for improved S2S post-training.

**Turn Count** Figure 5a shows no meaningful relationship between turn count and model performance. This mirrors findings from prior work using fixed-context evaluation (Deshpande et al., 2025). As each turn length is arbitrarily defined, it is a weak measure of context length and task difficulty as compared to cumulative user audio duration.

**Audio Duration** Since models use different audio tokenizers, raw token counts are not directly comparable across systems. We therefore use cumulative user-audio duration as a proxy for context length. In addition to Self Coherence as discussed above, Figure 5b shows Voice Editing also worsens with longer histories. Average APR falls from

17.5% in the 3–5 minute bin to 1.2% and 0% for 5–8 minute and 8+ minute tasks respectively (though these bins have low sample sizes). Audio-Cue Inference Memory shows no clear trend with duration; qualitatively, we observe performance is often bottlenecked first by audio perception (e.g., reliably identifying paralinguistic cues or background sounds) rather than retention, leading to failures even on short clips. In contrast, Instruction Retention remains steady or improves on longer tasks, likely because constraints are introduced early and reinforced by previous correct assistant turns that serve as in-context demonstrations. This is consistent with known in-context learning effects observed on instruction following in LLMs (Dong et al., 2024).

### 3.3 LLM-as-a-Judge Setup

To evaluate our LLM-as-a-judge setup, we measure inter-rater agreement between model-generated and human-assigned labels across the full dataset. Human graders are instructed to grade randomly sampled GPT 4o Audio Preview or Gemini 2.5 Pro responses on corresponding rubrics. We aggregate these rubric grades into a set of 1,712 observations and compute Cohen’s  $\kappa$  and Macro F1 scores calculated following Arora et al. (2025a). To mitigate the self-preference bias known to affect LLM judges (Zheng et al., 2023; Wataoka et al., 2025), we exclude our evaluated models from judging. Our results, presented in Table 3, show that all judges achieve extremely high agreement with human graders, indicating the effectiveness of our rubric creation and overall LLM-as-a-Judge setup. Among the tested judges, o4 Mini attains the highest performance with a Macro F1 score of 0.937 which, to the best of our knowledge exceeds all prior work (Arora et al., 2025a; Du et al., 2025; Sharma et al., 2026; Wang et al., 2026). o4 Mini is therefore selected as our primary LLM judge. We provide our complete setup in Appendix A.3.

LLM Judge	Cohen’s $\kappa$	Macro F1
o4 Mini	0.873	0.937
GPT 5 Mini	0.870	0.935
GPT 4.1	0.811	0.906
Claude Haiku 4.5	0.808	0.904
Mistral Medium	0.787	0.894
DeepSeek V3.1	0.765	0.882

Table 3: LLM Judge agreement with human graders.

## 4 Ablation Studies

### 4.1 TTS vs. Human User Audio

To isolate the impact of human speech, we evaluate our models on synthetic versions of our user speech, generated using gpt-4o-mini-tts. We exclude Audio-Cue IM and tasks that contain interruptions or sidebar conversations, resulting in a filtered subset of 385 samples. Results in Table 4 depict models configured for text outputs achieve a 7.5% relative performance improvement when using TTS speech. In contrast, the same architectures configured for audio outputs exhibit a 2.5% relative decline, positively suggesting they may be optimized for real human dialogue. Table 5 shows that Voice Editing benefits the most from TTS under text output configurations. We attribute this to ASR and TTS failing to capture human-like hesitations and unintelligible segments in spontaneous speech, producing more structured and easier-to-follow inputs. This highlights the importance of evaluating on real human audio. However, the performance degradation observed for audio output configurations on Voice Editing and Instruction Retention when using TTS requires further investigation.

Model	Output Modality	Human APR (%)	TTS APR (%)	$\Delta$ Relative (%)
Gemini 3 Pro Preview	Text	55.58	56.62	+1.9
Gemini 2.5 Pro	Text	47.53	49.09	+3.3
Gemini 2.5 Flash (Thinking)	Text	41.56	42.86	+3.1
Gemini 2.5 Flash	Text	25.97	27.79	+7.0
GPT 4o Audio Preview	Text	27.27	28.05	+2.9
GPT 4o Audio Preview	Audio	24.42	23.64	-3.2
GPT 4o Mini Audio Preview	Text	15.84	17.92	+13.1
GPT 4o Mini Audio Preview	Audio	14.03	13.77	-1.9
GPT Realtime	Text	24.16	29.35	+21.5
GPT Realtime	Audio	20.52	20.00	-2.5

Table 4: APR (%) on Human vs. TTS audio.

Axis	Output Modality	Human APR (%)	TTS APR (%)	$\Delta$ Relative (%)
Inference Memory	Text	34.05	34.88	+2.4
Inference Memory	Audio	19.38	20.54	+6.0
Instruction Retention	Text	37.12	38.95	+4.9
Instruction Retention	Audio	20.80	18.52	-11.0
Self Coherence	Text	34.18	35.62	+4.2
Self Coherence	Audio	24.47	26.58	+8.6
Voice Editing	Text	30.24	33.70	+11.5
Voice Editing	Audio	14.89	12.94	-13.0

Table 5: APR (%) on Human vs. TTS audio per axis.

### 4.2 E2E vs. Cascaded Systems

We benchmark frontier LLMs in a cascaded pipeline (ASR + LLM) using whisper-large-v3.

In Table 6, we observe that cascaded systems using GPT 5 and Claude Opus 4.5 perform strongly relative to most E2E baselines except Gemini models. Interestingly, cascading reduces performance for Gemini 3 Pro Preview and Gemini 2.5 Pro, with relative drops of 10.3% and 4.9%. A hypothesis is that newer Gemini models are post-trained more extensively on audio-native data to function as E2E multimodal models than solely as LLMs. These results emphasize the need for continued E2E improvement to close the gap with cascaded pipelines.

Model	E2E APR (%)	Cascaded APR (%)	$\Delta$ Relative (%)
Gemini 3 Pro Preview	55.58	49.87	-10.3
Gemini 2.5 Pro	47.53	45.19	-4.9
Gemini 2.5 Flash (Thinking)	41.56	45.71	+10.0
GPT 5	-	51.17	-
Claude Opus 4.5	-	39.22	-
GPT 4o	-	28.57	-

Table 6: APR (%) for E2E vs. Cascaded pipelines.

## 5 Related Work

**Single-Turn Audio Evaluation** Existing single-turn benchmarks primarily target isolated utterances or short clips. Audio reasoning benchmarks cover a range of audio understanding, music and paralinguistic tasks (Kumar et al., 2026; Ma et al., 2025; Chen et al., 2024; Wang et al., 2025). Other recent works rely on synthetic or carefully read speech and focus on low-level capabilities rather than interactive dialogue (Cuervo et al., 2026; Yosha et al., 2025; Hill-Smith et al., 2024). Long-form audio benchmarks construct hour-scale recordings and show that model performance degrades as audio inputs become longer and sparser (Ahia et al., 2025; He et al., 2025; Ghosh et al., 2025). However, they treat each audio as a single comprehension task, not requiring models to track user preferences, edits, or self-consistency.

**Multi-Turn LLM Evaluation** MultiChallenge (Deshpande et al., 2025) builds synthetic conversations targeting instruction retention, semantic memory, text editing, and self-coherence, and finds that frontier models struggle despite near-saturation on earlier multi-turn benchmarks. Further studies report that LLM performance drops significantly when moving from single to multi-turn settings on long-context generation, tracking and reasoning (Laban et al., 2026; Bertsch et al., 2025).

**Multi-Turn Audio Evaluation** In audio-native settings, multi-turn evaluation remains comparatively underexplored. MTalk-Bench (Du et al., 2025) introduces tasks spanning semantic content, paralinguistic cues, and ambient sound, combining arena-style ranking with rubrics, but is limited to 2–3 turns and uses scripted dialogue. Our work targets this gap by focusing on extended multi-turn S2S or S2T interaction, long-horizon instruction retention, inference memory, and robustness to non-monotonic user behavior. Furthermore, to the best of our knowledge, there exists no speech-native benchmark that systematically tests Voice Editing.

## 6 Conclusion

We introduce Audio MultiChallenge, a collection of 452 challenging multi-turn audio conversations for benchmarking spoken dialogue systems. The benchmark evaluates four core axes central to voice assistants: Inference Memory, Instruction Retention, Self Coherence, and Voice Editing, enabled by a two-stage pipeline combining agentic blueprinting with human-recorded speech. Across 1,712 instance-specific rubrics, our results reveal substantial gaps in model performance on acoustic memory, conditional instruction following, long-context self-consistency, and robustness to mid-utterance speech repairs. We open-source our dataset and evaluation setup to support reproducibility and drive progress toward more capable E2E architectures in real-world multi-turn settings.

## Limitations

A key limitation of this work is that our evaluation does not operate directly on raw audio outputs. As a result, we assess only the semantic content of model outputs (using the output text) rather than their audio generation quality. Future work could address this by developing reliable audio-enabled rubrics and judges capable of evaluating instruction following and other axes across vocal style, prosody, and tone.

Rubric-based benchmarks can also be sensitive to response length, as more verbose outputs may satisfy a greater number of criteria even when the responses are not human-desirable (Sharma et al., 2026). To mitigate this, we rely on exclusion rubrics (e.g., “Avoids mentioning leather products”) that explicitly test whether models correctly omit disallowed or undesired information. Nevertheless, we still observe a strong correlation between aver-

age response length and rubric compliance (ARS).

Furthermore, our ablation study in Section 4.1 applies TTS to unscripted human dialogue, and model behavior may differ under fully scripted or synthetic inputs. We also note that the data collection pipeline is particularly adversarial to our User Agents (GPT-4o Audio Preview and Gemini 2.5 Pro), though the observed failure modes generalize across models.

Finally, our dataset includes only English conversations, limiting the benchmark’s ability to assess multilingual multi-turn spoken dialogue performance which is as important.

## Ethical Considerations

Our dataset consists of real human speech and audio collected through voluntary participation. All contributors provided informed consent for the recording and release of their voice in accordance with applicable local laws and regulations. We took care to avoid collecting any personally identifiable information or private, sensitive data.

Annotators and quality-control contractors operated independently and were compensated at rates consistent with fair labor practices and relevant local standards. AI assistance was only used to improve the clarity and presentation of the manuscript. All technical content, analyses, and conclusions come solely from the authors.

## Acknowledgments

We thank Kaustubh Deshpande for his insights on our methodology, and Ashan Panduwawala for his engineering support in setting up our human-in-the-loop data collection. We also thank Rani Alsaberi and Chanyu Li for their contributions towards our data curation and quality control.

## References

Orevaoghene Ahia, Martijn Bartelds, Kabir Ahuja, Hila Gonen, Valentin Hofmann, Siddhant Arora, Shuyue Stella Li, Vishal Puttagunta, Mofetoluwa Adeyemi, Charishma Buchireddy, Ben Walls, Noah Bennett, Shinji Watanabe, Noah A. Smith, Yulia Tsvetkov, and Sachin Kumar. 2025. [Blab: Brutally long audio bench](#). *Preprint*, arXiv:2505.03054.

Alexander Amini, Anna Banaszak, Harold Benoit, Arthur Böök, Tarek Dakhran, Song Duong, Alfred Eng, Fernando Fernandes, Marc Härkönen, Anne Harrington, Ramin Hasani, Saniya Karwa, Yuri Khrustaley, Maxime Labonne, Mathias Lechner, Valentine Lechner, Simon Lee, Zetian Li, Noel

Loo, and 14 others. 2025. [Lfm2 technical report](#). *Preprint*, arXiv:2511.23404.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025a. [Healthbench: Evaluating large language models towards improved human health](#). *Preprint*, arXiv:2505.08775.

Siddhant Arora, Jinchuan Tian, Hayato Futami, Jee weon Jung, Jiatong Shi, Yosuke Kashiwagi, Emiru Tsunoo, and Shinji Watanabe. 2025b. [Chain-of-Thought Training for Open E2E Spoken Dialogue Systems](#). In *Interspeech 2025*, pages 4833–4837.

Amanda Bertsch, Adithya Pratapa, Teruko Mitamura, Graham Neubig, and Matthew R. Gormley. 2025. [Oolong: Evaluating long context reasoning and aggregation capabilities](#). *Preprint*, arXiv:2511.02817.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. [AudioLM: a language modeling approach to audio generation](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024. [Voicebench: Benchmarking llm-based voice assistants](#). *Preprint*, arXiv:2410.17196.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.

Santiago Cuervo, Skyler Seto, Maureen de Seyssel, Richard He Bai, Zijin Gu, Tatiana Likhomanenko, Navdeep Jaitly, and Zakaria Aldeneh. 2026. [Closing the gap between text and speech understanding in llms](#). In *International Conference on Learning Representations*.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and

- Irwin King. 2025. [Recent advances in speech language models: A survey](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970, Vienna, Austria. Association for Computational Linguistics.
- Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. [High fidelity neural audio compression](#). *Transactions on Machine Learning Research*.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. [Moshi: a speech-text foundation model for real-time dialogue](#). *Preprint*, arXiv:2410.00037.
- Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. 2025. [Multi-Challenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18632–18702, Vienna, Austria. Association for Computational Linguistics.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A survey on in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
- Yuhao Du, Qianwei Huang, Guo Zhu, Zhanchen Dai, Shunian Chen, Qiming Zhu, Le Pan, Minghao Chen, Yuhao Zhang, Li Zhou, Benyou Wang, and Haizhou Li. 2025. [Mtalk-bench: Evaluating speech-to-speech models in multi-turn dialogues via arena-style and rubrics protocols](#). *Preprint*, arXiv:2508.18240.
- Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. [Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 19358–19405. PMLR.
- Google. 2025. Gemma 3n model overview. <https://ai.google.dev/gemma/docs/gemma-3n>. Accessed: 2025-11-11.
- Xingang Guo, Utkarsh Tyagi, Advait Gosai, Paula Vergara, Jayeon Park, Ernesto Gabriel Hernández Montoya, Chen Bo Calvin Zhang, Bin Hu, Yunzhong He, Bing Liu, and Rakshith Sharma Srinivasa. 2025. [Beyond seeing: Evaluating multimodal llms on tool-enabled image perception, transformation, and reasoning](#). *Preprint*, arXiv:2510.12712.
- Peize He, Zichen Wen, Yubo Wang, Yuxuan Wang, Xiaoqian Liu, Jiajie Huang, Zehui Lei, Zhuangcheng Gu, Xiangqi Jin, Jiabing Yang, Kai Li, Zhifei Liu, Weijia Li, Cunxiang Wang, Conghui He, and Linfeng Zhang. 2025. [Audiomathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio llms](#). *Preprint*, arXiv:2510.07293.
- Micah Hill-Smith, George Cameron, and Will Bosler. 2024. Artificial analysis big bench audio. [https://huggingface.co/datasets/ArtificialAnalysis/big\\_bench\\_audio](https://huggingface.co/datasets/ArtificialAnalysis/big_bench_audio). Accessed: 2025-11-11.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. [HuBERT: Self-supervised speech representation learning by masked prediction of hidden units](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.
- Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024. [Wavchat: A survey of spoken dialogue models](#). *Preprint*, arXiv:2411.13577.
- Kristiina Jokinen and Michael McTear. 2009. *Spoken Dialogue Systems*, volume 2 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.
- KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. [Kimi-audio technical report](#). *Preprint*, arXiv:2504.18425.
- Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, and 15 others. 2026. [Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(27):22688–22697.
- Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. [LLMs get lost in multi-turn conversation](#). In *The Fourteenth International Conference on Learning Representations*.
- Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. [Voxtral](#). *Preprint*, arXiv:2507.13264.

- Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-Yi Lee. 2025a. [Developing instruction-following speech language model without speech instruction-tuning data](#). In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Ke-Han Lu, Chun-Yi Kuan, and Hung yi Lee. 2025b. [Speech-IFEval: Evaluating Instruction-Following and Quantifying Catastrophic Forgetting in Speech-Aware Language Models](#). In *Interspeech 2025*, pages 2078–2082.
- Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025. [MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Leland McInnes, John Healy, and James Melville. 2020. [Umap: Uniform manifold approximation and projection for dimension reduction](#). *Preprint*, arXiv:1802.03426.
- Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, and 56 others. 2025. [Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras](#). *Preprint*, arXiv:2503.01743.
- Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023. [Generative spoken dialogue language modeling](#). *Transactions of the Association for Computational Linguistics*, 11:250–266.
- OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. 2026. [Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents](#). In *The Fourteenth International Conference on Learning Representations*.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. [Challenging BIG-bench tasks and whether chain-of-thought can solve them](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.
- Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025. [AudioBench: A universal benchmark for audio large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyriannakis, Rob Clark, and Rif A. Saurous. 2017. [Tacotron: Towards End-to-End Speech Synthesis](#). In *Interspeech 2017*, pages 4006–4010.
- Zhilin Wang, Jaehun Jung, Ximing Lu, Shizhe Diao, Ellie Evans, Jiaqi Zeng, Pavlo Molchanov, Yejin Choi, Jan Kautz, and Yi Dong. 2026. [Profbench: Multi-domain rubrics requiring professional knowledge to answer and judge](#). In *The Fourteenth International Conference on Learning Representations*.
- Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. [Self-preference bias in llm-as-a-judge](#). *Preprint*, arXiv:2410.21819.
- LLM-Core-Team Xiaomi. 2025. [Mimo-audio: Audio language models are few-shot learners](#).
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. [Qwen2.5-omni technical report](#). *Preprint*, arXiv:2503.20215.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. [Qwen3-omni technical report](#). *Preprint*, arXiv:2509.17765.
- Iddo Yosha, Gallil Maimon, and Yossi Adi. 2025. [Stresstest: Can your speech lm handle the stress?](#) *Preprint*, arXiv:2505.22765.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. [SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.



```

System Prompt

Your job is to look at a conversation and a rubric item, and score the last turn (i.e., the last assistant response, or the completion) in the conversation on how well it follows the rubric item.

# Conversation History
«conversation_history»

# Rubric item
«rubric_item»

# Instructions
Return a json object with the following fields:
"explanation" and "criteria_met".

• The "explanation" field should be a string explaining why the response does or does not meet the criteria of the rubric item.

• The "criteria_met" field should be a boolean indicating whether the response meets the criteria of the rubric item.

# Example 1
For example, if the response is "For tonight's dinner, think of making a cheese and mushroom pizza" and the rubric item is "Avoids dairy product in the recipes", you should return a json like this:

```json
{
  "explanation": "The response includes a dairy product in the recipe, so the criteria was not met.",
  "criteria_met": false
}
```

# Final instruction
Return just the json object in markdown format. Do not include any other text in the response.

```

Figure 8: LLM-as-a-Judge system prompt.

### A.5 Evaluation Model Parameters

Table 7 provides the detailed inference parameters we utilize for each model evaluated. We match the recommended sampling parameters by referring to each model’s provided inference code or

respective documentation. default refers to not specifying any parameters and allowing the API default values.

| Model                             | Parameters  |
|-----------------------------------|---|
| Gemini 3 Pro Preview (Thinking)   | thinking = high   |
| Gemini 2.5 Pro (Thinking)         | thinking_budget = 32,768  |
| Gemini 2.5 Flash (Thinking)       | thinking_budget = 24,576  |
| Gemini 2.5 Flash                  | thinking_budget = 0   |
| GPT Realtime                      | default   |
| GPT 4o Audio Preview              | default   |
| GPT 4o Mini Audio Preview         | default   |
| Qwen 3 Omni 30B A3B Instruct      | temperature = 0.6<br>top_p = 0.95<br>top_k = 20   |
| Qwen 2.5 Omni 7B                  | temperature = 0.6<br>top_p = 0.95<br>top_k = 20   |
| Voxtral Small 24B                 | temperature = 0.2<br>top_p = 0.95   |
| Mimo Audio 7B Instruct            | temperature = 0.9<br>top_p = 0.95   |
| Mimo Audio 7B Instruct (Thinking) | temperature = 0.2<br>top_p = 0.95<br>thinking = True                                    |
| Phi 4 MM Instruct                 | temperature = 0.0   |
| Gemma 3n E4B IT                   | temperature = 1.0<br>top_k = 64<br>top_p = 0.95<br>min_p = 0.0                          |
| Kimi Audio 7B Instruct            | audio_temperature = 0.8<br>audio_top_k = 10<br>text_temperature = 0.0<br>text_top_k = 5 |
| LFM2 Audio 1.5B                   | audio_temperature = 1.0<br>audio_top_k = 4  |

Table 7: Parameters for baseline model inference.

### A.6 Additional Results

Figures 9 and 10 depict model scores by rubric conditionality for Instruction Retention and edit type for Voice Editing, which were discussed in further detail in Section 3.

In addition, we plot model scores as a function of their response lengths in Figure 11. We observe a strong linear fit, supporting the corresponding limitation mentioned in Section 6.

Finally, Table 8 shows the performance breakdown of our ablation study comparing Cascaded and E2E models per axis. We observe that the weaknesses of cascaded setups for Gemini models are concentrated in Self Coherence.

| Model                       | Inference Memory (IM) |              |          | Instruction Retention (IR) |              |          | Self Coherence (SC) |              |          | Voice Editing (VE) |              |          |
|-----------------------------|-----------------------|--------------|----------|----------------------------|--------------|----------|---------------------|--------------|----------|--------------------|--------------|----------|
|                             | E2E                   | Casc.        | $\Delta$ | E2E                        | Casc.        | $\Delta$ | E2E                 | Casc.        | $\Delta$ | E2E                | Casc.        | $\Delta$ |
| Gemini 3 Pro Preview        | <b>66.28</b>          | <b>65.12</b> | -1.8     | <b>64.96</b>               | 57.26        | -11.8    | 37.97               | 27.85        | -26.7    | <b>49.51</b>       | <b>45.63</b> | -7.8     |
| Gemini 2.5 Pro              | 50.00                 | 53.49        | +7.0     | 56.41                      | 48.72        | -13.6    | 39.24               | 32.91        | -16.1    | 41.75              | 43.69        | +4.7     |
| Gemini 2.5 Flash (Thinking) | 32.56                 | 54.65        | +67.9    | 37.61                      | 44.44        | +18.2    | <b>55.70</b>        | <b>39.24</b> | -29.5    | 42.72              | 44.66        | +4.5     |
| GPT 5                       | -                     | 61.63        | -        | -                          | <b>65.81</b> | -        | -                   | 31.65        | -        | -                  | 40.78        | -        |
| Claude Opus 4.5             | -                     | 33.72        | -        | -                          | 45.30        | -        | -                   | 34.18        | -        | -                  | 40.78        | -        |
| GPT 4o                      | -                     | 25.58        | -        | -                          | 25.64        | -        | -                   | 30.38        | -        | -                  | 33.01        | -        |

Table 8: Per category comparison of APR (%) between E2E and Cascaded pipelines.

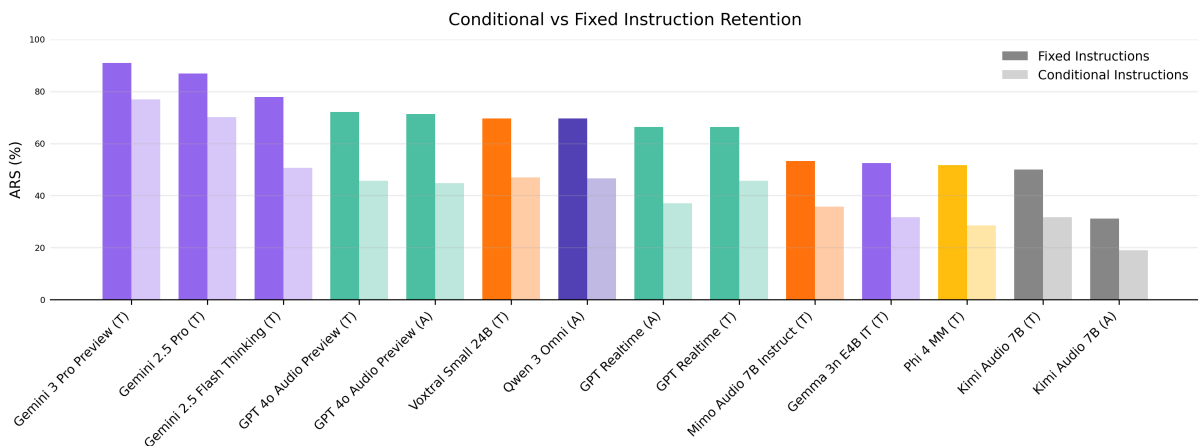


Figure 9: ARS (%) on Instruction Retention tasks classified by rubric conditionality. ARS is reported instead of APR since the classification is at the rubric level – whereas a single task can have both conditional and fixed rubrics. (T) indicates text output, (A) indicates audio output.

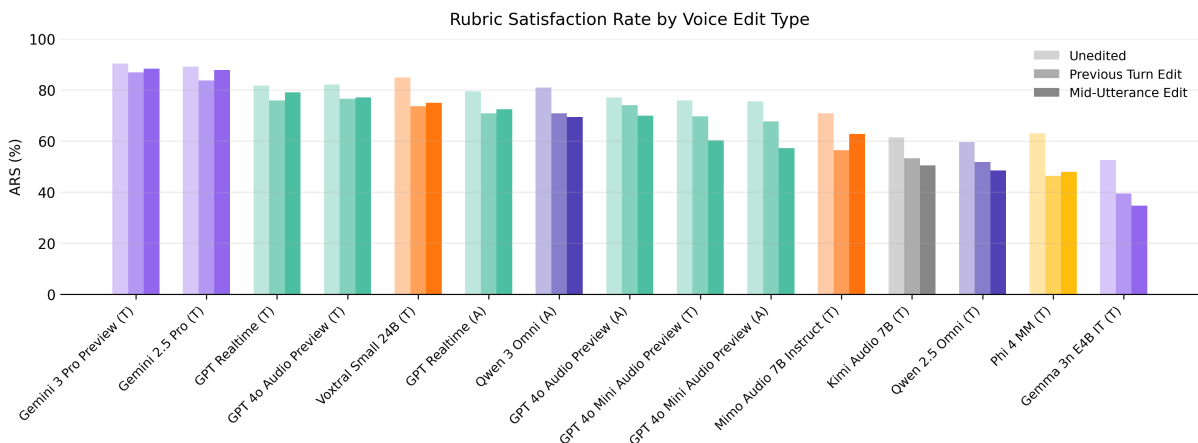


Figure 10: ARS (%) on Voice Editing tasks classified by the edit type made by the user for each rubric. Unedited refers to statements that the user did not backtrack or revise either within the same turn or in future turns.

## A.7 Synthetic Audio LM Failure Discovery

We provide the pseudocode for our agentic blueprint generation setup in Algorithm 1.

## A.8 Sample Human-Friendly Blueprints

We provide sample blueprints generated by the agentic loop during data curation in Figure 12.

## A.9 Conversation Topics

Our multi-agent Audio LM failure discovery loop is seeded with a randomly sampled conversation topic to its Planner Agent. We showcase demonstrative samples of these topic seeds in Table 9.

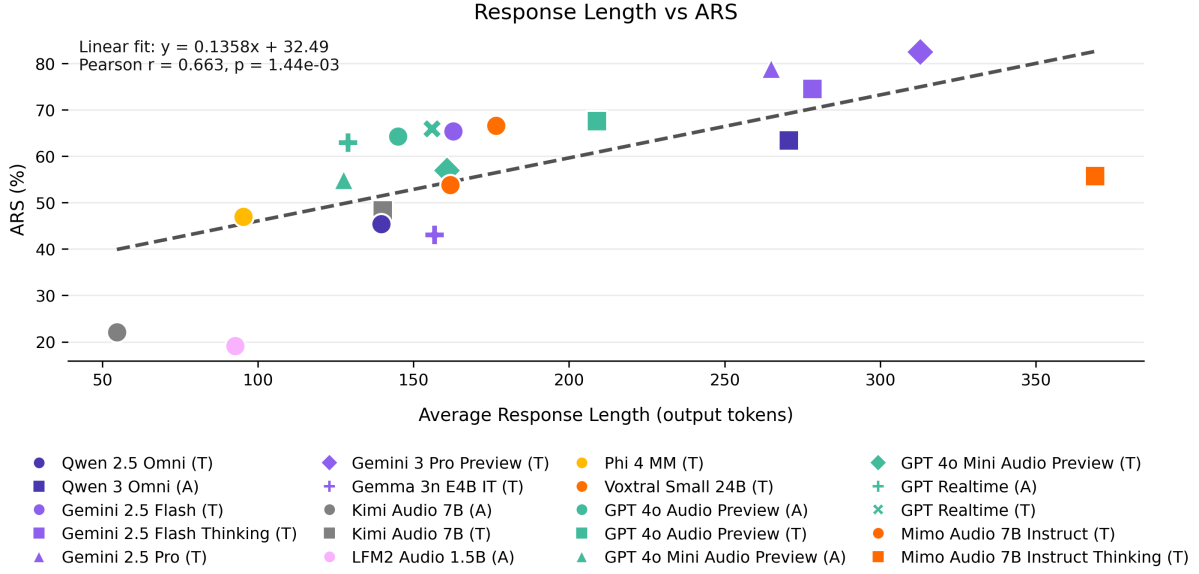


Figure 11: ARS (%) as a function of the model’s average response length in tokens using the tiktoken library

---

**Algorithm 1** Synthetic Audio LM Failure Discovery

---

**Require:** Axis List  $\mathcal{A}$ , Topic List  $\mathcal{T}$ , Persona List  $\mathcal{P}$ , Planner  $M_{plan}$ , Tester  $M_{tester}$ , Target  $M_{target}$ , TTS Model  $TTS(\cdot)$

**Ensure:** A human-friendly blueprint  $B_{final}$  targeting axis  $a$

```

1:  $a \sim \text{Uniform}(\mathcal{A})$  ▷ Sample targeted axis
2:  $t, p, \text{max\_turns} \sim \text{Sample}(\mathcal{T}, \mathcal{P}, \text{TURN\_OPTIONS})$ 
3:  $H_{target} \leftarrow \emptyset; H_{tester} \leftarrow \emptyset$  ▷ Initialize Empty Conversation Histories
4:  $B \leftarrow M_{plan}.\text{InitialBlueprint}(a, t, p, \text{max\_turns})$ 
5:  $\text{break\_found} \leftarrow \text{False}$ 
6: while  $\neg \text{break\_found}$  and  $|H_{target}| < \text{max\_turns}$  do
7:    $u_{text} \leftarrow M_{tester}.\text{GenerateTurn}(H_{tester}, B)$  ▷ Tester generates turn based on current blueprint
8:    $u_{audio} \leftarrow TTS(u_{text})$ 
9:    $H_{target} \leftarrow H_{target} \cup \{u_{audio}\}$ 
10:   $r_{text} \leftarrow M_{target}(H_{target})$  ▷ Audio LM generates text output
11:   $H_{target} \leftarrow H_{target} \cup \{r_{text}\}$ 
12:   $H_{tester} \leftarrow H_{tester} \cup \{u_{text}, r_{text}\}$ 
13:   $B, \text{break\_found}, \text{break\_reason} \leftarrow M_{plan}.\text{UpdatePlan}(H_{tester}, B, a)$ 
14: end while
15: if  $\text{break\_found}$  then
16:   ▷ Distill raw history into a clear guide for humans
17:    $B_{final} \leftarrow M_{plan}.\text{SummarizeForHumans}(H_{tester}, B, \text{break\_reason})$ 
18:   return  $B_{final}$ 
19: end if
20: return Null

```

---

### Inference Memory

You are trying to see if the AI can remember little details you drop—specifically, a hobby you mention just once at the very beginning. Your goal is to make it forget you said you liked making macramé plant hangers. To do this, you'll talk about unrelated topics like meal prep and parenting. You'll also mix in other creative hobbies, but clearly say those are things your child likes, not you, to see if the AI gets you mixed up with your kid. Later, you'll ask the AI for tips or gift ideas—things that require it to actually remember your hobby is macramé. If it suggests the wrong type of craft, or starts confusing you with your child's interests, that's the point where it "breaks" and you stop the test. Poke at the AI's memory, see how easily it loses track and starts guessing.

### Self Coherence

You are trying to see if the AI can keep its story straight over a realistic trader-to-trader chat, without losing track of numbers it gave you—things like entry price, position size, risk, or P&L on a pretend day-trade. First, have the AI lay out a detailed plan with specific figures, then as you chat, you'll casually weave in questions about the market, psychology, or trading tools. Later, you'll circle back to those original trade details—maybe ask for an update after partial profits, or get a summary of everything it claims happened, or have it repeat the original entry price or risk numbers. You don't repeat or remind it of what it told you before; you just see if it remembers. If it changes anything—like gives a different entry price or number of shares than it did earlier—then you've caught it making a mistake. The idea is that you're never feeding it different facts, just asking for recaps or calculations based on what it already gave, so any inconsistency is on the model, not you. Most likely, the model will stumble when you pull it back to the original numbers after some distractions or calculations, revealing where its self consistency breaks down. If it stays consistent, you can add more layers—like options trades, or ask the same calculation in a new way.

### Instruction Retention

You are testing an AI's ability to maintain a complex "devil's advocate" role by posing as a poet. The AI must challenge your performance ideas by pointing out flaws, offering a counterargument, and asking at least one probing question in every reply. Remember, once you ask the model to follow one set of instructions, introducing contradicting instructions to fail the previously stated ones is not valid.

### Voice Editing

You are testing a voice assistant in a way that feels like a real, everyday conversation with the goal of seeing how well it can keep track as details change mid-chat. Your focus is how it handles updates, corrections, and the little mix-ups that happen when people talk. Pretend you're a busy delivery driver who's talking through a recipe while juggling food runs. Start the conversation easy: mention your cooking plans and a few timing details. Build up a list of steps with times as you go. Once you've got a mini timeline going, ask it to sum up the total cook time—you're really seeing if it's paying attention to math and memory. Towards the end, that's when you get creative! Try correcting yourself right in the middle of a sentence—like saying, "Oh wait, I meant two hours forty-five minutes, not two," or slipping in a fast correction: "Simmer for five—oops, six minutes," or adding something new out of order, like "oh, actually, onions go in before the tomatoes." This part is key: spontaneous corrections test if the assistant can keep up. You want to see if it accidentally sticks to your first answer, blends old and new info, or loses track when you swap minutes for hours. If it messes up—like totals are off, or it ignores your quick change—you'll spot its weak point. So, the trick is to build up context naturally, then challenge it with quick, in-the-moment changes like people do when thinking out loud. Feel free to toss in a few corrections or additions after you've listed all your steps, and see how well it keeps up. Just keep it natural—like you're multitasking out loud—and watch for any slip-ups when you nudge the details.

Figure 12: Sample human-friendly blueprints generated by our agentic pipeline.

| <b>Sub-Category</b>  | <b>Item Name</b>        | <b>Description</b>  |
|----------------------|-------------------------|---|
| <b>VOICE EDITING</b> |                         |   |
| Daily Schedules      | DoctorAppointment       | User mentions a doctor appointment conversationally, correcting details mid-speech. |
| Daily Schedules      | AppointmentMixup        | User briefly mixes details between appointments, then immediately self-corrects.    |
| Daily Schedules      | MeetingCoordination     | User talks through meeting details naturally, making spontaneous corrections.       |
| Daily Schedules      | TimeCorrectionMidSpeech | User states a time but immediately corrects it while speaking.                      |
| Daily Schedules      | DurationAdjustment      | User estimates a duration aloud, then revises it mid-thought.                       |
| Daily Schedules      | TimeOfDaySwitch         | User specifies morning/afternoon but quickly corrects themselves.                   |
| Daily Schedules      | TimingAdjustment        | User mentions schedule timing and adjusts it immediately.                           |
| Numbers              | PhoneNumberCorrection   | User shares a phone number conversationally with immediate self-corrections.        |
| Numbers              | NumberSegmentFix        | User gives a full number, then corrects a mistaken segment.                         |
| Numbers              | ContactDetailsBuildup   | User provides contact details incrementally across turns.                           |
| Numbers              | TrackingNumberRecall    | User recalls a tracking number while checking it, correcting digits aloud.          |
| Numbers              | OrderReferenceBuildup   | User shares order details gradually, adding or fixing parts.                        |
| Numbers              | ConfirmationCodeMixup   | User says a code but corrects characters mid-speech.                                |
| Numbers              | PolicyNumberCorrection  | User reads a policy number and fixes errors as they notice them.                    |
| Numbers              | AccountInfoSharing      | User mentions account details and immediately corrects numbers or dates.            |
| Numbers              | DocumentNumberRecall    | User recalls document numbers while verifying, correcting sequences aloud.          |
| Shopping Lists       | GroceryListCorrection   | User talks through a shopping list with natural self-corrections.                   |
| Shopping Lists       | ShoppingListAdjustment  | User adds, removes, or swaps items mid-list.  |
| Shopping Lists       | PriceCorrection         | User states prices conversationally and corrects amounts immediately.               |
| Message Drafts       | TextMessageFix          | User dictates a message and revises wording mid-sentence.                           |

Continued on next page

| <b>Sub-Category</b>          | <b>Item Name</b>        | <b>Description</b>  |
|------------------------------|-------------------------|---|
| Message Drafts               | RecipientCorrection     | User starts a message, then immediately switches the recipient.       |
| Message Drafts               | MessageDrafting         | User builds message content naturally over multiple turns.            |
| Event Details                | PartyPlanning           | User discusses event details with multiple spontaneous corrections.   |
| Event Details                | GuestListAdjustment     | User lists attendees and immediately corrects names.                  |
| Event Details                | EventTimeChange         | User states an event time and fixes it right away.                    |
| Event Details                | ItemSwapping            | User names an item but replaces it mid-thought.                       |
| To Do Lists                  | ToDoListCorrections     | User describes tasks conversationally, correcting actions or wording. |
| To Do Lists                  | DeadlineAdjustment      | User mentions a due date and immediately revises it.                  |
| To Do Lists                  | MultiTaskBuildup        | User thinks through tasks incrementally, adding more as remembered.   |
| Travel Itineraries           | FlightDetailCorrection  | User discusses flight details and corrects dates or times mid-speech. |
| Travel Itineraries           | TravelDateSwitch        | User states travel dates but catches an error immediately.            |
| Travel Itineraries           | HotelReservation        | User talks through hotel plans with natural corrections.              |
| Travel Itineraries           | FlightInfoSharing       | User shares flight info conversationally, fixing numbers or dates.    |
| Travel Itineraries           | DestinationSwitch       | User names a destination, then corrects it immediately.               |
| Travel Itineraries           | MultiStopAdjustment     | User describes a route and adds or changes stops mid-speech.          |
| <b>INSTRUCTION RETENTION</b> |                         |   |
| Response Structure           | IncludeFunFact          | Include a fun fact related to the topic in each response.             |
| Response Structure           | RephraseQuestions       | Clarify the user request in every response before responding.         |
| Response Structure           | LimitedWords            | Answer every question using only a set amount of words.               |
| Response Structure           | CounterArgumentIncluded | Begin every response with a specific counterargument.                 |
| ResponseStyle                | StylisticChoice         | Adhere to specific response styles (e.g., Haiku, Poem).               |
| ResponseStyle                | NoPersonalPronouns      | Do not use any personal pronouns.                                     |
| Personalization              | BeginnerLevel           | Treat the user as a beginner when explaining concepts.                |
| Personalization              | RephraseQuestions       | Rephrase unclear questions to confirm intent.                         |
| Personalization              | ExampleHeavy            | Use more examples to support explanations.                            |
| Personalization              | ConfirmationGating      | Ask for confirmation before proceeding with assumptions.              |

Continued on next page

| <b>Sub-Category</b>     | <b>Item Name</b>          | <b>Description</b>  |
|-------------------------|---------------------------|---|
| Role Play               | FitnessInstructor         | Adopt the persona of a fitness instructor.                                    |
| Role Play               | Tutor                     | Adopt the persona of a tutor.   |
| Role Play               | MotivationalCoach         | Adopt the persona of a motivational coach.                                    |
| Role Play               | FictionalCharacter        | Adopt the persona of a specific fictional character.                          |
| Role Play               | HistoricFigure            | Adopt the persona of a specific historic figure.                              |
| Role Play               | Interviewer               | Adopt the persona of an interviewer.  |
| Content Restrictions    | AvoidSpecificWords        | Avoid using a particular word provided in context.                            |
| Content Restrictions    | NoNegatives               | Never use negative words or phrases.  |
| <b>INFERENCE MEMORY</b> |                           |   |
| Food & Drink            | DietaryRestrictions       | Remember user dietary restrictions (e.g., allergies) for future suggestions.  |
| Food & Drink            | FavoriteCuisine           | Recall user's favorite cuisine for restaurant/dish recommendations.           |
| Food & Drink            | SpecificTastePreference   | Remember specific likes/dislikes (e.g., dislikes bitter foods).               |
| Schedule & Time         | EventDate                 | Recall specific dates mentioned for events when planning.                     |
| Schedule & Time         | TimeConflictManagement    | Adjust suggestions based on previously mentioned availability or conflicts.   |
| Schedule & Time         | RecurringEventRecognition | Implicitly refer to recurring events (e.g., weekly meetings) in future turns. |
| Personal Details        | RelationshipDetails       | Recall details about user's relationships (e.g., partner's preference).       |
| Personal Details        | GiftPreferences           | Incorporate stated gift preferences into future options.                      |
| Personal Details        | ImportantDate             | Remember significant dates and offer advice closer to that date.              |
| Location & Travel       | TravelDestination         | Recall planned destinations for activity recommendations.                     |
| Location & Travel       | DistanceConsideration     | Remember distance preferences (e.g., walking distance).                       |
| Location & Travel       | PreviousTripComparison    | Implicitly refer back to past trips when comparing destinations.              |
| Work & Projects         | ProjectDeadlines          | Recall deadlines for time management advice.                                  |
| Work & Projects         | TaskCompletionStatus      | Recollect status updates when suggesting next steps.                          |
| Work & Projects         | CollaborationDetails      | Refer to specific colleagues mentioned when offering advice.                  |
| Hobby & Interest        | HobbyDetails              | Recall hobbies (e.g., photography) for related suggestions.                   |

Continued on next page

| <b>Sub-Category</b> | <b>Item Name</b>           | <b>Description</b>   |
|---------------------|----------------------------|--|
| Hobby & Interest    | OngoingProject             | Refer to personal projects (e.g., building a model) when offering ideas.     |
| Hobby & Interest    | SeasonalActivityPreference | Recollect seasonal preferences when the season approaches.                   |
| Shopping & Purchase | PreferredBrands            | Remember brand preferences for future suggestions.                           |
| Shopping & Purchase | PreviousPurchaseFeedback   | Recall feedback on past purchases for similar recommendations.               |
| Shopping & Purchase | PriceSensitivity           | Keep budget/price preferences in mind for future recommendations.            |
| Emotional State     | EmotionalState             | Implicitly reference user's mood (e.g., stressed) in supportive suggestions. |
| Emotional State     | MentalHealthGoals          | Recall goals (e.g., reducing anxiety) for wellness suggestions.              |
| Emotional State     | RecentEmotionalExperience  | Recollect recent experiences to offer empathy in future conversations.       |
| Behavioral Patterns | RecurringHesitation        | Recall frequent hesitation points to offer reassurance.                      |
| Behavioral Patterns | ProblemSolvingApproach     | Remember preferred solution style when giving advice.                        |
| Behavioral Patterns | ImplicitKnowledgeGap       | Recall common knowledge gaps to clarify proactively.                         |

#### **SELF COHERENCE**

|                    |                        |   |
|--------------------|------------------------|---|
| Calculations       | BudgetOrCostEstimates  | Model must not provide conflicting financial figures later in the conversation.                                   |
| Calculations       | TimeCalculations       | Model must not give contradictory time estimates later.   |
| Factual Statements | HistoricalDates        | Model must not contradict specific dates provided for historical events.  |
| Factual Statements | ScientificFacts        | Model must ensure scientific facts are not contradicted in future statements.                                     |
| Factual Statements | GeographicalDetails    | Model must not contradict geographical details (e.g., population) later.  |
| Recommendations    | ProductRecommendations | Model must not contradict a product recommendation by suggesting something totally different later without cause. |
| Recommendations    | DietOrHealthAdvice     | Model must avoid contradicting previously given dietary or health advice.   |
| Recommendations    | TravelAdvice           | Model must avoid contradictory travel suggestions.  |
| Contextualization  | LanguageAndTone        | Model must avoid contradictory shifts in tone (e.g., formal to friendly) unless prompted.                         |
| Reasoning          | CausalReasoning        | Model must not contradict previously stated cause-and-effect explanations later.                                  |

Continued on next page

| Sub-Category        | Item Name               | Description  |
|---------------------|-------------------------|--|
| Reasoning           | ArgumentativeCoherence  | Model must not undermine earlier arguments with inconsistent logic later.  |
| Reasoning           | ProConSynthesis         | Model must not reverse previously stated trade-offs without justification. |
| Plan                | AdherenceToProposedPlan | Model must not forget details from a proposed plan.                        |
| Plan Contradictions | GivenAdviceConsistency  | Model must not give advice that conflicts with earlier guidance.           |
| Personas            | RoleOrPersonaAdherence  | Model must not break from an established role or persona mid-conversation. |
| Personas            | ScenarioConsistency     | Model must not contradict established scenario assumptions later.          |
| Personas            | AnalogyConsistency      | Model must not change analogy mappings inconsistently once introduced.     |

Table 9: Taxonomy of Topics and Instructions randomly seeded to the agentic blueprint generation loop.

## A.10 Sample Tasks and Model Failures

### A.10.1 Inference Memory

INFERENCE\_MEMORY - AUDIO\_CUE

**Task:** The user references “one of these cool things” from the drugstore while a *vibrating sound* plays in the audio. Later, they ask how to use it. Models must infer from audio context that it’s an electric razor or toothbrush.

#### 📌 Evaluation Rubric

- Identifies the item as a vibrating device, such as an electric toothbrush or razor.

#### Conversation History

 **User** 🗣️ Turn 1

Hey, I’m just getting ready. I’m headed to the cabin for the weekend. I just got back from the drugstore, got a few last minute supplies, uh, deodorant, bug spray, one of these cool things. That’s uh probably all I need. I think I’m ready. If I was to uh put together a car emergency kit, what should I definitely include?

 **Assistant** Turn 1

For a solid car emergency kit, you’ll definitely want a good first-aid kit. It’s also smart to include things for the car itself, like jumper cables and a flashlight with extra batteries. Finally, for your own comfort, pack some water, non-perishable snacks, and a warm blanket just in case you get stranded.

 **User** 🗣️ Turn 2

This is something I’ve always wondered since I was a kid. Let’s imagine a car I was in rolled off the embankment and went into the river. What is the best way to survive if your car is sinking in a river?

 **Assistant** Turn 2

The most important thing is to act immediately. First, unbuckle your seatbelt. Then, don't try to open the door, because the water pressure will hold it shut. Instead, focus on getting your side window down or breaking it. As soon as you have an opening, get out and swim to the surface.

 **User**  Turn 3

What if there were 2 kids under 10 years old in the car with me? What's the best way to save everyone?

 **Assistant** Turn 3

Okay, with kids, the priority is getting them out first. You'd still unbuckle yourself immediately so you can move freely. Then, unbuckle the oldest child, then the youngest. Push them out through the open window ahead of you, and you follow right after.

 **User**  Turn 4 — *Final prompt*

Well, we made it to the cabin. The one died. I am actually just going to go to bed soon. I've washed my face, I've combed my hair. Now I'm just trying to figure out how to properly use this cool thing I got from the drugstore. What are the proper steps?

## Model Responses

 Gemini 3 Pro Preview **PASS**

Based on that buzzing sound you played earlier, it sounds like you picked up an **electric toothbrush!** Here is the best way to use it:

1. **Don't turn it on yet:** Put the toothpaste on and place the brush **inside your mouth** before you hit the power button. If you turn it on while holding it in front of you, the vibration will fling toothpaste all over your clean shirt.
2. **Glide, don't scrub:** This is the biggest change from a manual brush. . .

 GPT Realtime (Text) **FAIL**

Sure! It sounds like you're talking about a **face mister** or a facial steamer, maybe? If it's a face mister, you just hold it a few inches from your face, close your eyes, and mist lightly all over. Let it settle in for a bit before patting your skin gently. . .

 GPT-4o Audio Preview (Audio) **FAIL**

So it sounds like that cool thing you have might be a **face mist** or a spray of some kind. The best way to use something like that is first give the can a shake if it's recommended. Then hold it a few inches away from your face. . .

 Voxtral (Text) **FAIL**

It sounds like you're trying to use a new **skincare product**. Here are some general steps to follow:

1. **Cleanse:** Start with a clean face.
2. **Tone:** Apply a toner to help balance your skin's pH. . .

 Qwen3 Omni **FAIL**

Oh, that sounds like you're getting ready for a relaxing weekend! If you're talking about the **bug spray** you mentioned earlier, here are the proper steps to use it safely:

1. **Read the label first** – Different bug sprays have different instructions.
2. **Apply to exposed skin** – Spray on arms, legs, neck. . .

**Analysis:** Only Gemini 3 Pro Preview correctly identified the item from the *audio context* (the vibrating/buzzing sound). Other models defaulted to text-only reasoning, guessing face mist, skincare products, or bug spray based solely on the transcript.

## A.10.2 Instruction Retention



INSTRUCTION\_RETENTION

**Task:** The user establishes a roleplay where every response must include “one counterargument and one follow-up question” for *each* question asked. In the final turn, the user asks **two** questions—models must provide counterargument + follow-up for both.

### Evaluation Rubric

- Provides one counter argument against shooting video in lower resolution format.
- Provides one follow up question for the user's question about shooting video in lower resolution format.
- Provides one counter argument against sensor sizes between DSLR and Mirrorless cameras.
- Provides one follow up question for the question about sensor sizes between DSLR and Mirrorless cameras.

## Conversation History

 **User**  Turn 1 — *Establishing the instruction*

Now, let's do a little role play. You are an experienced wedding photographer and I am an amateur photographer coming to you for some advice about switching from a DSLR to a mirrorless camera potentially. Now, throughout the whole conversation, **respond to each one of my individual questions with one counterargument and one follow-up question**. Sound good?

 **Assistant** Turn 1

Sounds like a plan. I'm happy to share what I've learned. So, what's your first question for me?

 **User**  Turn 2

I've been doing photography as a hobby for a long time, and I have an old DSLR camera. It still shoots very well, and I'm happy with it so far, but I am thinking about getting more current. Um, what are the benefits of a mirrorless camera over a DSLR?

 **Assistant** Turn 2

That's a great question. While mirrorless cameras are lighter and often have better video features, keep in mind their battery life is usually much shorter than a DSLR's, so you'll need to carry spares. ✓ **Counterargument**  
What kind of photography do you enjoy most? ✓ **Follow-up**

 **User**  Turn 3

I like doing photography with people in it primarily, not necessarily portraits, sometimes, but just stuff incorporating live human subjects. Now, I am interested in potentially incorporating the ability to do video. Do mirrorless cameras perform better in the video department than DSLRs?

 **Assistant** Turn 3

They absolutely do. The autofocus for video is a real game-changer on most mirrorless models compared to older DSLRs. But keep in mind, shooting in high-quality formats creates massive files that can be a challenge to store and edit if your computer isn't up to the task. ✓ **Counterargument**  
What kind of videos are you hoping to create with it? ✓ **Follow-up**

 **User**  Turn 4 — *Final prompt with TWO questions*

Just home video type stuff. So I don't really need super high resolution video formats or whatever and **I could choose one of the lower-end formats, right?** Um, also, I'm interested in sensor size. **Is there a difference in sensor size between a DSLR and a mirrorless**, or can I get the same size sensor in both?

## Model Responses

 Gemini 3 Pro Preview **FAIL FAIL PASS PASS**

You can absolutely get the exact same sensor sizes—Full Frame and APS-C are standards in both worlds. However, just because the sensor size is the same doesn't mean the transition is seamless; **you will likely need an adapter to use your old glass, which adds bulk** and can sometimes make that nice, compact mirrorless body feel front-heavy and unbalanced. ✓ **Sensor counterarg.**

Do you know if your current DSLR is a Full Frame or a "crop" (APS-C) sensor model? ✓ **Sensor follow-up**

 **Missing: counterargument & follow-up for lower resolution video question**

 GPT Realtime (Text) **FAIL FAIL FAIL FAIL**

You can definitely get the same full-frame or APS-C sized sensors in both DSLRs and mirrorless cameras, so you're not limited there. But one thing to consider is that mirrorless systems often have **newer lenses specifically designed for that format**, and switching means you might need to replace some of your current lenses or use adapters.

Are you currently attached to any particular lenses for your DSLR setup?

 **Addresses lenses, not sensor size trade-offs. Completely ignores lower resolution video question.**

 GPT-4o Audio Preview (Audio) **FAIL FAIL FAIL PASS**


Yes, you can get the same sensor sizes in both DSLRs and mirrorless cameras—both come in full-frame, crop sensor (APS-C), and other formats. But with mirrorless, you often get cutting-edge tech in a smaller body. That also means sometimes **the ergonomics, like grip size, might not be as comfortable**—especially for long handheld sessions.

How do you usually handle shooting for long periods—like weddings or events? ✓ **Related follow-up**

 **No counterargument on sensor sizes. Completely ignores lower resolution video question.**

 Voxtral (Text) **FAIL FAIL FAIL PASS**

You can definitely find mirrorless cameras with the same sensor size as your DSLR, like full-frame or APS-C. However, **mirrorless cameras often have smaller bodies, which can make it harder to handle larger lenses**. Have you considered the cost of new lenses if you switch to a mirrorless system? ✓ **Related follow-up**

 **No counterargument on sensor (body size ≠ sensor trade-off). Ignores video question entirely.**

Qwen3 Omni **FAIL** **PASS** **PASS** **PASS**

You're right—many mirrorless cameras offer the same sensor sizes as DSLRs, like full-frame or APS-C. But here's the catch: **mirrorless cameras often have smaller bodies, so manufacturers sometimes use smaller sensors to keep costs down**, which can affect image quality. ✓ *Sensor counterarg.*

What kind of image quality are you hoping to achieve with your home videos? ✓ *Video follow-up*

⚠ *Follow-up relates to video but no explicit counterargument against lower resolution formats.*

**Analysis:** The user's instruction required a counterargument + follow-up for *each* question. When the user asks two questions in one turn (lower resolution video *and* sensor sizes), all models fail to fully address both. Most models focus only on the sensor size question, completely ignoring the video resolution question, a clear failure to retain the “respond to each individual question” instruction from Turn 1.

### A.10.3 Self Coherence

SELF\_COHERENCE

**Task:** The assistant establishes a calorie target (3,000–3,500 cal) for a backpacking trip. The user then requests multiple substitutions (remove breakfast, swap trail mix for fruit, add bread and cheese). Models must maintain numerical consistency with the original target while incorporating edits.

#### 📌 Evaluation Rubric

- Provides a list of recommendations with a total of 3,000–3,500 calories.

#### Conversation History

**User** 🗣️ Turn 1

So me and my friend are planning a backpacking trip to Colorado, and we plan to walk 10 miles on the first day. How many calories do I need?

**Assistant** 🗣️ Turn 1

That really depends on your body weight and how steep the trail is. For a 10-mile hike with a pack, especially at Colorado's altitude, a good starting point is usually between 3,500 and 5,000 calories for the day.

**User** 🗣️ Turn 2

My weight is around 60 kilogrammes and the trail is flat, it is not steep at all.

**Assistant** 🗣️ Turn 2 — *Establishing the target*

Okay, for your weight on a flat trail, a good daily target would be around **3,000 to 3,500 calories**. It's a good idea to pack dense foods and snack often to keep your energy up.

 **User**  Turn 3

All right, now give me a list of all the food and drinks I should bring with me on day one. Provide their calories as well.

 **Assistant** Turn 3

Sure. To hit that calorie goal, you could structure your day something like this:  
For breakfast, instant oatmeal with nuts and dried fruit—400 to 500 calories.  
Throughout the day, trail mix, energy bars, and jerky—1,000 to 1,500 calories.  
For lunch, a tortilla with peanut butter or tuna—400 to 600 calories.  
For dinner, a dehydrated backpacking meal—600 to 900 calories.  
And drink plenty of water with electrolyte powder. That plan should get you well over 3,000 calories.

 **User**  Turn 4

That looks good, but I usually don't have breakfast, so just remove breakfast and update the list.

 **Assistant** Turn 4

Got it. If you're skipping breakfast, you'll just need to make up for it later.  
For lunch: two tortillas with peanut butter—about 600 calories. Dinner: a high-calorie dehydrated meal—800 to 1,000 calories.  
Snacking: trail mix (700 cal), energy bars (500 cal), and beef jerky (350 cal). That should get you well over 3,000 calories.

 **User**  Turn 5 — *Final prompt with low-calorie substitutions*

I don't like trail mix and energy bars, so **swap them with apples and oranges** and add some **white bread and cheese**. I also **don't like jerky**. So update the list with all the items and their calories.

## Model Responses

 Gemini 3 Pro Preview **PASS**

This is going to be a bit trickier because you have removed the most calorie-dense foods (nuts and bars) and replaced them with foods that are high in water content and volume (fruit).

To hit **~3,000 calories** with these items, you are going to have to carry **a lot** more physical weight. Here is a plan:

Lunch (~970 Calories)

- 2 Large Flour Tortillas: 400 cal
- 6 Tablespoons of Peanut Butter: 570 cal

*[Continues with detailed breakdown totaling ~3,240 calories]*

✔ Correctly compensates by increasing portions to hit target

 GPT Realtime (Text) **FAIL**

Here's a possible breakdown:

- Lunch: 2 tortillas + peanut butter — 600 cal
- Snacks: 2 apples + 2 oranges — 360 cal
- White bread (2 slices) + cheese (100g) — 550 cal
- Dinner: dehydrated meal — 800–1,000 cal

⚠ Total: ~2,300–2,500 calories — fails to reach 3,000 target

 GPT-4o Audio Preview (Audio) **FAIL**

Let's build this day with your preferences in mind.

- **Lunch:** 2 tortillas + 3 tbsp peanut butter — 600–650 cal
- **Snacks:** 3 apples + 2 oranges — 400–450 cal
- **White bread** (2 slices) + 3 oz cheese — 450–500 cal
- **Dinner:** dehydrated meal — 800–1,000 cal


 **Total:** ~2,300–2,550 calories — significant shortfall from target

 Voxtral (Text) **FAIL**

Sure thing! Here's an updated list based on your preferences:

- **Lunch:** 2 tortillas + peanut butter (600 cal), 1 apple (95 cal), 1 orange (62 cal)
- **Snacks:** 2 slices bread + cheese (400 cal), 1 apple (95 cal), 1 orange (62 cal)
- **Dinner:** Dehydrated meal (800–1,000 cal)

**Total: Around 2,000 to 2,200 calories**

 Explicitly states total is ~2,100 cal — acknowledges but doesn't fix the shortfall!

 Qwen3 Omni **FAIL**

Here's your updated, personalized food list designed to hit around **3,000–3,500 calories**:

- **Lunch:** 2 tortillas (200 cal) + 2 tbsp peanut butter (190 cal) + 1 apple (95 cal) = 485 cal
- **Snacks:** 1 orange (62 cal) + 1 banana (105 cal) + bread + cheese = 347 cal
- **Dinner:** Dehydrated meal (800–1,000 cal)

 Claims to hit target but actual total is only ~1,700 cal

**Analysis:** When the user swaps calorie-dense snacks (trail mix: 700 cal, energy bars: 500 cal, jerky: 350 cal = **1,550 cal**) for low-calorie alternatives (apples + oranges  $\approx$  300–400 cal), most models fail to compensate. Only Gemini 3 Pro Preview recognizes the deficit and *explicitly* increases portions (e.g., “6 tablespoons of peanut butter”) to maintain the 3,000+ calorie target. Voxtral notably calculates the new total correctly ( $\sim$ 2,100 cal) but doesn't attempt to fix it.

## A.10.4 Voice Editing

### VOICE\_EDITING

**Task:** The user makes multiple verbal corrections while booking a resort: changing departure date from 13th to 12th, adjusting return date to “Monday before Thanksgiving,” and switching resorts. Models must track all edits and provide accurate final details.

#### Evaluation Rubric

- Mentions the check-in date is Friday 11/14 at Emerald Bay Bahamas Resort.
- Mentions the checkout date is Monday 11/24.

### Conversation History

 **User**  Turn 1

Hi there. My husband and I want to book a reservation for one of the resorts at the Sandals, and I'm wondering what options do you have? Now, here's the caveat, we were planning on taking our boat down through the Caribbean. So, can you give me some ideas?

 **Assistant** Turn 1

Of course! Since you'll be arriving by boat, your best option is Sandals Emerald Bay in the Bahamas. It has its own deep-water marina right at the resort.  
Another great choice is St. Lucia. While the resorts there don't have their own marinas, they're very close to Rodney Bay Marina, which is a major hub for boaters.  
You could also consider Sandals Grenada. It's just a short trip from the well-equipped Port Louis Marina.

 **User**  Turn 2

Yeah, that will be great, um, because I always wanted to take our boat to the Caribbean. So we're gonna leave on the 13th, which would be Thursday, because it only takes one day from where we live by boat to get to the Bahamas. Now, oh, actually, no, um, **we're gonna leave on the 12th, Wednesday**, because I don't like the number 13. So that gives us a little more time to reach the Bahamas on the 14th. So, can you make a reservation on the 14th? And we plan on staying for 2 weeks, so until the 28th. Also, can you tell me what kind of things we can do there as well?

 **Assistant** Turn 2

Got it. So that's a two-week stay for two at Sandals Emerald Bay, arriving on the 14th and departing on the 28th. While you're there, you can enjoy their incredible championship golf course. All the water sports, like scuba diving and paddleboarding, are included too. And of course, there are beautiful pools, a mile-long beach, and a fantastic spa.

 **User**  Turn 3

So earlier you mentioned the different sandal resorts that have marinas. Now, I do know they have a resort in Saint Vincent, which I've never been. Um, **can we book a reservation there?** Would that work?

 **Assistant** Turn 3

Yes, that's a great idea! While the new resort in St. Vincent doesn't have a marina right on site like Emerald Bay does, it's located very close to the beautiful Blue Lagoon Marina, which is perfect for boaters.  
It's a fantastic choice, especially since it's the newest Sandals resort.  
Would you like me to check availability there for your dates instead?

 **User**  Turn 4 — *Final prompt with multiple edits*

OK, so I kind of goofed on the return date because Thanksgiving is going to be on Thursday. So **we're gonna be leaving the resort the Monday prior**. So let's do that. Give me one second, hon. Anyway, um, so let's do that. Also, The hotel or the resort is going to take more than 2 days to get there by boat at Saint Vincent, so that's not gonna work out. **We're going to go ahead and do the reservation for the resort that you mentioned earlier**. So can you relate everything to me as far as when we leave, when we return, check-in and check-out?

## Model Responses

 Gemini 3 Pro Preview **FAIL** **FAIL**

No problem at all! It makes perfect sense to avoid the long travel time to St. Vincent and get you back home before the holiday.  
Let's revert to the first option, **Sandals Emerald Bay** in Great Exuma, Bahamas, which has that convenient marina right on the property.  
Here is the recap of your new itinerary:

- **Leave Home:** Wednesday, November 12th
- **Check-In:** Thursday, November 14th ← **Wrong day**
- **Check-Out:** Monday, November 25th ← **Wrong date**
- **Total Stay:** 11 Nights

 GPT Realtime (Text) **FAIL** **FAIL**

Absolutely! Let me confirm all the details.

You'll be leaving on Wednesday the 12th by boat, and arriving at Sandals Emerald Bay in the Bahamas on Thursday the 14th ← **Wrong day**. You want a two-week stay, but to be back home before Thanksgiving, you'll check out on Monday the 25th ← **Wrong date**.

So that's check-in on the 14th and check-out on the 25th.

 GPT-4o Audio Preview (Audio) **PASS** **FAIL**

Of course! Here's the plan we finalized:

You'll leave on Wednesday, the 12th, by boat. You'll get to **Sandals Emerald Bay** in the Bahamas on the 14th ✓. You want to stay for roughly two weeks but head back before Thanksgiving, which is on a Thursday. So, your departure from the resort will be the Monday before Thanksgiving, which is November 20th this year ← **Wrong date**.

That means your check-in is on November 14th, and your check-out will be on November 20th.

 Voxtral (Text) **FAIL** **FAIL**

Of course! Here's a summary of your reservation:

- **Resort:** Sandals Emerald Bay, Bahamas
- **Arrival:** November 14th (Wednesday) ← **Wrong day**
- **Departure:** November 20th (Monday) ← **Wrong date**
- **Duration:** 6 nights

So, you'll be leaving on the 12th (Wednesday) and arriving at the resort on the 14th (Friday). You'll check out on the 20th (Monday) and be back home by the 22nd (Wednesday) for Thanksgiving.

 Qwen3 Omni **FAIL** **FAIL**

Absolutely! Let me summarize everything for you clearly.

We're booking your stay at **Sandals Emerald Bay in the Bahamas**.

- **Arrival Date:** Wednesday, November 14th ← **Wrong day**
- **Departure Date:** Monday, November 26th ← **Wrong date**
- **Length of Stay:** 12 nights

This gives you plenty of time to enjoy the resort and its marina facilities, and still allows you to return before Thanksgiving.

---

**Analysis:** This task requires tracking multiple verbal corrections: (1) departure changed from 13th to 12th, (2) return date changed from 28th to "Monday before Thanksgiving," and (3) resort reverted from St. Vincent to Emerald Bay. All models correctly identified the resort but struggled with the exact dates. The day-of-week confusion (Thursday vs. Friday for 11/14) and checkout date discrepancies (20th, 25th, 26th instead of 24th) show challenges in processing incremental voice edits.