VOX-INFINITY: BENCHMARKING THE LIMITS OF LONG-CONTEXT SPOKEN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-context understanding remains a fundamental challenge for large language models, as excessively long inputs often lead models to forget salient information. This issue is even more pronounced in the speech domain, where audio, as a low-compression modality, requires substantially more embeddings than text to preserve both semantic content and acoustic cues. To address this challenge, we introduce **Vox-Infinity**, the first benchmark specifically designed to evaluate long-context understanding in spoken language models. Vox-Infinity systematically extends audio history along two dimensions: turn count and turn duration. It covers a diverse range of representative scenarios with varying interaction structures and semantic complexity. Crucially, Vox-Infinity provides explicit answer-provenance annotations and organizes samples according to the amount of historical context required to resolve each query, enabling precise and length-aware evaluation. Extensive evaluations of seven representative spoken language models reveal a general tendency toward better performance on more recent evidence: models generally achieve higher accuracy when answer-supporting evidence is closer to the query, but struggle to retrieve and use evidence located farther back in the dialogue history. Cases and datasets are available at `https://vox-infinity.github.io`.

1 INTRODUCTION

Spoken language systems (Kaplan & Haenlein, 2019; Hachman, 2019; Chu et al., 2023; 2024; Ghosh et al., 2025; Chen et al., 2024a) aim to engage in intelligent and natural interactions with humans, requiring the ability to comprehend not only semantic content but also paralinguistic cues embedded in speech. Early systems (Kaplan & Haenlein, 2019; Hachman, 2019) primarily relied on automatic speech recognition (ASR) (Yu & Deng, 2016) to convert audio into text, and achieved basic dialogue functionality through intent recognition (Liu et al., 2019; Zheng et al., 2017) and dialogue state tracking (Williams et al., 2016; Jacqmin et al., 2022) based on the transcribed content. With the emergence of large language models (LLM) (Bai et al., 2023a; Dubey et al., 2024; Fang et al., 2025; Yang et al., 2025a), some systems (SpeechTeam, 2024; Chen et al., 2025; Zhang et al., 2024a) have significantly improved in semantic understanding. Some recent approaches have integrated ASR and LLM to construct powerful multi-stage dialogue systems with strong interactive capabilities. However, as cascaded architectures, such systems (Zhang et al., 2024a; Chen et al., 2025) still struggle to fully capture and interpret the rich information present in speech, particularly acoustic features such as intonation, emotion, and emphasis. To overcome these limitations and advance audio understanding, recent studies (Tang et al., 2023; Xu et al., 2025a; Chen et al., 2024a; Ghosh et al., 2025; Zhang et al., 2024b) have proposed large audio language models that incorporate raw audio directly into the dialogue framework. This end-to-end integration allows models to better recognize and utilize nuanced acoustic signals (Ao et al., 2024; Cheng et al., 2025) that are often lost in ASR-based pipelines.

However, audio presents a significant challenge due to its low information density (Ji et al., 2024a). Spoken content typically requires 12 to 25 embeddings (Zeng et al., 2024; Ji et al., 2024b) per second to encode, resulting in much longer input sequences than text for conveying equivalent semantic information. This substantially increases the effective context length in spoken language tasks and exacerbates the difficulty of long-context modeling. Prior research (Bai et al., 2023b; Ding et al., 2025) on long-context language models has shown that performance tends to degrade as relevant

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Illustration of three representative dialogue scenarios in VOX-INFINITY. **(a) Ultra Multi-Turn Dialogue**: frequent back-and-forth exchanges requiring fine-grained information retrieval. **(b) Personal Monologue**: single-speaker long-form speech requiring reasoning over distributed evidence. **(c) Beyond-Semantic Dialogue**: extended conversations that require models to retain and retrieve emotions conveyed through vocal expression.

information moves farther back in the input, a limitation that similarly affects spoken dialogue systems. While long-context understanding has been widely explored in textual (Ding et al., 2024) and visual modalities (Chen et al., 2024b; Zhang et al., 2024b), research in the spoken language domain remains in its early stages. Most existing efforts (Goel et al., 2025) focus on extending audio input beyond 30 seconds, a constraint largely imposed by the capacity of audio encoders such as Whisper (Rouditchenko et al., 2024). However, a systematic evaluation of spoken language models under varying long-context conditions remains notably lacking.

To bridge this gap, we introduce VOX-INFINITY, the first benchmark specifically designed to evaluate long-context understanding in spoken language models across varying context lengths. Vox-Infinity systematically extends audio history along two dimensions: the number of dialogue turns and the duration of each turn. As illustrated in Figure 1, the benchmark encompasses three complementary scenarios: ultra multi-turn dialogues, personal monologues, and beyond-semantic dialogues. To enable precise evaluation, each question is annotated with the exact locations of its answer-supporting evidence in the audio history. We define the required context duration as the temporal span from the earliest supporting evidence to the query. Unlike total session duration, this quantity captures how far the evidence required to answer a query lies in the dialogue history. Vox-Infinity groups test samples according to their required context duration, enabling fine-grained analysis of model performance across different evidence distances. Using Vox-Infinity, we systematically evaluate seven representative spoken language models across diverse dialogue scenarios and required context durations. The results reveal a general tendency toward better performance on more recent evidence: models generally achieve higher accuracy when answer-supporting evidence is closer to the query, but struggle to retrieve and use evidence located farther back in the dialogue history. Vox-Infinity is available at `https://vox-infinity.github.io`. Our main contributions are as follows:

- We introduce Vox-Infinity, the first benchmark specifically designed to evaluate long-context understanding in spoken language models under varying context conditions. It systematically scales dialogue history along two axes, turn count and turn duration, and covers a diverse range of realistic long-context spoken language scenarios.
- Vox-Infinity provides explicit answer-provenance annotations and defines the required context duration of each query as the temporal span from the earliest supporting evidence to the query. This enables precise, length-sensitive evaluation across different evidence distances.
- We systematically evaluate seven representative spoken language models and reveal a general tendency toward better performance on more recent evidence. Models generally perform better

when answer-supporting evidence is closer to the query and struggle to retrieve and use evidence located farther back in the dialogue history.

2 Related Works

2.1 Spoken Dialogue Systems

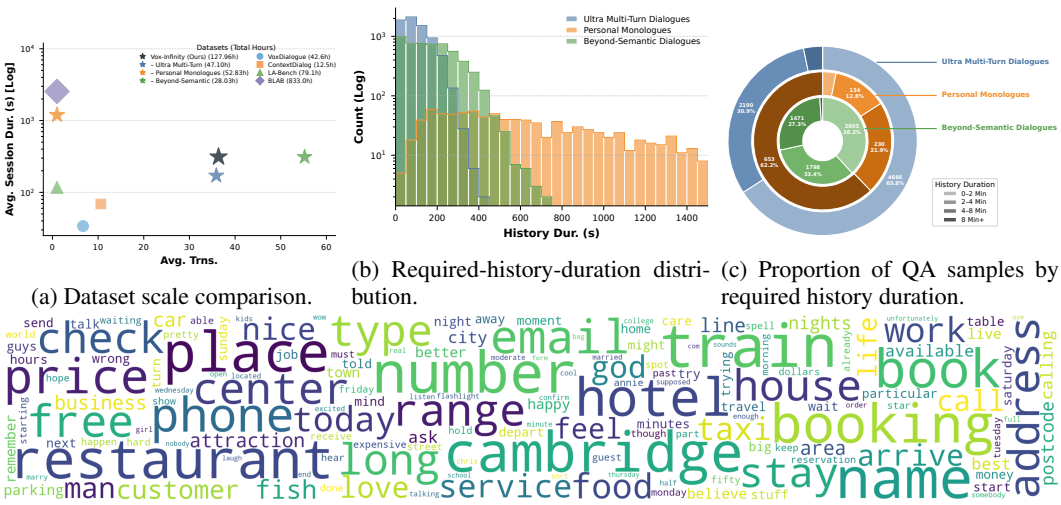
The rapid advancement of large language models (LLMs) has fueled the emergence of increasingly powerful spoken dialogue systems. A notable early milestone was SpeechGPT (Zhang et al., 2023), the first speech-centric LLM to integrate discrete speech units into a unified language modeling framework. Building on this foundation, large-scale audio language models such as Qwen-Audio 1/2 (Chu et al., 2023; 2024) extended capabilities to over 30 audio-related tasks, including speech recognition, speech translation, and audio event detection—laying the groundwork for more sophisticated spoken language understanding. These core audio modeling capabilities have since enabled the development of specialized dialogue systems. For example, StyleTalk (Lin et al., 2024) focuses on emotional dialogue, introducing the first model capable of generating speech with distinct emotional prosody. Models such as GLM-4-Voice (Zeng et al., 2024), LLaMA-Omni 1/2 (Fang et al., 2024; 2025), and the Qwen-Omni series (Xu et al., 2025a;b; Qwen Team, 2026) support end-to-end spoken interaction through unified speech understanding and generation.

Despite these advancements, most current systems remain constrained by limited context windows. Because audio is a low-compression modality, representing equivalent semantic content requires much longer sequences than text, posing serious challenges for long-context modeling. To address this challenge, the Audio Flamingo series (Ghosh et al., 2025; Goel et al., 2025; Ghosh et al., 2026) has progressively extended support for long-form audio understanding, with its latest version supporting complex audio inputs of up to 30 minutes. In parallel, models such as Slam-Omni (Chen et al., 2024a) and Step-Audio2 (Wu et al., 2025) alleviated audio compression bottlenecks by incorporating text-based history representations, which significantly improve memory efficiency and dialogue history retention. MiMo-Audio (Xiaomi, 2025) adopts a complementary approach by aggregating every four consecutive 25-Hz tokenizer frames into a 6.25-Hz representation for the LLM, thereby reducing the length of audio sequences.

2.2 Long-Context Modeling and Benchmarks

Long-context modeling (Li et al., 2024) has become a central research area in the development of LLMs, as extended inputs are essential for maintaining discourse coherence and retrieving information introduced earlier in the context. Prior work has expanded model context windows through architectural modifications, positional extrapolation, and context compression (Ding et al., 2024; Jiang et al., 2023). More recent work has further scaled context modeling through sparse and linear attention mechanisms. Qwen2.5-1M (Yang et al., 2025b) combines long-context training, length extrapolation, and sparse attention to support contexts of up to one million tokens. Kimi Linear (Team et al., 2025) introduces Kimi Delta Attention for efficient inference at the million-token scale, while Kimi K3 (Team et al., 2026) further scales this architecture with a one-million-token context window. However, a larger nominal context window does not necessarily imply that a model can reliably use information throughout the input. Prior studies have shown that model performance can vary substantially with the position of relevant evidence (Liu et al., 2024).

To systematically evaluate whether models can effectively use extended contexts, a number of long-context benchmarks have been developed. Early efforts such as LAMBADA (Paperno et al., 2016) and Long Range Arena (Tay et al., 2020) focused on controlled settings for testing long-range dependencies. More comprehensive benchmarks, including LongBench (Bai et al., 2023b) and L-Eval (An et al., 2023), expanded evaluation to summarization, question answering, reasoning, and dialogue. Subsequent benchmarks further increased context scale and task difficulty: InfiniteBench (Zhang et al., 2024c) extends evaluation beyond 100K tokens, LongBench-V2 (Bai et al., 2024) emphasizes realistic and challenging long-context tasks, and RULER (Hsieh et al., 2024) evaluates models under configurable sequence lengths and task complexities. Together, these studies show that nominal context size alone is insufficient to characterize long-context capability, as performance can vary with sequence length, task complexity, and evidence position.



(a) Dataset scale comparison.

(b) Required-history-duration distribution.

(c) Proportion of QA samples by required history duration.

(d) Word cloud visualization of common topics.

Figure 2: Statistics and characteristics of VOX-INFINITY. **(a)** Dataset-scale comparison by average turns per session and average complete session duration; marker size encodes total hours, and stars denote VOX-INFINITY and its three subsets. **(b)** Distribution of required history durations using 50-second bins and a logarithmic count axis. **(c)** Percentage of QA samples in the 0–2, 2–4, 4–8, and 8+ minute required-history-duration bins. **(d)** Word cloud of common topics.

Despite these advancements, existing evaluations remain almost exclusively text-only, overlooking the distinctive challenges of spoken language, such as acoustic cues (Ao et al., 2024), and significantly longer sequence lengths due to audio's low information density (Ji et al., 2024a). These factors make spoken language understanding a fundamentally different long-context problem. To address this gap, we propose Vox-Infinity, the first benchmark explicitly designed to evaluate long-context reasoning in spoken language models. Vox-Infinity scales audio history along two axes (turn count and duration) and incorporates a diverse set of realistic spoken scenarios. By providing detailed annotations and organizing test cases by answer provenance and context length, Vox-Infinity enables fine-grained context-aware evaluation, pushing the boundaries of long-context modeling beyond the text domain to better reflect real-world usage of spoken language systems.

3 VOX-INFINITY: LONG-CONTEXT SPOKEN LANGUAGE BENCHMARK

3.1 DATASET CONSTRUCTION

Stage 1: Long-Context Spoken Dialogue Generation. VOX-INFINITY encompasses three complementary spoken dialogue scenarios. ***(I) Ultra Multi-Turn Dialogues*** consist of conversations with more than 20 turns, constructed from SpokenWOZ (Si et al., 2024), a task-oriented dialogue dataset recorded by real speakers. Questions in this subset evaluate whether models can retrieve relevant information from dense, long-turn dialogue histories. ***(II) Personal Monologues*** feature single-speaker speeches longer than two minutes, derived from GigaSpeech (Chen et al., 2021). Based on timestamp annotations, we reconstruct short segments into their original long-form monologues. The corresponding questions require multi-hop reasoning over evidence distributed throughout the long-form speech. ***(III) Beyond-Semantic Dialogues*** evaluate whether models can retain paralinguistic emotion cues over extended dialogue histories. We construct this subset from IEMOCAP (Busso et al., 2008). After isolating the target speaker using the MossFormer2 speech separation model (Zhao et al., 2024), we associate each turn with its fine-grained emotion annotation. However, emotions in naturally occurring speech can often be inferred from lexical and semantic content alone, allowing models to answer correctly without relying on acoustic cues. To disentangle acoustic emotion recognition from text-based semantic inference, we additionally select neutral turns and use Qwen-TTS to synthesize the same textual content in different emotional styles. The synthesized turns are independently validated by Gemini and Qwen-Omni, and only samples for which both models recognize the intended emotion are retained. By combining naturally emotional

Table 1: Comparison of existing long-context benchmarks across textual and audio modalities. **Avg. Tok.** denotes the average number of tokens per session, **Avg.Dur.** denotes the average duration per session, and **Avg.Trns.** denotes the average number of turns per session. **Prov.** indicates whether answer provenance annotations are available. **Length distribution** refers to the context-length ranges used for evaluation; ‡ for datasets without explicit definitions, we adopt the near-upper quartile of context length as the upper bound. For audio benchmarks, average duration is calculated as total duration divided by the number of sessions; † token counts are estimated assuming 25 tokens per second.

Benchmarks	Session	QAs	Dur.(h)	Avg. Tok.†	Avg. Dur.(s)	Avg. Trns.	Prov.	Length Distribution‡
Long Context Benchmark (Textual)								
LongBench (Bai et al., 2023b)	4,550	4,550	-	8.1k	-	1.00	✗	0-4k, 4k-8k, 8k+
∞ BENCH (Zhang et al., 2024c)	3,946	3,946	-	200k	-	1.00	✗	0-200k
LongBench-V2 (Bai et al., 2024)	503	503	-	54k	-	1.00	✗	0-32k, 32k-128k, 128k+
Long Context Benchmark (Audio)								
VoxDialogue (Cheng et al., 2025)	4,526	4,526	42.56	0.8k	33.9	6.70	✗	0-2min
ContextDialog (Kim et al., 2025)	653	2,612	12.5	1.7k	68.9	10.60	✗	0-2min
LA-Bench (Kong et al., 2024)	2,429	2,429	79.12	2.9k	117.3	1.00	✗	0-3min+
BLAB (Ahia et al., 2025)	1,176	1,600	833	63.8k	2,550.0	1.00	✗	0-60min+
Vox-Infinity (ours)	1,474	13,516	127.96	7.8k	312.5	36.38	✓	0-2min, 2-4min, 4-8min, 8min+
- *Ultra Multi-Turn Dialogue*	991	7,086	47.10	4.3k	171.1	35.89	✓	0-2min, 2-4min, 4-8min
- *Personal Monologues*	159	1,050	52.83	29.9k	1,196.1	1.00	✓	0-2min, 2-4min, 4-8min, 8min+
- *Beyond-Semantic Dialogue*	324	5,380	28.03	7.8k	311.4	55.23	✓	0-2min, 2-4min, 4-8min, 8min+

speech with semantically controlled synthetic speech, this subset evaluates whether models can retrieve and recognize emotional cues expressed in distant dialogue turns.

Stage 2: Question–Answer Pair Generation. In the second stage, we generate multiple question-answer (QA) pairs for each session to evaluate fine-grained retrieval, multi-hop reasoning, and paralinguistic memory across varying required context durations. Since model responses may evolve as the dialogue progresses, we employ GPT-4 to generate diverse questions based on the whole dialogue to ensure each yields a unique and unambiguous answer. To precisely control the required context, GPT-4 is also instructed to specify which dialogue turns must be referenced when answering each question. Detailed prompt design and generation strategies are provided in the appendix B.1.

Stage 3: Data Verification and Quality Control. Finally, to ensure benchmark quality, each QA pair is manually examined to verify that its answer is correct and unique within the corresponding dialogue context. The associated answer-provenance annotation is further checked to confirm that the annotated turns provide the supporting evidence required to answer the question. Samples with ambiguous answers, factual inconsistencies, or inaccurate provenance annotations are regenerated and subjected to another round of manual verification. This process yields a reliable long-context spoken language benchmark covering diverse scenarios and capabilities.

3.2 Dataset Statistics

Detailed Statistics of Vox-Infinity. To provide a comprehensive overview of Vox-Infinity, Figure 2 presents detailed dataset statistics. As shown in Figure 2a, we compare the three subsets of Vox-Infinity with existing long-audio benchmarks in terms of average turns per session and average complete session duration. Across its three subsets, Vox-Infinity contains 1,474 sessions, 13,516 QA pairs, and 127.96 hours of audio, with an average session duration of 312.5 seconds. Beyond-Semantic contains 324 sessions and 5,380 QA pairs, with an average of 55.23 turns and 311.4 seconds per session. Together, the three subsets cover complementary regimes ranging from dense multi-turn interactions to long single-speaker monologues. Figure 2b illustrates the distribution of required history durations across the three subsets. Figure 2c summarizes the same QA-level distributions using the evaluation bins adopted throughout the paper. Ultra Multi-Turn is concentrated below four minutes, Personal Monologues contain predominantly eight-minute-plus dependencies, and Beyond-Semantic spans all four required-history-duration ranges. Finally, Figure 2d presents a word cloud of common topics in Vox-Infinity. The dataset features a high concentration of queries targeting key contextual information—such as phone numbers, train IDs, and restaurants—all of

which have explicit answers, making them ideal for assessing a model's ability to retain and utilize long-term contextual memory. Beyond-Semantic additionally introduces long-range paralinguistic emotion cues, emphasizing challenges that cannot be captured by text-only dialogue histories.

Comparison with Existing Benchmarks. Table 1 situates Vox-Infinity within the landscape of long-context benchmarks across text and audio modalities. Text-based benchmarks such as LongBench (Bai et al., 2023b; 2024) enable performance evaluation under varying context lengths. In contrast, existing audio benchmarks (Goel et al., 2025) primarily aim to overcome the 30-second input limit of encoders like Whisper (Radford et al., 2023), focusing on extending input length rather than analyzing performance across different context regimes. Vox-Infinity fills this gap by being the first audio benchmark to explicitly evaluate models under multiple context-length conditions, allowing fine-grained analysis of long-context reasoning. Compared with benchmarks like BLAB (Ahia et al., 2025), whose recordings span 15 minutes to 2 hours, Vox-Infinity adopts more practical dialogue durations while preserving contextual diversity. This makes it better suited for realistic and scalable spoken dialogue evaluation.

Another limitation of prior QA-based benchmarks (Bai et al., 2024; Goel et al., 2025) is the absence of answer provenance annotations, which makes the true amount of context required to answer a question ambiguous. Vox-Infinity addresses this issue by explicitly labeling provenance, enabling precise measurement of the actual context length needed to support each answer. By jointly considering context-length granularity and answer provenance, Vox-Infinity establishes a benchmark that is more analytically rigorous for long-context modeling.

3.3 Evaluation Suite

Following the evaluation suite of LongAudioBench (Ghosh et al., 2025), we adopt Accuracy as the primary evaluation metric, which measures the proportion of correctly answered samples among all test cases. Given the high diversity of responses generated by Spoken Dialogue Systems, we follow prior work (Duan et al., 2024; Ghosh et al., 2025) and employ GPT-4o (OpenAI, 2024) to assess the model outputs. Responses are categorized into three labels: Correct, Incorrect, and Unknown. The Unknown label allows GPT-4o to indicate that a response cannot be reliably judged as correct or incorrect, avoiding forced binary decisions in ambiguous cases. Detailed prompt templates for this evaluation process are provided in Appendix B.2.

4 Benchmarking Long-Context Spoken Language Models

4.1 Benchmarking SLMs Across Context Lengths

4.1.1 Comparison Models.

We evaluate seven representative spoken language models: Qwen3-Omni, GLM-4-Voice (Zeng et al., 2024), AudioFlamingo3 (Goel et al., 2025), AudioFlamingo Next, MiMo-Audio (Xiaomi, 2025), Gemini-3.1-Pro, and Qwen3.5-Plus. Each model is evaluated with dialogue history represented entirely as text or audio. Text histories are transcribed using the open-source SenseVoice model (SpeechTeam, 2024). For locally deployed models with a configured positional limit below 32,768, we increase `max_position_embeddings` to 32,768 to accommodate longer audio-history sequences. All dialogue histories are provided in full, without input truncation.

4.1.2 Main Results

Table 2 compares model performance across the three subsets of Vox-Infinity. An overall recency trend is observed in Ultra Multi-Turn and Beyond-Semantic Dialogues, although its strength varies across models. Figure 3 provides a finer-grained view of audio-history performance across 30-second intervals of required context duration, revealing local fluctuations rather than strictly monotonic declines. In Table 2, the average audio-history accuracy across all seven models decreases from 48.9% to 43.2% on Ultra Multi-Turn Dialogues between the first two minutes and the four-to-eight-minute range. Over the same ranges, Beyond-Semantic accuracy falls from 46.3% to 39.7%, reaching 23.7% beyond eight minutes. These results are consistent with difficulties in distant fine-grained semantic information retrieval and paralinguistic memory, with a larger observed decline

Table 2: Accuracy of spoken language models across dialogue scenarios and required-context-duration bins. *Beyond-Semantic* uses multiple-choice questions, whereas *Ultra Multi-Turn* and *Personal Monologues* use open-ended questions. Here, *m* denotes minutes and *Avg.* denotes the overall question-level accuracy within each subset, weighted by the number of QA samples in each duration bin. All values are percentages, and † marks closed-source models.

Model	*Ultra Multi-Turn*				*Personal Monologues*					*Beyond-Semantic*				
	0–2m	*2–4m*	*4–8m*	*Avg.*	*0–2m*	*2–4m*	*4–8m*	*8m+*	*Avg.*	*0–2m*	*2–4m*	*4–8m*	*8m+*	*Avg.*
Text-as-history														
Qwen3-Omni	85.5	84.2	77.0	84.8	84.8	86.6	84.8	85.5	85.5	49.5	46.2	42.4	25.0	46.2
GLM-4-Voice	61.4	56.3	48.3	59.4	27.3	36.6	26.1	26.3	27.6	43.5	42.4	37.0	10.7	41.0
AudioFlamingo3	45.6	40.7	23.0	43.4	42.4	58.2	47.8	47.8	49.0	20.0	20.1	15.0	8.9	18.5
AudioFlamingo Next	76.1	73.2	67.0	74.9	45.5	62.7	63.0	62.6	62.2	46.7	40.6	35.1	16.1	41.2
MiMo-Audio	76.1	78.1	70.9	76.6	63.6	79.9	74.3	63.1	67.7	32.5	30.2	26.8	16.1	30.0
†Gemini-3.1-Pro	80.3	77.6	66.1	79.0	45.5	81.3	80.9	85.3	82.6	58.2	54.2	46.4	19.6	53.2
†Qwen3.5-Plus	82.2	79.7	71.3	81.1	57.6	81.3	84.3	86.4	84.4	57.0	52.9	48.4	17.9	52.9
Audio-as-history														
Qwen3-Omni	56.0	51.9	66.5	55.1	51.5	76.1	73.0	69.5	70.5	56.6	53.1	49.4	32.1	53.2
GLM-4-Voice	36.9	31.5	23.5	34.8	12.1	20.1	19.1	9.8	13.2	24.1	23.2	19.2	8.9	22.3
AudioFlamingo3	35.4	34.3	29.1	34.9	60.6	64.9	49.1	37.1	44.0	26.6	26.8	23.0	12.5	25.5
AudioFlamingo Next	68.9	63.8	61.7	67.1	45.5	51.5	49.1	42.9	45.4	50.9	46.7	42.2	19.6	46.8
MiMo-Audio	24.6	19.4	13.9	22.6	24.2	6.0	22.2	10.9	13.2	42.7	38.4	37.6	21.4	39.6
†Gemini-3.1-Pro	63.2	62.6	54.8	62.8	33.3	62.7	60.4	60.8	60.1	62.7	57.7	51.0	32.1	57.5
†Qwen3.5-Plus	57.2	56.6	53.0	56.8	39.4	64.9	64.8	65.7	64.6	60.8	57.8	55.3	39.3	58.1

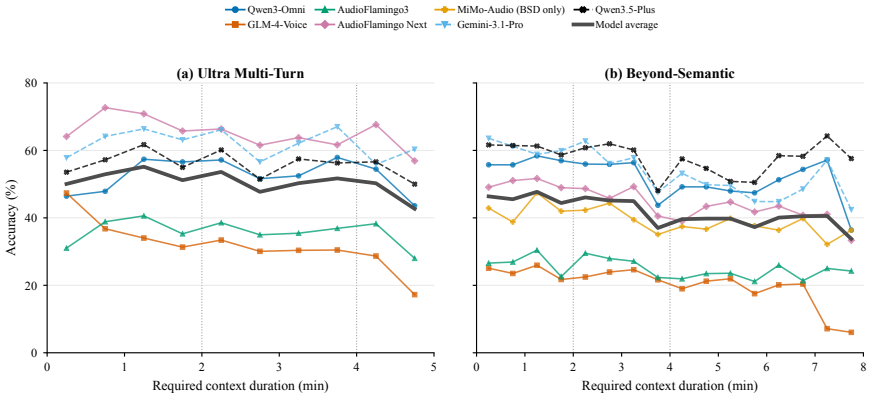


Figure 3: Fine-grained accuracy under audio history across 30-second required-context-duration intervals on **(a)** Ultra Multi-Turn and **(b)** Beyond-Semantic Dialogues. Thin curves represent individual models, while thick gray curves show their mean within each interval. Intervals containing fewer than 20 examples are omitted, and MiMo-Audio is included only in **(b)**. Despite local fluctuations, both subsets exhibit an overall recency effect.

on Beyond-Semantic Dialogues over the shared duration ranges. Personal Monologues show a less consistent pattern across duration ranges, where differences in question difficulty may also affect accuracy. In particular, the first two-minute interval contains fewer than 40 QA pairs, making its accuracy estimates more sensitive to sampling variability and question composition.

Across tasks that primarily depend on linguistic content, models generally perform better with text history, with the clearest advantages observed on Ultra Multi-Turn Dialogues and Personal Monologues. This pattern suggests that compact transcripts support the retention and use of long-range semantic information more effectively than substantially longer audio sequences. The gap is especially pronounced on Ultra Multi-Turn Dialogues, where questions often require recovering exact slot values, such as phone numbers and train identifiers, from densely interleaved turns. Text history outperforms audio history for every evaluated model on this subset, with mean accuracies of 71.3% and 47.7%, respectively. This substantial difference indicates that current spoken language models struggle to retain and recover precise lexical evidence from long audio histories.

However, this advantage does not extend to queries that depend on paralinguistic evidence. Transcripts preserve linguistic content but discard prosodic and vocal cues, including intonation, rhythm, and intensity. Accordingly, audio history outperforms text history for six of the seven evaluated models on Beyond-Semantic Dialogues. For example, MiMo-Audio achieves 30.0% with text his-

tory and 39.6% with audio history. These results reveal complementary strengths between the two history representations. Text history more effectively supports the retention and retrieval of long-range semantic information, whereas audio history retains paralinguistic evidence that is unavailable in transcripts. Taken together, the three subsets expose distinct limitations in fine-grained retrieval, multi-hop reasoning, and paralinguistic memory, demonstrating the diagnostic value of Vox-Infinity beyond a single aggregate score.

4.1.3 Disentangling Acoustic Emotion from Semantic Cues

To distinguish acoustic emotion recognition from semantic inference, we divide Beyond-Semantic Dialogues into natural and semantically controlled synthetic samples. Natural samples retain the original IEMOCAP recordings, in which emotion may be conveyed by both lexical content and vocal expression. Controlled samples instead use semantically neutral utterances synthesized in designated emotional styles, making the intended emotion difficult to infer from text alone. Table 3 compares text and audio histories under both settings.

Table 3: Accuracy on natural ($n = 2{,}758$) and semantically controlled ($n = 2{,}622$) samples from Beyond-Semantic Dialogues. Text and Audio denote the history modality, and Δ reports Audio minus Text in percentage points. All values are percentages.

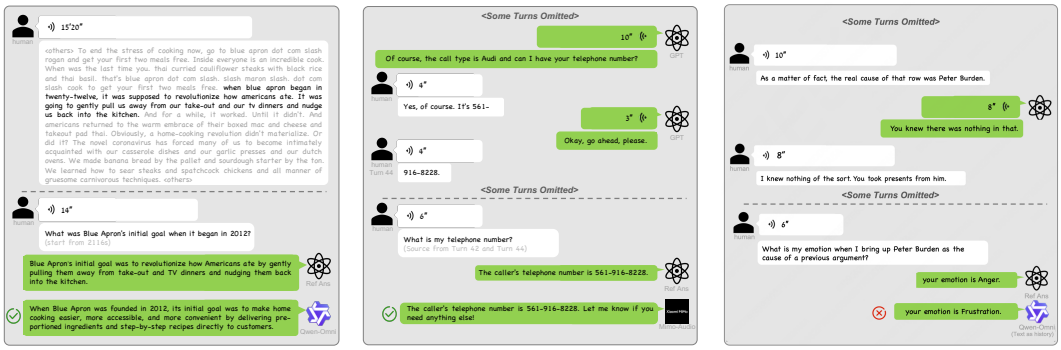
	Natural			*Semantically Controlled*		
Model	**Text**	**Audio**	**Δ**	**Text**	**Audio**	**Δ**
Qwen3-Omni	65.3	68.3	+3.0	26.1	37.3	+11.2
GLM-4-Voice	52.9	28.4	-24.6	28.5	16.0	-12.5
AudioFlamingo3	28.9	35.1	+6.1	7.6	15.4	+7.9
AudioFlamingo Next	57.8	57.9	+0.1	23.6	34.9	+11.3
MiMo-Audio	41.4	53.8	+12.4	18.0	26.0	+8.1
†Gemini-3.1-Pro	74.0	74.9	+0.9	31.4	39.2	+7.9
†Qwen3.5-Plus	73.6	71.4	-2.2	31.2	44.1	+12.9

For most models, audio history outperforms text history, and this advantage becomes substantially clearer when semantic cues are controlled. On natural samples, audio history performs better for five of the seven models, but the model-averaged accuracies of audio and text histories are nearly identical at 55.7% and 56.3%, respectively. This small gap is consistent with many natural utterances containing lexical evidence of emotion, which text history can preserve and retrieve effectively over long contexts. On controlled samples, audio history outperforms text history for six of the seven models and raises model-averaged accuracy from 23.8% to 30.4%, a gain of 6.6 percentage points. Because their lexical content is emotionally neutral, these samples require models to rely more heavily on prosody and vocal intensity, such as the raised volume and sharp intonation associated with anger, to retain and recover the intended emotion. Since the controlled samples contain synthesized speech, we interpret within-group modality gaps rather than cross-group differences in absolute accuracy.

GLM-4-Voice is the main exception, with text history outperforming audio history in both groups. A likely reason is that its discrete speech tokenizer and semantically oriented training pipeline prioritize linguistic content while retaining limited emotion-related prosody (Zeng et al., 2024). This exception highlights that the benefit of audio history depends on whether the underlying speech representation preserves paralinguistic information.

4.1.4 Human Validation of LLM-Based Evaluation

To assess LLM-judge reliability, we randomly sample 300 audio-history outputs from Qwen3-Omni and Audio Flamingo Next, with 50 examples per subset per model. Beyond-Semantic samples are drawn from the combined natural and semantically controlled subsets. Sampling is without replacement, and all outputs undergo human review. Table 4 reports model accuracies and label agreement with human judgments. Sample-level accuracies may differ from the full-benchmark results in Table 2.



(a) Personal Monologues. (b) Ultra Multi-Turn Dialogues. (c) Beyond-Semantic Dialogues.

Figure 4: Representative examples drawn from the three subsets of VOX-INFINITY.

Table 4: Agreement between LLM-based evaluation and human judgments on 300 randomly sampled audio-history outputs. Each subset cell contains 50 examples. Agreement denotes the percentage of matching labels.

Method	Ultra Multi-Turn	Personal Monologues	Beyond-Semantic	Overall
Qwen3-Omni				
LLM-eval	48.0	74.0	54.0	58.7
Human-eval	46.0	76.0	54.0	58.7
Agreement	98.0	98.0	100.0	98.7
Audio Flamingo Next				
LLM-eval	68.0	50.0	52.0	56.7
Human-eval	68.0	52.0	52.0	57.3
Agreement	100.0	98.0	100	99.3

The LLM judge agrees with human judgments on 297 of the 300 sampled outputs, corresponding to an overall agreement of 99.0%. All Beyond-Semantic labels agree for both models. The three discrepancies are isolated: two false negatives on Personal Monologues, one for each model, and one false positive on Ultra Multi-Turn for Qwen3-Omni. We observe no systematic subset- or model-specific disagreement in the audited sample.

4.2 LONG-CONTEXT SPOKEN DIALOGUE CASES IN VOX-INFINITY

To illustrate Vox-Infinity, Figure 4 presents one example from each subset. In each case, the light-gray region within the question indicates the answer's location, thereby specifying the amount of historical context required. The examples illustrate the distinct evidence patterns targeted by the three subsets. In particular, the Beyond-Semantic example in Figure 4c shows why text-only histories cannot preserve paralinguistic cues. Additional examples are available on our demo page.

5 CONCLUSION

In this work, we present Vox-Infinity, a benchmark for evaluating long-context understanding in spoken language models. Vox-Infinity extends spoken histories along turn count and turn duration and provides answer-provenance annotations for evidence-distance-aware evaluation across fine-grained retrieval, multi-hop reasoning, and paralinguistic memory. Evaluations of seven representative models reveal an overall recency effect, with model performance generally declining as answer-supporting evidence appears farther from the query. The results further show that fine-grained semantic evidence is difficult to retrieve from long audio histories, while text histories preserve lexical information more effectively but lose speech-specific paralinguistic cues. These findings establish Vox-Infinity as a diagnostic testbed for studying how spoken language models retain and use semantic and acoustic information over extended interactions.

Ethical Considerations

Vox-Infinity was developed with careful attention to ethical and responsible research practices. All speech data used in this benchmark are publicly available, and no personally identifiable information is included. To further protect speaker privacy, the data are anonymized and used exclusively for research purposes.

We recognize that spoken dialogue corpora may reflect social, cultural, or gender biases, which could be inherited by models trained or evaluated on them. While Vox-Infinity is intended to advance research on long-context understanding, it should not be regarded as free of bias. We encourage future researchers to critically examine and mitigate potential harms that may arise from evaluation on this benchmark.

Finally, as large language and speech models become increasingly powerful, they also present risks of misuse, such as generating misleading content or enabling surveillance. Vox-Infinity is therefore released solely for academic and responsible industrial research, with the goal of fostering dialogue systems that are beneficial, transparent, and aligned with human values.

Reproducibility Statement

We are committed to ensuring the reproducibility of our work. A portion of the dataset is already publicly available on our demo page, and the full release will be made available after paper acceptance. We provide detailed descriptions of the dataset construction process, evaluation protocols, and experimental settings in the main text and Appendix. In addition, the evaluation tools required for benchmarking will also be released upon acceptance to facilitate replication and future research.

References

Orevaoghene Ahia, Martijn Bartelds, Kabir Ahuja, Hila Gonen, Valentin Hofmann, Siddhant Arora, Shuyue Stella Li, Vishal Puttagunta, Mofetoluwa Adeyemi, Charishma Buchireddy, et al. Blab: Brutally long audio bench. *arXiv preprint arXiv:2505.03054*, 2025.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*, 2023.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *arXiv preprint arXiv:2406.13340*, 2024.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023b.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.

Junjie Chen, Yao Hu, Junjie Li, Kangyue Li, Kun Liu, Wenpeng Li, Xu Li, Ziyuan Li, Feiyu Shen, Xu Tang, et al. Fireredchat: A pluggable, full-duplex voice interaction system with cascaded and semi-cascaded implementations. *arXiv preprint arXiv:2509.06502*, 2025.

Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, et al. Slam-omni: Timbre-controllable voice interaction system with single-stage training. *arXiv preprint arXiv:2412.15649*, 2024a.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024b.

Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, Shengpeng Ji, Rongjie Huang, Boyang Zhang, Tao Jin, et al. Voxdialogue: Can spoken dialogue systems understand information beyond words? In *The Thirteenth International Conference on Learning Representations*, 2025.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*, 2024.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. *arXiv preprint arXiv:2505.02625*, 2025.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*, 2025.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music, 2026. URL https://arxiv.org/abs/2604.10905.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

Mark Hachman. The microsoft-amazon deal leaves cortana speakers with one advantage: Skype. *PCWorld Retrieved October*, 23, 2019.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Léo Jacqmin, Lina M Rojas-Barahona, and Benoit Favre. " do you follow me?": A survey of recent approaches in dialogue state tracking. *arXiv preprint arXiv:2207.14627*, 2022.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*, 2024a.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024b.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*, 2023.

Andreas Kaplan and Michael Haenlein. Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business horizons*, 62(1): 15–25, 2019.

Heeseung Kim, Che Hyun Lee, Sangkwon Park, Jiheum Yeom, Nohil Park, Sangwon Yu, and Sungroh Yoon. Does your voice assistant remember? analyzing conversational context recall and utilization in voice interaction models. *arXiv preprint arXiv:2502.19759*, 2025.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*, 2024.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. *arXiv preprint arXiv:2402.12786*, 2024.

Jiao Liu, Yanling Li, and Min Lin. Review of intent detection methods in the human-machine dialogue system. In *Journal of physics: conference series*, volume 1267, pp. 012059. IOP Publishing, 2019.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

OpenAI. Gpt-4o system card. *https://cdn.openai.com/gpt-4o-system-card.pdf*, 2024.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. Whisper-flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation. *arXiv preprint arXiv:2406.10082*, 2024.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *Advances in Neural Information Processing Systems*, 36, 2024.

Tongyi SpeechTeam. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*, 2024.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, Wentao Li, Enzhe Lu, Weizhou Liu, Yanru Chen, Weixin Xu, Longhui Yu, Yejie Wang, Yu Fan, Longguang Zhong, Enming Yuan, Dehao Zhang, Yizhi Zhang, T. Y. Liu, Haiming Wang, Shengjun Fang, Weiran He, Shaowei Liu, Yiwei Li, Jianlin Su, Jiezhong Qiu, Bo Pang, Junjie Yan, Zhejun Jiang, Weixiao Huang, Bohong Yin, Jiacheng You, Chu Wei, Zhengtao Wang, Chao Hong, Yutian Chen, Guanduo Chen, Yucheng Wang, Huabin Zheng, Feng Wang, Yibo Liu, Mengnan Dong, Zheng Zhang, Siyuan Pan, Wenhao Wu, Yuhao Wu, Longyu Guan, Jiawen Tao, Guohong Fu, Xinran Xu, Yuzhi Wang, Guokun Lai, Yuxin Wu, Xinyu Zhou, Zhilin Yang, and Yulun Du. Kimi linear: An expressive, efficient attention architecture, 2025. URL https://arxiv.org/abs/2510.26692.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guangyu Chen, Guanzheng Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kexin Chen, Peng Chen, Ruijue Chen, Wentao Chen, Xin Chen, Yang Chen, Yanru Chen, Yifei Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Dazhi Cheng, Yean Cheng, Jialei Cui, Jingbing Cui, Anqi Dai, Jiaqi Deng, Hao Ding, Rui Ding, Shaofeng Ding, Mengfan Dong, Mengnan Dong, Yuhao Dong, Yuxin Dong, Angang Du, Chenzhuang Du, Dikang Du, Jusen Du, Yulun Du, Yu Fan, Jing Feng, Qiulin Feng, Yichen Feng, Kelin Fu, Qiang Fu, Fuxuan Gao, Hongcheng Gao, Jingyue Gao, Tong Gao, Weijia Gao, Shangyi Geng, Jie Gong, Linhu Gong, Shengao Gong, Xiaochen Gong, Qizheng Gu, Yicheng Gu, Shuhao Guan, Haiqing Guo, Shiqi Guo, Xiang Guo, Zhengyan Guo, Beixi Hao, Wenxin Hao, Xiaoru Hao, Dailan He, Haotian He, Lehan He, Qi He, Weiran He, Xinran He, Xinyi He, Yibo He, Yunjia He, Chao Hong, Tiange Hong, Hao Hu, Jiaxi Hu, Ruikun Hu, Weiming Hu, Yangyang Hu, Zhenxing Hu, Liang Hua, Jinbin Huang, Ke Huang, Ruiyuan Huang, Siying Huang, Weixiao Huang, Yan Huang, Zhengjie Huang, Zhiqi Huang, Yulong Hui, Chaobo Jia, Yutong Jiang, Zhejun Jiang, Zuoyou Jiang, Wenyi Jin, Xinyi Jin, Yu Jing, Huanjun Kong, Guokun Lai, Aidi Li, Cheng Li, Chengyuan Li, Cong Li, Fang Li, Guanyu Li, Haoyang Li, Jia Li, Junxiong Li, Lei Li, Letian Li, Lincan Li, Weihong Li, Wentao Li, Xintong Li, Yang Li, Yishen Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxi Li, Zheming Li, Zhengxiao Li, Zhiyuan Li, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zichao Lin, Ziyan Lin, Bill Liu, Boxiao Liu, Chuan Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yipeng Liu, Zhengying Liu, Zhiheng Liu, Enzhe Lu, Haoyu Lu, Linqiang Lu, Tingzhan Lu, Zhiyuan Lu, Aotian Luo, G. Luo, Junyu Luo, Yifan Luo, B. Lyu, Wenzhou Lyu, Shaoguang Mao, Yuan Mei, Xin Men, Minqing Ni, Yixuan Niu, Siyuan Pan, Shujun Peng, Zhangyang Qi, Ruoyu Qin, ZeChao Qin, Zeyu Qin, Haiquan Qiu, Jianxin Qiu, Jiezhong Qiu, Bowen Qu, Yuhao Qu, Zeyu Shang, Youbo Shao, Han Shen, Jincheng Shi, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Wingchun Siu, Pengwei Song, Xiaoxi Song, Jianlin Su, Yunfeng Su, Zhaochen Su, Lin Sui, Jingsong Sun, Junyao Sun, Shaoning Sun, Shuzhe Sun, Tongyu Sun, Yujun Sun, Yunpeng Tai, Chuning Tang, Heyi Tang, Sirui Tang, Zecheng Tang, Chaoran Tian, Rongpeng Tian, Yu Tian, Wei Tu, Chensi Wang, Chuang Wang, Chunjie Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hao Wang, Huaqing Wang, Hui Wang, Jiayi Wang, Jinglong Wang, Jinhong Wang, Jiuzheng Wang, Linian Wang, Shaobo Wang, Shenzhi Wang, Shuyi Wang, Si Wang, Siyuan Wang, Tianfu Wang, Wenjue Wang, Xingran Wang, Xinmei Wang, Xinyuan Wang, Xusheng Wang, Yalin Wang, Yangkun Wang, Yao Wang, Yaoyu Wang, Yejie Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang,

Zhengtao Wang, Zhenhao Wang, Zhongsheng Wang, Zifan Wang, Chu Wei, Ming Wei, Shouxin Wei, Zichen Wen, Fan Wu, Haoning Wu, Rucong Wu, Wenhao Wu, Xiaoxue Wu, Yingcong Wu, Yongqi Wu, Yuxin Wu, Zijian Wu, Xinglang Xian, Chenxuan Xiang, Yuye Xiang, Bocheng Xiao, Chenjun Xiao, Xin Xiao, Jin Xie, Xiaotong Xie, Yifeng Xie, Zhe Xie, Bowei Xing, Yiming Xiong, Baosheng Xu, Boyu Xu, Jiale Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Qingtao Xu, Shuyao Xu, Suting Xu, Tiantian Xu, Tianxiang Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ye Xu, Yueni Xu, Ziyao Xu, Haonan Xue, Junjie Yan, Yaoyao Yan, Fan Yang, Guangyao Yang, Hao Yang, Junwei Yang, Ruoyu Yang, Wenjie Yang, Xiaofei Yang, Xinyu Yang, Yi Yang, Yiling Yang, Ying Yang, Yuchen Yang, Zhen Yang, Zhilin Yang, Zian Yang, Zuhao Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhanbo Ye, Bohong Yin, Haoxiang Yin, Xietong Yin, Chengzhen Yu, Haozhen Yu, Longhui Yu, Shengnan Yu, Shuying Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Tongtian Yue, Wei Yue, Yang Yue, Dunyuan Zha, Haobing Zhan, B. H. Zhang, Dehao Zhang, Fei Zhang, Hao Zhang, Haoyuan Zhang, Huanyu Zhang, Jiapei Zhang, Jiaxuan Zhang, Jin Zhang, Kaiyi Zhang, Miaozhen Zhang, Puqi Zhang, Qinglei Zhang, Rong Zhang, Rui Zhang, Shaoshuai Zhang, Shiyi Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yangkun Zhang, Ye Zhang, Yichi Zhang, Yikun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Zijing Zhang, Bin Zhao, Chenguang Zhao, Feifan Zhao, Jinglun Zhao, Jinxiang Zhao, Shuai Zhao, Wenshuo Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Haozhi Zheng, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Haofeng Zhong, Lei Zhong, Longguang Zhong, M. Zhou, Qiankang Zhou, Runjie Zhou, Ruozhang Zhou, Xinyu Zhou, Yiqiao Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yangjunfeng Zhu, Yuxuan Zhu, Zhen Zhu, Chen Zhuang, Weiyu Zhuang, and Xinxing Zu. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Jason D Williams, Antoine Raux, and Matthew Henderson. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33, 2016.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

LLM-Core-Team Xiaomi. Mimo-audio: Audio language models are few-shot learners, 2025. URL https://github.com/XiaomiMiMo/MiMo-Audio.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. Qwen2.5-1m technical report, 2025b. URL https://arxiv.org/abs/2501.15383.

Dong Yu and Lin Deng. *Automatic speech recognition*, volume 1. Springer, 2016.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023.

Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, Qipeng Guo, Haodong Duan, Xin Chen, Han Lv, Zheng Nie, Min Zhang, Bin Wang, Wenwei Zhang, Xinyue Zhang, Jiaye Ge, Wei Li, Jingwen Li, Zhongying Tu, Conghui He, Xingcheng Zhang, Kai Chen, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2.5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions. *arXiv preprint arXiv:2412.09596*, 2024a.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024b.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. ∞ bench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*, 2024c.

Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma. MossFormer2: Combining transformer and RNN-free recurrent network for enhanced time-domain monaural speech separation. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10356–10360, 2024. doi: 10.1109/ICASSP48485.2024.10445985.

Yang Zheng, Yongkang Liu, and John HL Hansen. Intent detection and semantic parsing for navigation dialogue language processing. In *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)*, pp. 1–6. IEEE, 2017.

A USE OF LLM

In this work, we employed large language models (LLMs) both for generating QA data and for conducting model evaluation. Moreover, as Vox-Infinity is proposed as a benchmark, we also report the performance of several LLMs on our dataset to provide reference baselines.

B PROMPT TEMPLATE

B.1 PROMPT TEMPLATE FOR QA GENERATION.

We show the prompt template for QA generation in Figure 5. In addition to generating question–answer pairs, the process also records the source locations of the information needed for answering, ensuring traceability.

B.2 EVALUATION PROMPT TEMPLATES

Figures 6 and 7 illustrate the standardized prompts used to evaluate multiple-choice and open-ended questions, respectively.

LIMITATION

Vox-Infinity focuses on retaining and using evidence within a self-contained spoken session. It does not yet evaluate persistent long-horizon interactions in which relevant information is updated, contradicted, or revisited across multiple sessions. Such settings require models not only to retrieve distant evidence, but also to distinguish current information from superseded content and maintain coherent memory as an interaction evolves. Extending Vox-Infinity to dynamic and cross-session spoken interactions would complement its current evaluation of fine-grained retrieval, multi-hop reasoning, and paralinguistic memory.

You are a researcher studying large speech models and their ability to **retain and recall information from historical voice dialogues**. Your goal is to evaluate whether a model can answer user questions based on previous dialogues after many conversational turns. To do this:

- You will receive **transcriptions** of several audio samples.
- Each transcription contains the spoken content (speech only).
- Use the transcriptions to reconstruct the conversation and gather information.

Example input format:

```
{
  "Audio 0": { "Transcription": "hi , my name is john . i'm looking for info in cambridge ." },
  ...
  "Audio 39": { "Transcription": "okay , thank you so much for calling . goodbye ." }
}
```

Your task: Generate questions and answers based on the dialogue.
Requirements:

1. Questions must relate to the audio content.
2. Answers should be **short, clear, and concise** for easy evaluation.
3. Questions should require **reasoning, world knowledge, and information extraction**, not trivial restatement.
4. Do **not** refer to the transcription explicitly; assume the model only "hears" the audio.
5. Each answer should reflect the **final resolved information** (not partial mid-conversation states).
6. Avoid ambiguity: every question should have a **single, clear answer**.
7. Add a "Source" field citing the audio(s) that provide the answer.
8. Context length control:
    - First question comes from the most recent stage, the second from the next, and so on.
    - If an answer spans multiple audios, ask at the stage where the information is complete.
    - As turns increase, "Source" audios should progressively move earlier.

Output format:

```
{
  "QA-Pair 1": {
    "Question": "What are the open hours of the theatre?",
    "Answer": "The theatre is open all the time.",
    "Source": { "Audio 35": "okay , actually they are open all the time ." }
  }
  ...
}
```

Figure 5: Prompt template for QA generation.

Prompt Template

You are an AI assistant tasked with judging the correctness of a model's response to a multiple-choice question. You are given the question, its answer options, the correct answer, and the model's output. Your evaluation must follow these rules:

- If the model's output matches the correct answer, output "Correct".
- If the model's output matches one of the options but not the correct answer, output "Incorrect".
- If the model's output does not match any of the options, output "Unknown".

Your output must be a single word chosen strictly from: {Correct, Incorrect, Unknown}.

Question: {question}
Options: {options}
Correct Answer: {correct_answer}
Model Output: {model_output}

Figure 6: Standardized prompt template used for multiple-choice QA evaluation.

Prompt Template

You are evaluating a model's response to an open-ended question. You will receive the question, a reference answer, and the model's response. Judge whether the response conveys the same essential answer as the reference.

- Output "Correct" if the response gives the correct essential answer. Accept paraphrases and harmless differences in formatting. For names, identifiers, dates, and numbers, accept formatting variations only when the underlying value is unchanged.
- Output "Incorrect" if the response gives a wrong or contradictory answer, omits an essential part, does not answer the question, or says it does not know.
- Output "Unknown" only if the question or reference answer is too ambiguous or incomplete to determine correctness reliably. Do not use "Unknown" merely because the model's wording differs from the reference.
- Do not mark a response correct based only on keyword overlap. Do not require the response to reproduce the reference wording verbatim.
- Output exactly one label: Correct, Incorrect, or Unknown.

Question: {question}
Reference Answer: {correct_answer}
Model Response: {model_output}

Figure 7: Standardized prompt template used for open-ended QA evaluation.