BENCHMARKING AUDIO-LANGUAGE MODELS ON REAL-WORLD LONG-FORM UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large audio language models (LALMs) have made remarkable progress at the second- and minute-level scale, yet understanding hour-level audio remains a fundamental bottleneck. Existing benchmarks rely on short clips or artificially concatenated segments, and so fail to assess long-range comprehension in real-world scenarios such as podcasts or lengthy broadcasts. We introduce VOICEGIRAFFE, a bilingual, open-domain benchmark for extreme long-context audio understanding. It comprises 1,500 curated triplets under a dual-level taxonomy: single-hop questions for fine-grained grounding and acoustic perception, and multi-hop questions requiring cross-segment evidence aggregation. We evaluate a broad suite of open-source and proprietary LALMs against human performance, and complement the primary multiple-choice protocol with open-ended evaluation that removes candidate options. Three findings emerge. First, VOICEGIRAFFE is far from saturation, and only one end-to-end LALM surpasses the human reference. Second, inference paradigms are model-dependent. End-to-end inference favours models with native long-context capacity, and cascaded aggregation stabilizes models overwhelmed by hour-scale audio. Third, long-range memory persistence is the key bottleneck, where current models reason over prominent evidence once localized but fail to retain event states over hour-scale context. VOICEGIRAFFE thus offers a challenging, diagnostic testbed for long-form audio understanding.

1 INTRODUCTION

Large Audio Language Models (LALMs) (Chu et al., 2023; 2024; Kong et al., 2024; Ghosh et al., 2025; Yu et al., 2025) have rapidly emerged as a unifying paradigm for speech, sound and music understanding, combining an audio encoder with a large language model to address diverse auditory tasks such as automatic speech recognition, audio captioning and audio question answering (Sakshi et al., 2025; Yang et al., 2024; Ma et al., 2025). As LALMs are increasingly deployed as audio assistants for meetings, podcasts, live broadcasts, and long interviews, reasoning over *hour-scale* audio inputs has become a crucial capability. However, current LALMs face fundamental challenges in long-context understanding, stemming from the high audio token rate and insufficient pre-training on long-horizon audio understanding. The computational overhead imposed by dense tokens, coupled with limited exposure to longer contexts during training. Consequently, models exhibit significant performance degradation when applied to long-form sequences.

A critical reason the community has struggled to tackle this challenge is that evaluation protocols lag far behind both model capabilities and real-world requirements. As summarized in Table 1a, existing short-form benchmarks such as MMAU (Sakshi et al., 2025), MMAR (Ma et al., 2025), AudioBench (Wang et al., 2025), and AIR-Bench (Yang et al., 2024) provide robust and comprehensive evaluation of audio understanding and reasoning for clips of 10–30 seconds, yet severe limitations emerge at longer temporal scales. While recent efforts have extended evaluation to minute-level contexts, such as AudioMarathon (He et al., 2025), ChronosAudio (Luo et al., 2026), and LongSpeech (Yang et al., 2026), these benchmarks still fall short of real-world scenarios requiring hour-level fine-grained understanding, such as sports events, TV dramas, and podcasts. More importantly, several benchmarks construct long-context inputs by concatenating short audio clips rather than using native long-form recordings, making it difficult to capture the natural temporal continuity and sparse long-range dependencies present in real-world audio. Beyond the duration

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.

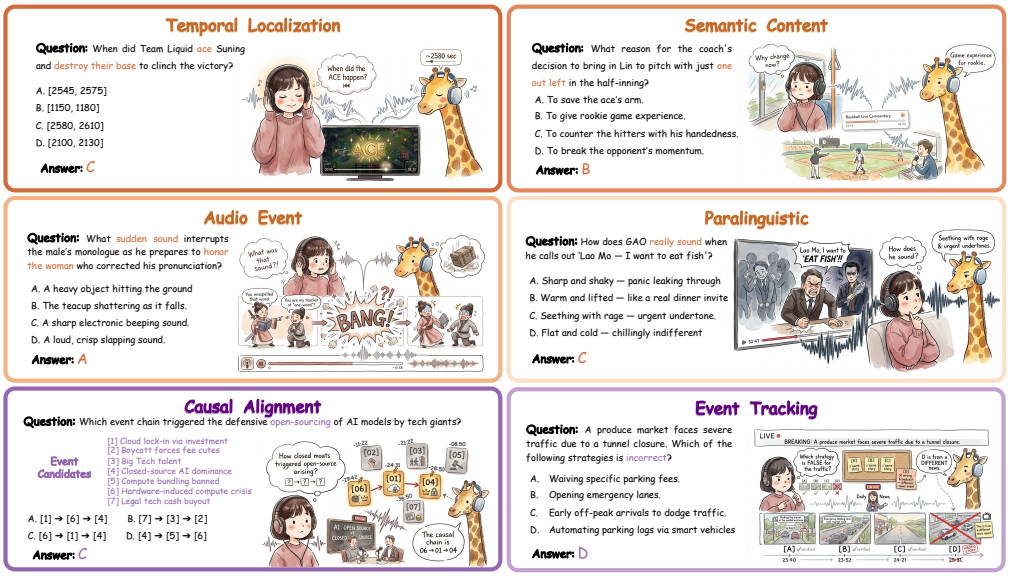


Figure 1: Examples of VOICEGIRAFFE across five real-world domains: e-sports commentary, sports broadcast, TV drama, news, and podcast. *Orange* panels show single-hop questions (temporal localization, semantics, audio events, paralinguistics). *Purple* panels show multi-hop questions (causal reasoning and event tracking) requiring evidence aggregation across long spans.

and authenticity gap, existing benchmarks tend to focus on single aspects and lack comprehensive integration of bilingual coverage, cross-domain diversity, and multi-hop reasoning.

To address these gaps, we introduce VOICEGIRAFFE, the first extremely long-context audio question-answer (AQA) benchmark for evaluating long-context understanding capacities in LALMs. As illustrated in Figure 1 and Table 1, VOICEGIRAFFE is built on three pillars. **First, hour-scale context**. VOICEGIRAFFE collects 123 long-form recordings that total approximately 113.1 hours, with an average duration of 55.2 minutes and 34% of recordings exceeding one hour. To our knowledge, it is the first open-domain AQA benchmark at this scale. **Second, bilingual, open-domain, and full-modality coverage.** The benchmark includes five domains across bilingual: sports, e-sports commentary, TV dramas, news, and podcasts. All recordings contain natural mixtures of speech, sound effects, and background music, challenging models on both verbal comprehension and cross-modal reasoning. **Third, a two-tier task taxonomy from single-hop perception to multi-hop reasoning.** Tier-1 questions assess single-hop perception across four categories: temporal localization, semantic content, audio event detection, and paralinguistic analysis. Tier-2 questions target multi-hop reasoning through causal event tracing and state-trajectory modeling, requiring evidence synthesis from 2 to 4 non-contiguous segments spanning tens of minutes.

Based on VOICEGIRAFFE, we evaluate 9 open-source and 4 proprietary audio-capable models together with a human reference. Our analysis reveals four key findings. First, VOICEGIRAFFE presents substantial challenges to current LALMs. Among the models capable of hour-scale E2E inference, only Qwen3.5-Omni-Plus surpasses the human reference, while open-source systems remain below the passing threshold under cascaded caption aggregation without any enhancement. Second, the optimal inference paradigm is highly model-dependent. E2E inference can better preserve fine-grained perceptual cues for models with strong native long-context audio understanding ability, whereas cascaded caption aggregation provides a more stable way for models overwhelmed by long-form audio. Reasoning-enhanced cascading further improves open-source models by compensating for weak long-range aggregation, but can also bottleneck strong proprietary systems when the external reasoning model is less capable than the evaluated model itself. Third, long-term memory persistence remains a key bottleneck. Unlike humans, who are better at tracking sparse states over extended context, LALMs are stronger at reasoning over salient causal cues but struggle to memorize and retrieve states across distant clips. In summary, our contributions are listed as follows:

- We present VOICEGIRAFFE, the first bilingual, hour-scale audio benchmark for LALMs, covering five real-world scenarios with naturally interleaved speech, sound and music.

Table 1: (**a**) Comparison of VOICEGIRAFFE with representative audio benchmarks across four key dimensions: average per-sample duration, heterogeneous modality coverage, bilingual contexts, and multi-hop reasoning. (**b**) Preliminary statistics of VOICEGIRAFFE.

(a) *Benchmark Comparison*

Benchmark	Duration	Modality			Bilingual	Multi-Hop
		Speech	Sound	Music		
MMAU Sakshi et al. (2025)	∼10 sec	✔	✔	✔	✗	✗
MMAR Ma et al. (2025)	∼20 sec	✔	✔	✔	✗	✔
AudioBench Wang et al. (2025)	∼14 sec	✔	✔	✔	✗	✗
AIR-Bench Yang et al. (2024)	∼35 sec	✔	✔	✔	✗	✗
AudioMarathon He et al. (2025)	∼4 min	✔	✔	✔	✗	✔
LongSpeech Yang et al. (2026)	∼10 min	✔	✗	✗	✔	✗
ChronosAudio Luo et al. (2026)	∼5 min	✔	✗	✗	✗	✗
BLAB Ahia et al. (2025)	∼51 min	✔	✗	✗	✗	✗
VOICEGIRAFFE *(ours)*	**∼1 hr**	✔	✔	✔	✔	✔

(b) *Benchmark Statistics*

Statistics	Value
Total recordings	123
Total audio duration (h)	113.1
Avg. duration (min)	55.2
Sub-tasks	6
Data-domains	5
Languages	EN / ZH
Total QA items	1,500
Single-hop items	1,000
Multi-hop items	500

- We design a two-tier task taxonomy with 1,500 carefully designed AQA pairs that progress from perception to reasoning, supporting systematic evaluation of long-context understanding.
- We conduct comprehensive analysis and comparison experiments across broad suite of LALMs together with a human reference, identifying long-context degradation, long-form memory failures and language bias as the key bottlenecks of current models.

2 RELATED WORK

2.1 LARGE AUDIO-LANGUAGE MODELS

The landscape of Large Audio Language Models (LALMs) has expanded rapidly along both open-source and proprietary tracks. On the open-source side, the Qwen-Audio series (Chu et al., 2023; 2024) pioneered general audio understanding, later extended by Qwen2.5-Omni and Qwen3-Omni (Xu et al., 2025; Team, 2025c) to unified audio–vision–language reasoning. Audio Flamingo (Kong et al., 2024) and its successor Audio Flamingo 2 (Ghosh et al., 2025) introduced few-shot and long-audio capabilities, while SALMONN (Tang et al., 2024), GAMA (Ghosh et al., 2024), Baichuan-Omni (Inc., 2025), Phi-4-Multimodal (Abouelenin et al., 2025), and MiMo-Audio (Xiaomi, 2025) contribute diverse architectural designs audio encoders, mixture-of-experts, and end-to-end streaming. Among proprietary systems, GPT-4o-Audio (OpenAI, 2024) and the Gemini 2.5 series family (Team, 2025a) demonstrate strong integrated multimodal reasoning, and Qwen3.5-Omni-PLUS now supports a 256K-token context window (Team, 2025b). Despite these ever-growing context capacities and reasoning abilities, *no systematic evaluation protocol exists for hour-scale naturalistic audio*, a setting that is increasingly common in real-world applications such as podcasts, meetings, live broadcasts, and long interviews. This gap motivates us to establish a rigorous evaluation framework for hour-scale audio understanding.

2.2 AUDIO UNDERSTANDING BENCHMARKS

Second-Level Benchmarks. AudioBench (Wang et al., 2025), AIR-Bench (Yang et al., 2024), MMAU (Sakshi et al., 2025), and MMAR (Ma et al., 2025) evaluate LALMs on clips of 10–30 seconds, providing comprehensive coverage of perception, instruction following, and reasoning within a single auditory scene. While rigorous in scope, short-context evaluation fundamentally limits assessment of two critical long-context capabilities: (*i*) maintaining memory of important information across extended listening spans, and (*ii*) aggregating evidence from multiple non-contiguous segments separated by minutes or hours. In contrast, we bridge this gap with VOICEGIRAFFE designed to probe these capabilities in the extreme long-context regime.

Minute-Level Benchmarks. Recent benchmarks have begun extending evaluation timescales into the minute regime, with efforts spanning 3–10 minutes: AudioMarathon (He et al., 2025), LongSpeech (Yang et al., 2026), ChronosAudio (Luo et al., 2026). While these efforts mark valuable progress, they fall short of addressing the real-world challenge users increasingly face, namely rapidly extracting information from hour-long podcasts, lectures, and sports commentary. More importantly, several benchmarks construct long-context inputs by concatenating short audio clips rather than using native long-form recordings. Even BLAB (Ahia et al., 2025), which extends to 51 minutes,

remains confined to speech-specific evaluation, without testing multi-domain acoustic understanding or complex temporal reasoning over hour-scale inputs. To bridge this gap, our benchmark targets the hour scale with multi-domain coverage and questions requiring multi-hop aggregation.

As summarised in Table 1, VOICEGIRAFFE addresses three orthogonal limitations simultaneously: (***i***) it evaluates on **hour-scale naturalistic audio** drawn from intact real-world recordings rather than synthetic concatenations; (***ii***) it covers **bilingual** (English + Chinese) content across **five domains** with naturally interleaved speech, music, and environmental sound; and (***iii***) it introduces a **perception-to-reasoning** two-tier ontology for comprehensive evaluation across long-form contexts.

3 VOICEGIRAFFE

3.1 OVERVIEW

VOICEGIRAFFE is a benchmark designed to evaluate the long-context understanding of LALMs. Figure 2 and Table 1b summarize the specific statistics across data and task, while Table 2 presents evaluation results across state-of-the-art LALMs.



(a) Data Distribution (b) Task Taxonomy (c) Duration Statistics across Tasks

Figure 2: (**a**) Data distribution across five domains encompassing modalities of speech, sound, and music events. (**b**) Task taxonomy partitioning evaluation into foundational single-hop perception and complex multi-hop reasoning tiers. (**c**) Statistical consistency of hour-scale durations across diverse task types to ensure balanced evaluation in duration.

Domain Coverage. Unlike prior benchmarks that artificially splice short clips (Yang et al., 2026; He et al., 2025), VOICEGIRAFFE preserves entire long-form recordings spanning authentic real-world scenarios across five diverse domains in English and Chinese. As illustrated in Figure 2a, our corpus encompasses: *(i) Sports Commentary*, featuring rapid play-by-play narration with dense temporal events; *(ii) E-sports Casting*, characterized by overlapping team communications, and chaotic multi-speaker environments; *(iii) TV Dramas*, presenting complex character interactions, emotional arcs, and rich background soundscapes; *(iv) News Broadcasts*, offering structured yet information-dense content with formal register and topic transitions; and *(v) Interview Podcasts*, exhibiting natural conversational flow, turn-taking dynamics, and extended argumentative discourse.

Task Coverage. As illustrated in Figure 1 and 2b, we organize questions into a two-tier taxonomy: **Tier-1** establishes foundational audio-language understanding through four perceptual pillars: *(i) Temporal Localization* requires precise timestamp retrieval of events within hour-long streams, testing fine-grained temporal indexing; *(ii) Semantic Content* probes factual comprehension and topic tracking across extended discourse; *(iii) Audio/Acoustic Events* evaluates non-speech sound and music recognition; and *(iv) Paralinguistic Analysis* assesses speaker characteristic understanding (emotion, age, gender, timbre, pitch) beyond its transcription. **Tier-2** raises the challenge to multi-hop reasoning over non-contiguous segments, covering two complementary patterns. *Event Tracking* serves as a stress test for long-term memory, requiring models to exhaustively scan the audio and aggregate sparse, often implicit events, such as counting how many times an event occurs or determining which event never happens. In contrast, *Causal Alignment* evaluates logical coherence and narrative reasoning, requiring models to reconstruct causal chains from distributed evidence, identify temporal dependencies, and detect logical inconsistencies.

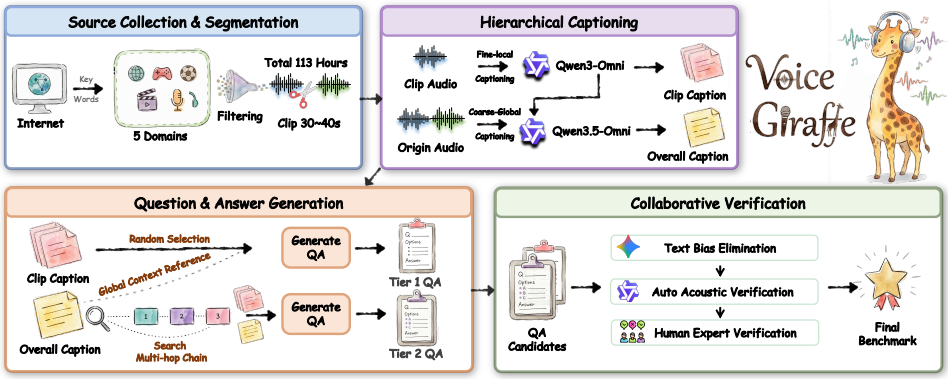


Figure 3: The four-stage data construction pipeline of VOICEGIRAFFE: (1) source collection and VAD-guided segmentation; (2) hierarchical captioning with state-of-the-art Omni model; (3) question and answer construction across two tiers; (4) collaborative model-to-human verification.

Together, these six task categories comprehensively evaluate the capabilities required for hour-scale audio understanding, ranging from low-level perceptual analysis to high-level reasoning, and from local temporal grounding to global narrative aggregation.

3.2 DATA CURATION PIPELINE

As shown in Figure 3, We adopt a rigorous multi-stage framework to construct VOICEGIRAFFE that transforms raw long-form recordings into audited QA pairs through Voice Activity Detection (VAD) segmentation, clip-level captioning, overall captioning, QA generation, and multimodal verification.

Step 1. Source Collection and Segmentation. We collect complete long-form recordings from public platforms across five domains. For each domain, we use keyword lists to retrieve content. Retrieved recordings undergo manual filtering to exclude those outside 30 minutes to 2 hours, or with domain misalignment. Retained recordings are segmented into 30–40-second clips using `pyannote` VAD (Bredin et al., 2020), establishing global time anchors for consistent annotation across all stages.

Step 2. Hierarchical Captioning. Each clip is captioned by `Qwen3-Omni` (Team, 2025c) covering voices (timestamp, speaker ID, transcript, paralinguistic cues), sound effects (category, timestamp), and background music. To further capture dependencies beyond individual clips, we then input all clip captions together with the original long-form audio into `Qwen3.5-Omni` (Team, 2025b), generating a holistic caption for building global themes, narrative structure, and cross-segment event relationships. This design enables both local detail and global context for question construction.

Step 3. QA Construction. Based on the annotated metadata, we employ `Gemini-3.0` (Team, 2025a) with an evidence-grounded pipeline. For Tier-1 perception tasks, we pair the target clip caption with its overall caption, ensuring that fine-grained perceptual questions are grounded in local evidence while remaining consistent with the global context reference. For Tier-2 reasoning tasks, we adopt a two-stage synthesis process. We first identify potential multi-hop reasoning pathways from the overall caption, and then retrieve the corresponding clip captions to anchor the final QA generation. This retrieval-anchored strategy reduces hallucination by ensuring that complex reasoning chains are supported by verified clip-level evidence. Each question is paired with four multiple-choice options, including adversarial distractors designed to target partial reasoning shortcuts.

Step 4. Collaborative Verification. Each candidate question undergoes a three-stage collaborative filtering process, in which specialized models and human experts verify complementary quality dimensions. *(i) Text-bias elimination*: a text-only LLM assesses whether the question can be answered without listening to the audio to discard audio-independent items. *(ii) Automated acoustic verification*: an omni-modal model checks whether each QA pair is grounded in specific acoustic evidence from the corresponding audio clips, where questions without verifiable audio cues are removed. *(iii) Human expert review*: 8 human annotators independently review the benchmark to ensure that each question is grounded in the audio, answerable from the provided evidence, and paired with an accurate answer. Only questions that pass all three stages are included in the benchmark.

3.3 ACCESS AND COPYRIGHT COMPLIANCE

To strictly adhere to copyright regulations and platform Terms of Service (e.g., YouTube ToS), VOICEGIRAFFE is released under a framework that completely avoids the redistribution of raw copyrighted audio. We provide three complementary access paths designed to ensure full reproducibility while maintaining rigorous copyright compliance: ***(i) Canonical path for local retrieval.*** We strictly do not host or distribute raw audio files. Instead, we release recording identifiers (public video IDs), QA annotations, and automated preprocessing scripts. Users retrieve the audio directly from the original public sources into their local environments. This standard practice ensures that data access complies with the original platforms' terms, while allowing evaluation on any LALM without architectural restrictions; ***(ii) Feature extraction path***. To facilitate seamless evaluation and avoid issues arising from broken data links, we operate a feature preprocessing service. Researchers can submit their custom audio encoder (code and weights) via an online request form. We process the raw long audio entirely on our secure local servers and return only the extracted acoustic features. This allows downstream models to be trained and tested while guaranteeing that the raw copyrighted audio never leaves our servers. Both paths are complementary. The former ensures unrestricted local evaluation, and the extraction service resolves the testing bottleneck without violating redistribution constraints.

4 EXPERIMENTS

Benchmark Models. We benchmark a comprehensive suite of contemporary models against human performance as an upper-bound reference. The suite comprises two model families: *large audio language models* designed for audio-language understanding including MiMo-Audio (Xiaomi, 2025), MOSS-Audio (Team, 2026), Audio-Flamingo-3 (Goel et al., 2025), and *omni-modal language models* supporting fully multimodal input/output including QwenOmni series (Xu et al., 2025; Team, 2025c;b), Gemini series (Team, 2025a), Phi-4-Multimodal (Abouelenin et al., 2025), MiniCPM-o-4.5 (Yao et al., 2024). Furthermore, we introduce a reasoning-enhanced cascading setting, where the Large Reasoning Model (LRM) `GPT-5.2-Thinking` (OpenAI, 2025) performs inference-time reasoning over clip-level audio captions generated by above LALMs to answer the corresponding QA pairs. The LRM is a strong *external* reasoner, distinct from the evaluated LALM, and operates on the caption transcript rather than on the original audio; formal definitions of all three paradigms and the motivation for the LRM-Cascaded pipeline are given in Appendix I.

Inference Settings. We implement three inference settings to evaluate models with different context capacities. *(i) End-to-End (E2E).* For models that natively support hour-scale context windows, we directly feed the complete audio signal together with the query for inference. *(ii) Cascaded Caption Aggregation.* For models limited to short audio inputs, we adopt a cascaded pipeline that performs sliding-window captioning over the full audio timeline with a 30-second window size. Each window is captioned with speech transcripts, speaker information (gender, age, emotion, etc.), sound events, and music cues. The resulting clip-level captions are concatenated into a textual description and processed together with the query by the original LALM. *(iii) Reasoning-Enhanced Cascading.* This setting extends the cascaded pipeline by feeding the aggregated LALM captions and the query into a stronger large reasoning model (OpenAI, 2025), enabling test-time reasoning over the audio caption.

Evaluation Settings. Given that all VOICEGIRAFFE questions follow a multiple-choice format, we adopt macro-average accuracy as the primary metric. Specifically, each model receives the audio recording, question text, and candidate options as input, and we assess whether its selected option matches the ground truth. Following established practices in MMAU (Sakshi et al., 2025) and MMAR (Ma et al., 2025), we employ regular expression matching and string alignment to extract and compare model predictions against gold labels. For models without explicit reasoning output, we directly evaluate the final output. For models with a thinking mode that generate an intermediate chain-of-thought steps, we strip the reasoning content and evaluate only the final answer, ensuring fair and consistent comparison across different model architectures.

Human Reference. We engage eight annotators with native-level proficiency in both English and Chinese. Given the hour-scale listening burden, we improve efficiency by sampling multiple questions per audio recording across all six tasks, allowing comprehensive evaluation from a single listening. Sampling is stratified across domains and languages. Each annotator independently evaluates a randomly assigned subset of 150 questions (75 per language), sampled via stratified random sampling to preserve the benchmark's proportional distribution across domains and languages.

Table 2: VOICEGIRAFFE results across 13 models and 6 task categories: 4 single-hop tasks—*Temp. Loc.* (Temporal Localization), *Sem. Con.* (Semantic Content), *Aco. Evt.* (Acoustic Event), and *Paralin.* (Paralinguistic), and 2 multi-hop tasks—*Causal* (Causal alignment) and *Evt. Trk.* (Event Tracking). **Size**: number of activated parameters. **E2E**: end-to-end audio inference without caption cascading. **Think**: native extended thinking mode. **LRM**: augmentation for inference with an external large reasoning model over streaming clip audio captions. Best results are highlighted in setting-specific **bold** colors: **red** for end-to-end inference, **yellow** for cascaded caption aggregation, and **blue** for reasoning-enhanced cascading.

Model	Size	Thinking	E2E	LRM	Single-hop Perception					Multi-hop Reasoning			Overall
					Temp. Loc.	Sem. Con.	Aco. Evt.	Paralin.	Avg.	Causal	Evt. Trk.	Avg.	Avg.
Open-source Models													
MiniCPM-o-4.5	9B	✗	✗	✗	3.60	4.00	4.50	2.29	3.40	5.20	3.60	4.40	3.73
				✓	69.60	91.00	51.00	68.86	69.90	42.40	55.60	49.00	**62.93**
Phi-4-Multimodal	5.6B	✗	✗	✗	30.40	30.00	31.50	33.71	31.70	30.00	23.60	26.80	30.07
				✓	32.80	39.50	44.00	50.86	42.70	39.60	22.80	31.20	38.87
Audio-Flamingo-3	8B	✗	✗	✗	35.60	43.50	40.50	42.29	40.50	64.40	40.40	52.40	44.47
				✓	65.60	52.00	49.00	58.86	57.20	38.40	33.60	36.00	50.13
Qwen2.5-Omni	7B	✗	✗	✗	38.00	48.00	31.50	50.57	43.10	43.20	30.80	37.00	41.07
				✓	74.40	92.00	39.50	62.86	66.90	41.60	52.40	47.00	60.27
MiMo-Audio	7B	✗	✗	✗	34.00	39.50	43.50	46.29	41.30	40.40	23.60	32.00	38.20
				✓	**84.00**	88.00	48.50	65.43	**71.20**	42.40	50.00	46.20	62.87
MOSS-Audio	8B	✗	✗	✗	28.80	40.50	39.00	47.43	39.70	35.20	26.00	30.60	36.67
				✓	62.40	72.50	51.00	64.86	63.00	42.00	40.40	41.20	55.73
		✓	✗	✗	31.20	36.00	44.00	60.00	44.80	65.60	27.20	46.40	45.33
				✓	73.20	76.50	52.50	67.43	67.70	40.80	41.60	41.20	58.87
Qwen3-Omni	30B(A3B)	✗	✗	✗	30.80	69.00	48.50	61.43	52.70	45.60	47.20	46.40	50.60
				✓	39.20	91.50	51.00	68.29	62.20	46.00	52.40	49.20	57.87
		✓	✗	✗	20.00	58.00	43.00	46.57	41.50	63.20	36.00	49.60	44.20
				✓	38.80	68.50	50.50	64.29	56.00	39.20	42.40	40.80	50.93
Proprietary Models													
Gemini-2.5-Pro	–	✗	✗	✗	72.80	95.50	58.00	68.29	72.80	74.40	**70.00**	72.20	72.60
			✗	✓	37.60	95.50	59.50	**70.57**	65.10	**46.80**	60.40	**53.60**	61.27
			✓	✗	33.20	61.00	43.50	52.57	47.60	52.00	37.20	44.60	46.60
Gemini-3.1-Pro	–	✗	✗	✗	**79.60**	**97.50**	57.00	66.29	**74.00**	**85.20**	69.60	**77.40**	**75.13**
			✗	✓	32.80	**96.50**	57.00	68.00	62.70	45.60	**61.20**	53.40	59.60
			✓	✗	32.80	53.00	33.50	46.29	41.70	36.40	40.00	38.20	40.53
Qwen3.5-Omni-Flash	–	✗	✗	✗	31.60	79.50	60.00	64.57	58.40	64.00	49.60	56.80	57.87
			✗	✓	37.20	83.00	57.00	68.29	61.20	45.20	52.00	48.60	57.00
			✓	✗	72.80	84.50	56.50	59.71	67.30	70.00	47.60	58.80	64.47
Qwen3.5-Omni-Plus	–	✗	✗	✗	44.80	92.00	**60.50**	**69.43**	66.00	76.40	56.80	66.60	66.20
			✗	✓	39.60	92.50	**64.00**	67.43	64.80	42.80	55.20	49.00	59.53
			✓	✗	**92.00**	**94.00**	**64.00**	**72.00**	**79.80**	**77.20**	**59.60**	**68.40**	**76.00**
Human Reference					63.89	90.74	79.63	79.17	77.38	52.24	61.11	56.20	70.51

4.1 EXPERIMENTAL RESULTS

Table 2 presents model performance of VOICEGIRAFFE, evaluated via multiple-choice accuracy (%). Results reveal several critical findings regarding the benchmark's difficulty and LALM's capabilities.

Finding 1: Hour-scale understanding of VOICEGIRAFFE remains extremely challenging for all models. Among the four models capable of hour-scale E2E inference, only Qwen3.5-Omni-Plus surpasses the human reference of 70.51%, while the remaining three fall substantially below human performance. For open-source models that rely solely on cascaded caption aggregation without LRM enhancement, the best overall accuracy reaches merely 50.60% (Qwen3-Omni), with most models clustered between 30% and 45%. Reasoning-enhanced cascading then improves performance markedly, suggesting that current LALMs possess reasonable short-term perception ability but still face major bottlenecks in long-term reasoning and memory aggregation. Furthermore, our ablation contrasting full recordings against 30-second evidence-centered clips (Appendix K) confirms that while models can answer localized queries accurately, they often fail to retrieve the same evidence from hour-scale inputs. This wide gap confirms VOICEGIRAFFE as a challenging benchmark that is far from saturation. We further verify that the multiple-choice format does not overstate model ability: under an open-ended (OQA) protocol without candidate options, every model drops substantially (9.9–18.1 points overall), confirming that closed-ended QA accuracy is only a lower bound on task difficulty (Appendix J).

Finding 2: The optimal inference paradigm is model-dependent. Table 2 reveals a fundamental trade-off between perceptual fidelity and long-range semantic reach. End-to-end (E2E) inference preserves fine-grained acoustic and paralinguistic cues, heavily favoring models with robust native long-context capabilities. For instance, Qwen3.5-Omni-Plus achieves the highest overall score,

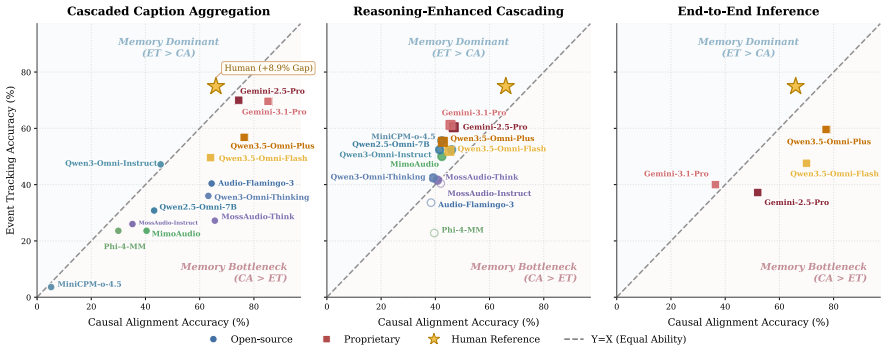


Figure 4: Performance asymmetry between Causal Alignment (CA) and Event Tracking (ET). The diagonal ($Y = X$) represents balanced capability. Current LALMs across both native Cascaded and E2E settings overwhelmingly fall into the lower "Memory Bottleneck" region ($CA > ET$).

outperforming its cascaded variant by 9.8%. Conversely, without sufficient hour-scale modeling, extensive audio context causes information dilution; Gemini-3.1-Pro suffers a 34.6% performance drop under E2E compared to cascading. Cascaded inference circumvents this by compressing audio into temporally ordered captions, aiding long-range semantic integration but introducing an irreversible bottleneck that discards subtle acoustic details. Reasoning-enhanced cascading applies an external large reasoning model (LRM) over these captions, yielding asymmetric effects. It substantially benefits open-source LALMs (improving average scores from 37.2% to 55.4%) by patching their text-aggregation deficits. However, for proprietary models with already superior long-context reasoning, the external LRM becomes a bottleneck and degrades performance (Sec. 4.2). Ultimately, E2E inference suits tasks requiring acoustic fidelity in robust models, while cascaded frameworks are preferable for semantic aggregation when native processing capacity is limited.

Finding 3: Long-range memory remains a key bottleneck, in contrast to human. As shown in Figure 4, while human annotators excel at Event Tracking over Causal Alignment (+8.9%) due to robust long-term episodic memory, models exhibit the exact reverse trend with gaps up to 38%, highlighting their reliance on localized cues over persistent event tracking. The asymmetry suggests that current LALMs are relatively capable of reconstructing causal relations from salient semantic cues, but struggle to preserve and aggregate sparse event states across hour-scale contexts. In other words, their failures are less about causal reasoning and more about persistent memory over audio evidence across long-term contexts. Notably, LRM enhancement partially mitigates this imbalance. Under reasoning-enhanced cascading, Event Tracking often surpasses Causal Alignment, suggesting that strong LRMs possess more robust long-context memory and evidence aggregation capabilities than LALMs themselves when reasoning over temporally ordered captions. However, relying on external text reasoners introduces a strict information bottleneck that explains the simultaneous degradation in Causal Alignment. While text-based LRMs excel at aggregating sparse events from captions, they lack the native audio priors required to interpret acoustic signals. Deep causal understanding inherently depends on non-semantic nuances. These critical acoustic cues are inevitably omitted or distorted during intermediate captioning, resulting in loss for causal tasks.

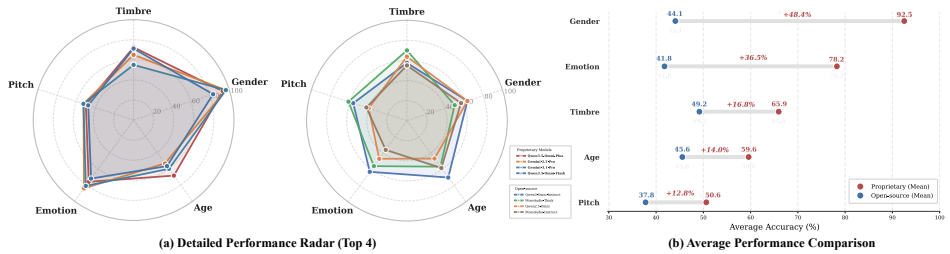
Overall, these findings underline the urgent need for further innovation in long-form audio understanding, to address fundamental challenges in memory persistence and acoustic perception that current models exhibit at the hour scale.

4.2 Large Reasoning Model for Enhanced Cascading Inference

Table 3 shows that external LRM selection substantially impacts performance, proving reasoning-enhanced cascading is not a plug-and-play solution. While Gemini-3.1-Pro consistently boosts all LALMs (averaging +22% overall and +28% on open-source models), GPT-5.2 exhibits sharp asymmetry—improving weak models but degrading strong proprietary ones by 9%, particularly on multi-hop tasks due to spurious causal reasoning over text captions. Thus, while an LRM serves as an effective surrogate memory for weaker models, an ill-matched LRM imposes a bottleneck by introducing extra hallucinations to stronger systems. This highlights the need for audio-grounded reasoning mechanisms capable of re-listening to original audio or capturing audio-grounded caption for evidence verification rather than naively routing captions to generic LRMs.

Table 3: Ablation study on LRM selection under cascade caption aggregation mode. Each LALM is paired with different large reasoning models. **No LRM** denotes the pure cascade baseline. Gains over baseline are shown as ↑ (improvement) or ↓ (degradation).

LALM	**No LRM** (baseline)			**GPT-5.2**			**Gemini-3.1-Pro**		
	Single-hop	Multi-hop	Overall	Single-hop	Multi-hop	Overall	Single-hop	Multi-hop	Overall
Open-source LALMs									
MiniCPM-o-4.5 (9B)	3.40	4.40	3.70	69.9 ↑66.5	49.0 ↑44.6	62.9 ↑59.2	75.8 ↑72.4	78.0 ↑73.6	76.5 ↑72.8
Phi-4-Multimodal (5.6B)	31.7	26.8	30.1	42.7 ↑11.0	31.2 ↑4.4	38.9 ↑8.8	41.6 ↑9.9	29.8 ↑3.0	37.7 ↑7.6
Audio-Flamingo-3 (8B)	40.5	52.4	44.5	57.2 ↑16.7	36.0 ↓16.4	50.1 ↑5.6	57.5 ↑17.0	39.2 ↓13.2	51.4 ↑6.9
Qwen2.5-Omni (7B)	43.1	37.0	41.1	66.9 ↑23.8	47.0 ↑10.0	60.3 ↑19.2	71.1 ↑28.0	73.6 ↑36.6	71.9 ↑30.8
MiMo-Audio (7B)	41.3	32.0	38.2	71.2 ↑29.9	46.2 ↑14.2	62.9 ↑24.7	74.9 ↑33.6	74.4 ↑42.4	74.7 ↑36.5
MOSS-Audio (8B)	39.7	30.6	36.7	63.0 ↑23.3	41.2 ↑10.6	55.7 ↑19.0	68.3 ↑28.6	61.0 ↑30.4	65.9 ↑29.2
MOSS-Audio-Think (8B)	44.8	46.4	45.3	67.7 ↑22.9	41.2 ↓5.2	58.9 ↑13.6	74.9 ↑30.1	65.4 ↑19.0	71.7 ↑26.4
Qwen3-Omni (30B/A3B)	52.7	46.4	50.6	62.2 ↑9.5	49.2 ↑2.8	57.9 ↑7.3	77.2 ↑24.5	74.8 ↑28.4	76.4 ↑25.8
Qwen3-Omni-Think (30B/A3B)	41.5	49.6	44.2	56.0 ↑14.5	40.8 ↓8.8	50.9 ↑6.7	67.2 ↑25.7	58.2 ↑8.6	64.2 ↑20.0
Proprietary LALMs									
Gemini-2.5-Pro	72.8	72.2	72.6	65.1 ↓7.7	53.6 ↓18.6	61.3 ↓11.3	75.4 ↑2.6	77.8 ↑5.6	76.2 ↑3.6
Gemini-3.1-Pro	74.0	77.4	75.1	62.7 ↓11.3	53.4 ↓24.0	59.6 ↓15.5	74.0 0.0	77.4 0	75.1 0.0
Qwen3.5-Omni-Flash	58.4	56.8	57.9	61.2 ↑2.8	48.6 ↓8.2	57.0 ↓0.9	74.6 ↑16.2	69.8 ↑13.0	73.0 ↑15.1
Qwen3.5-Omni-Plus	66.0	66.6	66.2	64.8 ↓1.2	49.0 ↓17.6	59.5 ↓6.7	79.3 ↑13.3	76.4 ↑9.8	78.3 ↑12.1



(a) Detailed Performance Radar (Top 4)

(b) Average Performance Comparison

Figure 5: Paralinguistic understanding across fine-grained attributes in the cascaded inference setting.

4.3 Paralinguistic Understanding Analysis

Figure 5 reveals a pronounced performance gap between proprietary and open-source models across paralinguistic dimensions. Proprietary models achieve substantially higher accuracy on coarse-grained demographic and affective perception, reaching 92.5% on Gender and 78.2% on Emotion on average. In contrast, open-source models lag behind most severely on Gender prediction, with a 48.4% gap, indicating a clear weakness in modeling basic speaker characteristics. However, this gap is not uniform across all attributes. Open-source models are relatively more competitive on Age and Timbre recognition, suggesting that some salient acoustic cues can still be captured reasonably well.

Most notably, Pitch perception emerges as a shared failure mode for both model families, with the lowest average accuracy among all paralinguistic attributes for open-source models (37.8%) and proprietary models (50.6%). This indicates that current LALMs remain limited in modeling fine-grained frequency dynamics, even when they can capture semantic or coarse acoustic cues. Overall, these results highlight the need for future LALMs to better understand paralinguistic domains.

5 Conclusion

We introduce VoiceGiraffe, a bilingual benchmark for evaluating hour-scale audio understanding in LALMs across real-world scenarios with interleaved speech, sound, and music. Through a systematic evaluation of broad LALMs, we show that this setting remains far from solved. Most models fall below the passing threshold, with major weaknesses in paralinguistic perception and multi-hop reasoning that require perceptual fidelity and persistent memory. Our analyses further reveal key bottlenecks, including duration-induced performance degradation, model-dependent inference paradigms, limited long-form memory persistence, language bias, and weakness in pitch perception. These findings suggest that future LALMs require stronger long-context audio understanding and reasoning, and more robust memory mechanisms. We release the benchmark to support future research on hour-scale audio understanding.

AI USE STATEMENT

Required disclosures. We used generative AI tools for the following tasks: *data processing and filtering* (as described in Section 3.2, model-generated captions and candidate QA pairs were produced with Qwen-Omni and Gemini models, then processed and filtered programmatically), *implementing methods* (portions of the evaluation and annotation-inspection scripts were drafted with AI assistance), and *cleaning and reformatting the dataset* (normalising timestamps, captions, and QA records).

Not used. We did not use generative AI tools to generate synthetic audio, to formulate mathematical claims, or to assist in the writing of proofs; these are not applicable to this work.

Recommended disclosures. Additionally, we used generative AI tools to create and modify figures and tables, to draft and edit portions of this paper for readability.

Review and responsibility. We have reviewed all AI-assisted work. All model-generated captions and candidate QA pairs were subsequently verified by human annotators who re-listened to the source audio and rewrote items as needed (Section 3.2, Step 4), so no item in the released benchmark reflects unverified model output. LLM-generated code was tested against the released annotations by two authors. All literature references were checked against the original publications.

ETHICS STATEMENT

VOICEGIRAFFE is built from publicly available broadcast audio (sports commentary, e-sports casting, news programs, TV dramas, interview-style podcasts). We acknowledge three classes of ethical considerations.

Copyright. The audio sources in VOICEGIRAFFE remain the property of their respective rights holders and are *not* redistributed with this paper. What we release is our own contribution: recording identifiers (public video IDs with in-recording time anchors), the QA annotations and derived captions, and the preprocessing scripts, all of which are our original work. Users obtain the underlying audio directly from the original public sources under those sources' own terms, exactly as they would when viewing or listening to that content. To keep the benchmark stable over time, we re-verify every source at release and periodically thereafter; if a recording becomes unavailable, it is replaced by a matched substitute that satisfies the same domain, language, and duration guidelines (Appendix A).

Speaker Privacy. All recordings come from public broadcasts or public-facing interviews. We do not attempt speaker re-identification beyond what is already public in the source material, and we do not release derived voice prints.

Dual-Use. Long-form audio understanding technology has clear benefits but also risks. VOICEGIRAFFE is a passive *evaluation* resource with no generative models or voice-cloning capability, and we explicitly discourage its use for any applications that would infringe on speaker privacy or facilitate synthetic impersonation.

Human Subjects. The human reference scores reported in Section 4 were collected from eight annotators engaged under a standard research-annotation agreement. Participation was voluntary and compensated, the task involved listening to publicly broadcast material only, and no personal or sensitive information about the annotators was retained.

REPRODUCIBILITY STATEMENT

We release the full evaluation suite, including the source selection criteria (Appendix A), the bilingual audio inventory (Appendix B), the VAD segmentation configuration, the data-construction pipeline (Section 3.2), the clip-level captioning and overall-caption synthesis prompts (Appendices P and Q), the T1/T2 QA generation templates (Appendix R), the leakage and hallucination audit protocols (Appendix D), the human-evaluation instructions (Appendix F), and the per-model inference configurations (Appendix G). The three inference paradigms and their exact definitions are given in Appendix I, and the benchmark access paths are described in Section 3.3. All hyperparameters, decoding configurations, and evaluation settings are specified in the main text or appendix. With the released scripts, the construction pipeline and the evaluation protocol can be reproduced end-to-end from the original source recordings.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *CoRR*, 2025.

Orevaoghene Ahia, Martijn Bartelds, Kabir Ahuja, Hila Gonen, Valentin Hofmann, et al. BLAB: brutally long audio bench. *CoRR*, 2025.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote.audio: Neural building blocks for speaker diarization. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 7124–7128, 2020.

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. *CoRR*, 2023.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *CoRR*, 2024.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S. Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 6288–6313, 2024.

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. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research, 2025.

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

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. Audiomarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio llms. *CoRR*, 2025.

Baichuan Inc. Baichuan-omni-1.5 technical report. *CoRR*, abs/2501.15368, 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. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pp. 25125–25148, 2024.

Kaiwen Luo, Liang Lin, Yibo Zhang, Moayad Aloqaily, Dexian Wang, Zhenhong Zhou, Junwei Zhang, Kun Wang, Li Sun, and Qingsong Wen. Chronosaudio: A comprehensive long-audio benchmark for evaluating audio-large language models. *CoRR*, 2026.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, et al. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *CoRR*, 2025.

OpenAI. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024.

OpenAI. Gpt-5.2. `https://deploymentsafety.openai.com/gpt-5-2/introduction`, 2025.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

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. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, 2025a.

OpenMOSS Team. Moss-audio technical report. `https://github.com/OpenMOSS/MOSS-Audio`, 2026. GitHub repository.

Qwen Team. Qwen3.5-omni technical report. *CoRR*, 2025b.

Qwen Team. Qwen3-omni technical report. *CoRR*, 2025c.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 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, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 4297–4316, 2025.

LLM-Core Xiaomi. Mimo-audio: Audio language models are few-shot learners. *CoRR*, 2025.

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. Qwen2.5-omni technical report. *CoRR*, 2025.

Fei Yang, Xuanfan Ni, Renyi Yang, Jiahui Geng, Qing Li, Chenyang Lyu, Yichao Du, Longyue Wang, Weihua Luo, and Kaifu Zhang. Longspeech: A scalable benchmark for transcription, translation and understanding in long speech. *CoRR*, 2026.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 1979–1998, 2024.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. Salmonn-omni: A standalone speech LLM without codec injection for full-duplex conversation. *CoRR*, 2025.

APPENDIX OVERVIEW

This appendix is organised as follows. Appendix A gives the per-domain content guidelines for data selection. Appendix B lists the preliminary long-form audio inventory. Appendix C provides the task taxonomy for VOICEGIRAFFE. Appendix D details the textual and multimodal leakage audit protocols. Appendix E shows the model evaluation prompt. protocol for annotation verification. Appendix G summarises per-model inference configurations. Appendix H provides baseline model cards. Appendix I defines the three inference paradigms and their interpretation. Appendix J reports the open-ended QA study. Appendix K reports the short-clip vs. long-form comparison. Appendix L reports the memory-vs-reasoning comparison across paradigms. Appendix M reports the audio-token compression study. Appendix N gives the per-task Chinese/English breakdown. Appendix O discusses limitations. Appendix P provides the clip-level captioning prompt. Appendix Q provides the overall caption synthesis prompt. Appendix R gives the T1 / T2 QA generation templates.

A DATA COLLECTION GUIDELINES

Table 4 summarises the per-domain content guidelines used during source curation, covering target counts, minimum durations, content sub-distributions and time-range constraints. The guidelines apply jointly to the English and Chinese halves of VOICEGIRAFFE; the Chinese equivalents of each domain follow the same acoustic and discourse criteria but on Chinese-medium broadcast sources.

Table 4: Per-domain data-selection guidelines for VOICEGIRAFFE. Each English (E-) / Chinese (C-) domain lists its target number of recordings, minimum per-recording duration, content sub-distribution and time-range constraint. All recordings satisfying these guidelines are eligible for the collection pool.

ID	Domain	Target	Min. dur.	Content sub-distribution	Time range
E-01	Sports-casting	20	≥ 50 min	NBA/NFL (American): 40%; English Premier League / Champions League: 40%; Boxing / Tennis / Baseball etc.: 20%.	2018–2025
E-02	E-sports (EN)	20	≥ 50 min	LoL (Worlds / MSI): 50%; Dota 2 (TI): 30%; CS:GO / Valorant Major: 20%. Finals / semi-finals preferred.	2020–2025
E-03	TV Dramas (EN)	20	≥ 50 min	*Friends*: 40%; *Big Bang Theory*: 20%; *Modern Family*: 20%; other supplements.	Original air year
E-04	News (EN)	20	≥ 30 min	BBC World News / BBC Radio 4: 60%; CNN International / Al Jazeera English: 40%. On-the-hour or long-form specials only.	2020–2025
E-05	Interview (EN)	20	≥ 50 min	Celebrity interviews: 60% recordings; multi-speaker podcasts / deep-dive interviews: 40% recordings (speakers > 1).	2020–2025
C-01	Sports Commentary (ZH)	20	≥ 50 min	Basketball (CBA / NBA ZH): 50%; Football (CSL / World Cup / Euro ZH): 50%.	2019–2025
C-02	Game Casting (ZH)	20	≥ 50 min	LoL league (finals / playoffs): 70%; KoF / Dota / HoK: 30%.	2020–2024
C-03	TV Dramas (ZH)	20	≥ 50 min	*My Own Swordsman* 20%, *Home with Kids* 10%, *In the Name of the People* 10%, *Ode to Joy* 20%, *iPartment* 20%, *The Knockout* 20%.	Original air year
C-04	News (ZH)	20	≥ 30 min	CCTV news: 40% recordings; local-station news: 60% recordings.	2020–2024
C-05	Interview (ZH)	20	≥ 50 min	Deep-dive interview shows (e.g. *Thirteen Invitations* 50%, *Business Interview Record* 50%) ; speakers > 1.	2020–2025

Table 5: Preliminary duration distribution of VOICEGIRAFFE at the recording level. Columns: number of recordings in each duration bucket, total seconds, total hours, recording ratio and duration ratio. All 126 recordings are at least 30 min long and the longest bucket reaches 1.5 h.

Duration range	# Recordings	Total seconds	Total hours	Recording ratio	Duration ratio
30–45 min	14	32,936	9.15	11.1%	7.9%
45–60 min	67	206,654	57.40	53.2%	49.5%
60–90 min	45	177,891	49.41	35.7%	42.6%
Total	**126**	**417,481**	**115.97**	100%	100%

Table 6: Preliminary per-domain recording counts and total hours in VOICEGIRAFFE, aggregated across the ten bilingual sub-domains. EN and ZH columns report the English-medium and Chinese-medium subsets respectively. A finer sub-category breakdown (including league / show / network provenance) is given in Table 7.

Domain	English (EN)		Chinese (ZH)		Total	
	#	Hours	#	Hours	#	Hours
Sports commentary	13	12.09	14	12.74	27	24.83
E-sports casting	24	23.89	16	14.81	40	38.70
TV dramas	11	9.49	8	6.22	19	15.71
News broadcasting	8	6.12	8	6.82	16	12.94
Interview / podcast	14	13.29	10	10.52	24	23.81
Total	**70**	**64.88**	**56**	**51.11**	**126**	**115.99**

B PRELIMINARY LONG-FORM AUDIO INVENTORY

We list the preliminary bilingual inventory below (126 recordings across five, totalling 116 h). Each entry includes a source identifier, sub-category and reported duration. Availability of each entry at release time will be re-verified prior to camera-ready, and any unavailable recording will be replaced by a matched substitute satisfying the same guidelines in Appendix A.

Table 5 reports the recording-level duration distribution across three buckets, and Table 6 aggregates per-domain recording counts and total hours for the two language halves. A finer sub-category breakdown (including league, show and network provenance, with italic Chinese originals for Chinese-medium entries) is given in Table 7.

C TASK TAXONOMY

VOICEGIRAFFE decouples perception from reasoning through a two-tier ontology. T1 items isolate fine-grained acoustic and semantic cues in a single local window; T2 items force evidence aggregation across non-contiguous windows that span the full recording.

Tier 1: Single-Hop Perception. *Temp. Loc.* (Temporal Localization) requires returning the precise start–end timestamps of a segment described in natural language. *Sem. Cont.* (Semantic Conetent) probes pragmatic intent beyond verbatim transcription, with context drawn from the semantic content within the recording. *Aco. Event* (Acoustic Event) targets non-linguistic signals, such as music genre, environmental category and specific event markers (goal whistles, gunshots, applause). *Paralin.* (Paralinguistic) covers speaker-level traits (age, gender, pitch, timbre, emotion) that are not recoverable from ASR transcripts alone.

Tier 2: Multi-Hop Reasoning. *Causal* (Causal Alignment) evaluates logical coherence and narrative reasoning, requiring models to reconstruct causal chains from provided evidence, identify temporal dependencies, and detect logical inconsistencies. *Evt. Trk,* (Event Tracking) serves as a

Table 7: Preliminary bilingual inventory of VOICEGIRAFFE, broken down by domain and sub-category. "#" counts unique long-form recordings; "Hours" reports the summed duration of those recordings (rounded to two decimals). Chinese-medium (*ZH*) show titles and column names are printed by their English title or conventional romanised name; English-medium show titles follow conventional italic typography. All recordings satisfy the per-domain selection guidelines in Appendix A, and all pass the six-stage audit pipeline (Section 3.2).

Domain	**Lang.**	**Sub-category**	**#**	**Hours**
E-sports	EN	Dota 2 (TI)	8	7.80
	EN	LoL (Worlds / MSI)	8	7.99
	EN	CS:GO / Valorant Major	8	8.10
	ZH	Fighting / MOBA (KOF / Dota / Honor of Kings)	8	6.98
	ZH	LoL: league finals / playoffs	8	7.83
Interview	EN	Multi-speaker podcasts	8	7.63
	EN	Celebrity interviews	6	5.66
	ZH	*Business Interview Record*	7	7.31
	ZH	*Thirteen Invitations*	3	3.21
News	EN	Al Jazeera English	2	1.83
	EN	BBC Radio 4	2	1.63
	EN	BBC World News	2	1.47
	EN	CNN International	2	1.19
	ZH	Local-station news	6	5.65
	ZH	CCTV News	2	1.17
Sports	EN	NBA / NFL (American)	9	8.32
	EN	Other: Boxing / Tennis / Baseball	4	3.77
	ZH	Football (CSL / World Cup / Euro, ZH streams)	8	6.81
	ZH	Other: Boxing / Tennis / Baseball (ZH)	4	3.76
	ZH	Basketball (CBA / NBA, ZH streams)	2	2.17
TV Dramas	EN	*The Big Bang Theory*	7	6.59
	EN	*Friends*	2	1.65
	EN	*Modern Family*	2	1.25
	ZH	*Ode to Joy*	2	1.44
	ZH	*My Own Swordsman*	2	1.40
	ZH	*iPartment*	2	1.51
	ZH	*The Knockout*	2	1.87
Total			**126**	**115.99**

stress test for long-term memory, requiring exhaustive scanning to aggregate sparse, non-explicit events (e.g., "How many times did X occur?" or "Which event never happened?")..

D LEAKAGE AND HALLUCINATION AUDIT PROTOCOL

Textual leakage audit. To ensure that questions require genuine audio understanding, we evaluate every candidate QA pair using a strong text-only LLM (detailed in Appendix H) with access solely to the question and the *union of all clip captions*. Items that can be answered with high confidence from textual captions alone are excluded from the audio-only evaluation set, though retained as a separate text-baseline subset for comparison. For T2 items, we enforce an additional constraint: no single clip caption may individually support the correct answer. If any isolated caption proves sufficient, the item is re-anchored to require stronger multi-hop evidence chains.

Multimodal hallucination audit. To detect caption-level hallucinations, we re-prompt a multimodal LALM (Qwen3-Omni) with only the claimed evidence *audio* windows paired with the proposed answer. The LALM is instructed to verify whether each acoustic claim is substantiated by the audio content. Items containing evidence-inconsistent claims are either re-anchored to verifiable audio segments or removed entirely. This audit explicitly targets hallucinations introduced by the caption generator, which may have described acoustic events, speaker states, or environmental sounds not actually present in the audio signal.

Human expert review. All candidate items that pass the automated audits undergo manual inspection by eight annotators. Reviewers verify question clarity, answer correctness, and evidence sufficiency. Items with ambiguous phrasing, factual errors, trivial distractors, or insufficient audio evidence are flagged for revision or removal. This human-in-the-loop validation step filters out edge cases that automated checks may miss, ensuring high benchmark quality.

E MODEL EVALUATION PROMPT

Auto Model Evaluation Prompt

Role: You are an expert Comprehensive Reasoning Engine, equipped with both advanced logical deduction (IQ) and high emotional intelligence (EQ). Your primary task is to evaluate diverse questions--ranging from rigorous factual problems to nuanced emotional or social scenarios--and solve them accurately.

Instructions:

1. Language Check (EN/ZH Only): Carefully read the provided Question and Options. Evaluate the primary language used in the text.
- If the QA pair is NOT primarily written in either Chinese or English, you MUST output ONLY NULL and terminate immediately.
- If the QA pair is written in Chinese or English, proceed to Step 2.

2. Adaptive Analysis & Step-by-Step Thinking: Determine the core nature of the question and apply the appropriate reasoning strategy:
- *For Logical/Analytical Tasks:* Conduct a rigorous reasoning process. Break down logical hops, verify temporal/causal dependencies, and eliminate distractors based on factual or structural flaws.
- *For Emotional/Contextual Tasks:* Analyze the underlying emotional tone, interpersonal dynamics, speaker intent, and situational empathy. Eliminate distractors that are tone-deaf, out of character, or socially inappropriate, ensuring the chosen answer aligns with natural human communication.

3. Format Restrictions: Your final output must strictly follow one of these two formats (DO NOT output the reasoning steps):
- If unsupported language (Not EN/ZH): Output exactly NULL.
- If supported language: Output exactly ONE correct option letter (e.g., A, B, C, D).



Figure 6: Interface of user study with Chinese.

```
Input:
Question and Options:
```

F Human Evaluation Protocol

We engage eight annotators with native-level proficiency in both English and Chinese. Given the hour-scale listening burden, we improve efficiency by sampling multiple questions per audio recording across all six tasks, allowing comprehensive evaluation from a single listening. Sampling is stratified across domains and languages. Each annotator independently evaluates a randomly assigned subset of 150 questions (75 per language), sampled via stratified random sampling to preserve the benchmark's proportional distribution across domains and languages.

G Per-Model Inference Configurations

Table 8 lists the per-model inference configuration, including maximum supported audio length, the long-audio strategy used for recordings exceeding this length (sliding window, token streaming, chunked inference), context length, decoding temperature and prompt template. All settings follow the models' official recommendations; deviations are noted in the table.

H Model Details

We provide short cards for each evaluated LALM, covering architecture, training regime, modality support, and the public endpoint or weight release used in our evaluation. Full cards are released alongside the code repository; we summarise key details here.

- **Phi-4-Multimodal-Instruct**: multi-modal extension of Phi-4-Mini with modality-specific LoRA adapters over a frozen base LM.
- **Qwen2.5-Omni-7B / Qwen3-Omni-30B-A3B-Instruct / -Thinking**: Thinker–Talker architecture with time-aligned multimodal rotary position embedding.
- **Audio-Flamingo-3**: fully open LALM with AF-Whisper unified audio encoder supporting long audio and chain-of-thought reasoning.
- **MiMo-Audio-7B-Instruct / -Thinking**: open audio instruction model from Xiaomi MiMo.
- **MOSS-Audio-8B-Instruct / -Thinking**: open instruction-tuned / reasoning-enabled audio model from OpenMOSS.
- **MiniCPM-o-4.5-9B**: compact open multimodal model supporting audio input.
- **Qwen3.5-Omni-Flash / -Plus**: closed-source Qwen omni-modal variants.
- **Gemini-2.5-Pro; Gemini-3.1-Pro**: Google's hybrid-reasoning multimodal family; we report the exact version snapshot used in each run.

Table 8: Per-model inference configuration used to evaluate VOICEGIRAFFE. "Max audio" reports the model's officially supported maximum audio context; "Long-audio strategy" is the method applied when the input exceeds this length (sliding window with overlap, token-level streaming, or chunked inference with aggregation). All decoding follows the official implementation. 'Unknown' means that the maximum length can not be found from the technical paper. The following numbers reflect the actual configuration used in our experiments.

Model	Long-audio strategy	Input length
Phi-4-Multimodal-Instruct	Sliding Captioning + Cascaded Inference	40 seconds (30 mins for maximum)
Qwen2.5-Omni-7B	Sliding Captioning + Cascaded Inference	30 seconds (32k length)
Qwen3-Omni-30B-A3B-Instruct	Sliding Captioning + Cascaded Inference	30 seconds (32k length)
Audio-Flamingo-3	Sliding Captioning + Cascaded Inference	600 seconds
MiMo-Audio-7B-Instruct	Sliding Captioning + Cascaded Inference	30 seconds (unknown)
MiMo-Audio-7B-Thinking	Sliding Captioning + Cascaded Inference	30 seconds (unknown)
MOSS-Audio-8B-Instruct	Sliding Captioning + Cascaded Inference	30 seconds (unknown)
MOSS-Audio-8B-Thinking	Sliding Captioning + Cascaded Inference	30 seconds (unknown)
MiniCPM-o-4.5-9B	Sliding Captioning + Cascaded Inference	30 seconds (unknown)
Qwen3.5-Omni-Flash	End-to-End	API (>1hr)
Qwen3.5-Omni-Plus	End-to-End	API (>1hr)
Gemini-2.5-Pro	End-to-End	API (>1hr)
Gemini-3.1-Pro	End-to-End	API (>1hr)

I INFERENCE PARADIGMS

We define the three inference paradigms referenced throughout the main text. All three share the same question set and differ only in how the hour-scale audio reaches the answering model.

(i) E2E (End-to-End). The raw audio and the query are fed directly to the model in a single native pass. This is available only for models whose native context window accommodates hour-scale audio.

(ii) SC (Self-Cascaded). The model itself captions the audio clip by clip (30-second sliding window), and then the *same* model reasons over its own concatenated captions together with the query. Information thus passes through the model's own textual bottleneck.

(iii) LC (LRM-Cascaded). The clip-level captions produced by the target LALM are passed to a strong *external* large reasoning model (LRM), which performs the QA. The external reasoner is a different system from the evaluated LALM and operates purely on text.

The LC pipeline. (1) A LALM slides a fixed window over the hour-scale audio and emits a streaming caption per clip; (2) the per-clip captions are concatenated in temporal order into a single long transcript; (3) this transcript, together with the query, is fed to a strong external large reasoning model that performs the QA.

Motivation. LC is a training-free compromise. Many LALMs are trained only on short-context audio and cannot natively ingest hour-scale input, so LC lets such models still be evaluated and improved on hour-scale understanding without any additional training or fine-tuning. It also offloads long-range memory to the text channel, mitigating the context-length and long-horizon memory limits of the audio encoder.

Scope of interpretation. Because LC relays information through text, LC scores for context-limited models should be read as performance under a *text-relay* setting, where a note-taker LALM converts audio into streaming captions and a separate reasoner operates over the resulting transcript. They are a lower bound on, rather than a direct measure of, a model's native long-audio ability. Reporting the three modes separately is what keeps native audio processing and text-relay reasoning from being conflated.

J OPEN-ENDED EVALUATION

The multiple-choice (CQA) format is our primary protocol because it yields objective and reproducible scoring across models and inference modes, and it is hardened against shortcuts through adversarial distractors, option-order randomization, and a text-only leakage check that removes any item answerable without listening. To test whether this format understates task difficulty, we additionally run an open-ended QA (OQA) study.

In OQA no options are provided: the model receives only the question and a description of the target sub-task, must produce a free-form answer over the hour-scale context, and is scored by an ensemble of LLM judges on semantic correctness rather than option matching. We evaluate two open-source and two proprietary models under the SC (Self-Cascaded) paradigm. Table 9 reports per-category accuracy and Table 10 contrasts overall CQA and OQA accuracy.

Table 9: Open-ended QA (OQA) per-category accuracy (%) by model, under the SC (Self-Cascaded) paradigm, where the model reasons over its own clip captions.

Model	Mode	Temp. Loc.	Sem. Con.	Aco. Evt.	Paralin.	Single-hop	Causal	Evt. Trk.	Multi-hop	Overall
MossAudio	SC	1.60	24.00	22.50	44.57	25.30	7.20	6.00	6.60	19.07
Qwen3Omni	SC	28.80	41.00	38.00	54.57	42.10	26.00	5.60	15.80	33.33
Gemini-2.5-Pro	SC	27.20	92.00	63.50	62.00	59.60	65.60	22.80	44.20	54.47
Gemini-3.1-Pro	SC	59.60	95.50	63.00	62.57	68.50	77.20	40.00	58.60	65.20

Table 10: CQA vs. OQA overall accuracy (%). Δ is the change from the closed-ended to the open-ended format.

Model	Mode	CQA	OQA	Δ
MossAudio	SC	36.67	19.07	-17.60
Qwen3Omni	SC	50.60	33.33	-17.27
Gemini-2.5-Pro	SC	72.60	54.47	-18.13
Gemini-3.1-Pro	SC	75.13	65.20	-9.93

Open-ended generation is substantially harder than the closed-ended format, and every model drops under it. Without candidate options the model must produce the answer entirely on its own across hour-scale context, so its output is more prone to drift or partial matches. For the same reason OQA is more discriminative and a more faithful measure of genuine understanding, and CQA accuracy should be read as a lower bound on task difficulty rather than an upper bound.

K SHORT-CLIP VS. LONG-FORM INPUT

To isolate whether the difficulty of VOICEGIRAFFE stems from long-form context rather than from intrinsic question hardness, we contrast two input regimes on the *identical* question set: (i) the full hour-scale recording, and (ii) a $\sim$30 s window randomly slid within the QA's annotated evidence timestamp. Because multi-hop tasks require jumping across multiple non-contiguous windows, a single fixed clip cannot fairly contain their evidence; we therefore restrict this study to single-hop (T1) tasks, whose evidence lies in one localizable window. Table 11 reports the results.

Table 11: End-to-end accuracy (%) on single-hop tasks: full hour-scale audio vs. a $\sim$30 s evidence-centred clip, on the same question set. Δ is the gain from the clip regime.

Model	Type	Full (hour)	Clip ($\sim$30 s)	Δ
Qwen3Omni	Open-source	36.89	71.95	$+35.06$
Gemini-3.1-Pro	Proprietary	40.53	66.95	$+36.42$

Long-form context is the dominant difficulty driver. Feeding only the $\sim$30 s evidence clip lifts Qwen3Omni from 36.89% to 71.95% on identical questions, and Gemini-3.1-Pro shows a comparable gain. Since the question and the required evidence are unchanged, this gap isolates long-context

degradation as the primary bottleneck: models can answer when the evidence is localized, but fail to retrieve it from hour-scale input. This corroborates the core claim of VOICEGIRAFFE and complements the duration analysis of the main text.

L MEMORY VS. REASONING

Causal Alignment and Event Tracking are deliberately constructed to separate two abilities. Causal Alignment reconstructs relations from salient, locally available cues, whereas Event Tracking requires aggregating and persisting sparse event states across hour-scale spans, and therefore stresses long-range memory rather than reasoning. Their gap acts as a direct memory-vs-reasoning probe across the three inference paradigms. Table 12 reports the comparison, averaged over models within each group.

Table 12: Causal Alignment vs. Event Tracking accuracy (%), averaged over models within each group and inference paradigm. $\Delta = \text{Causal} - \text{Event Tracking}$; $\Delta > 0$ indicates a memory bottleneck, $\Delta < 0$ memory dominance. E2E is reported only for proprietary models, as the open-source models cannot natively ingest hour-scale audio.

Setting	Method	Causal	Event Tracking	Δ	Regime
E2E (native)	Proprietary	58.9	46.1	$+12.8$	Memory bottleneck
Cascaded (SC)	Open-source	43.6	28.7	$+14.9$	Memory bottleneck
	Proprietary	75.0	61.5	$+13.5$	Memory bottleneck
LRM-Cascaded (LC)	Open-source	41.4	43.5	-2.1	Memory dominant
	Proprietary	45.1	57.2	-12.1	Memory dominant
–	Human reference	52.2	61.1	-8.9	Memory dominant

Without external reasoning enhancement, both open- and closed-source models score markedly higher on Causal Alignment than on Event Tracking, whereas humans show the opposite pattern ($\Delta = -8.9\%$). Models reconstruct causal relations from salient cues but fail to preserve sparse states over long contexts, which is the memory-persistence bottleneck discussed in the main text.

The Δ reversal under LC has two distinct causes. For open-source models, Event Tracking genuinely rises (28.7 to 43.5) while Causal Alignment stays flat; LC supplies the long-context capacity these short-context models lack, a real memory gain. For proprietary models, however, LC is worse overall, and the negative Δ is not a memory improvement but an artefact of Causal Alignment falling faster, since the LRM is a text-only reasoner. This is consistent with the finding that text-relay reasoning sacrifices perceptual fidelity, and confirms that it cannot substitute for native long-audio memory in strong capable models.

M AUDIO-TOKEN COMPRESSION

The preceding paradigms either feed the full audio (E2E) or relay it through text (SC/LC). A third option keeps the audio modality but compresses it: rather than routing information through a text intermediate, we retain the model's native continuous audio embeddings and compress them directly. For each 30 s segment we keep only a fraction of audio tokens and feed the concatenated audio embeddings to Qwen3-Omni, with no text relay at any stage. The motivation is concrete: a full recording produces a median of roughly 42K audio tokens (up to 62K). Table 13 reports overall accuracy at varying keep-ratios, with full-audio E2E and the SC text-only baseline as references.

Compressing the audio token stream instead of dropping to text recovers most of the signal, confirming that preserving the audio modality is valuable. It does not yet match text-only cascading (36.2% vs. 50.60% at best), because naive pruning still discards evidence. Closing this gap is a promising direction, and likely requires additional training of the audio LLM to compress rather than drop evidence.

Table 13: Qwen3-Omni overall accuracy (%) when fed compressed audio tokens (no text relay), at varying keep-ratios. Full-audio E2E and the SC text-only baseline are shown for reference.

Setting (audio-token)	**keep 10%**	**keep 25%**	**keep 50%**	**keep 75%**
Segment-wise top-k	30.6	32.0	34.9	36.2
Top-k	31.9	33.0	35.9	35.5
Reference			**Overall**	
Full E2E w/o compression			34.87	
SC text-only			50.60	

N LANGUAGE BIAS: PER-TASK BREAKDOWN

To systematically evaluate cross-lingual generalization, we analyze the performance disparity between English and Chinese inputs, defined as $\Delta = \mathrm{Acc}_{\mathrm{EN}} - \mathrm{Acc}_{\mathrm{ZH}}$. Table 14 reports the full per-task breakdown.

Language bias is generally moderate at the overall level, but exhibits clear model- and task-dependent patterns. All four Chinese-origin models obtain negative overall Δ, indicating a stable Chinese-favored tendency, whereas U.S.-origin models are more heterogeneous: the two Gemini models slightly favor Chinese while the open-source ones favor English. This suggests that language bias is not determined solely by model origin, but is also affected by multilingual training coverage and audio-language alignment quality. At the task level, the largest cross-lingual discrepancies appear in tasks that depend on paralinguistic cues, with U.S.-origin open-source models showing positive paralinguistic gaps of $+11.4\%$ and $+12.6\%$.

These results reveal that cross-lingual robustness in long-form audio understanding is not merely a text-level issue. Even tasks that should rely heavily on acoustic evidence exhibit substantial language-dependent variation, suggesting that current LALMs do not learn fully language-agnostic acoustic representations. Overall, these findings highlight the need for more balanced multilingual audio data and stronger acoustic-language disentanglement to improve cross-lingual generalization.

Table 14: Task-level accuracy (%) on Chinese vs. English. $\Delta = \mathrm{Acc}_{\mathrm{EN}} - \mathrm{Acc}_{\mathrm{ZH}}$. Negative Δ indicates Chinese-favored performance and positive Δ indicates English-favored. *Overall* aggregates all six task categories at the sample level (i.e., $\sum$correct/$\sum$total).

Model	**Temp. Loc.**			**Sem. Con.**			**Aco. Evt.**			**Paralin.**			**Causal**			**Evt. Trk.**			**Overall**		
	ZH	EN	Δ	ZH	EN	Δ	ZH	EN	Δ	ZH	EN	Δ	ZH	EN	Δ	ZH	EN	Δ	ZH	EN	Δ
LALMs from Chinese																					
Qwen3.5-Omni-Plus	51.2	38.4	-12.8	94.0	90.0	-4.0	64.0	57.0	-7.0	72.6	66.3	-6.3	76.8	76.0	-0.8	56.8	56.8	+0.0	68.8	63.6	-5.2
Qwen3-Omni-Instruct	30.4	31.2	+0.8	75.0	63.0	-12.0	52.0	45.0	-7.0	61.7	61.1	-0.6	48.0	43.2	-4.8	51.2	43.2	-8.0	52.9	48.3	-4.7
Qwen2.5-Omni	39.2	36.8	-2.4	49.0	47.0	-2.0	34.0	29.0	-5.0	54.3	46.9	-7.4	42.4	44.0	+1.6	31.2	30.4	-0.8	42.5	39.6	-2.9
MOSSAudio-Instruct	28.0	29.6	+1.6	43.0	38.0	-5.0	43.0	35.0	-8.0	46.9	48.0	+1.1	38.4	32.0	-6.4	25.6	26.4	+0.8	37.7	35.6	-2.1
LALMs from American																					
Gemini-3.1-Pro	81.6	77.6	-4.0	96.0	99.0	+3.0	63.0	51.0	-12.0	69.1	63.4	-5.7	84.0	86.4	+2.4	68.8	70.4	+1.6	76.4	73.9	-2.5
Gemini-2.5-Pro	79.2	66.4	-12.8	94.0	97.0	+3.0	60.0	56.0	-4.0	69.1	67.4	-1.7	74.4	74.4	+0.0	68.0	72.0	+4.0	73.6	71.6	-2.0
Audio-Flamingo-3	34.4	36.8	+2.4	40.0	47.0	+7.0	44.0	37.0	-7.0	36.6	48.0	+11.4	61.6	67.2	+5.6	37.6	43.2	+5.6	42.0	46.9	+4.9
Phi-4-Multimodal	30.4	30.4	+0.0	28.0	32.0	+4.0	27.0	36.0	+9.0	27.4	40.0	+12.6	31.2	28.8	-2.4	22.4	24.8	+2.4	27.7	32.4	+4.7

O LIMITATIONS

While VOICEGIRAFFE represents a significant step toward evaluating hour-scale audio understanding, several limitations warrant acknowledgement and inform future research directions. *(i) Domain Scope.* In this work, we cover five representative domains. However, other real-world scenarios, such as multi-party meetings and medical consultations, remain unexplored. These settings introduce additional challenges, including dense speaker conversation, domain-specific jargon, and privacy-sensitive content; and *(ii) Open-ended Evaluation.* We adopt a multiple-choice format to enable objective and reproducible evaluation. While we mitigate shortcut biases through adversarial distractor design, option randomization, and text-bias verification, this format may not fully capture free-form reasoning capabilities. Open-ended generation evaluation remains a complementary direction.

P CLIP-LEVEL CAPTIONING PROMPT FOR DIVERSE TASKS

The clip-level captioning prompt is fed to **Qwen3-Omni-30B-A3B-Instruct** (Team, 2025c). We reproduce it verbatim here to support exact reproduction. The prompt operates on a single VAD (Bredin et al., 2020) clip at a time (around 30 — 40 seconds).

Clip-Level Captioning Prompt

Task: Output a clear, strictly evidence-based audio description in natural English.

Global Requirements:
1. All content must be strictly based on **audible evidence**.
2. Begin with one sentence (≤40 words) summarising the overall audible content.
3. Provide precise start/end timestamps (in seconds) for every speech/singing segment and every short sound effect.
4. For human speech or singing, output **exact original words** (no paraphrasing). If Chinese, output Simplified Chinese.
5. Assign distinct identity IDs (SPEAKER_001, SPEAKER_002, ...) based strictly on timbre and identity.
6. Do not infer or guess any information that is not clearly audible.
7. The [Global VAD Data] defines the definitive global temporal scope. The initial start and terminal end times are absolute; intermediate timestamps are non-binding. NO_SPEECH denotes no human voice; SPEECH means ≥1 human voice.
8. **Strict Global Timestamp Anchoring:** Base Offset = first start_time from VAD. All timestamps = [Offset + Relative Audio Time], total seconds, 2 decimal places. MM:SS format is a **fatal error**.

Three Description Dimensions:

(a) Human Voices (speech or singing)
For each segment:
- [start, end] with offset
- Transcript (exact original language, Simplified Chinese if Chinese)
- Speaker ID (timbre-based identity)
- **Paralinguistic and voice traits:** gender (male/female), age (child/teenager/young adult/middle-aged/elderly), pitch (low/normal/high), timbre (husky/deep/bright/sweet/warm/nasal/mellow/gravelly/delicate), volume (whisper with breath / soft projection / conversational / loud with strong resonance), speed (slow/fast/moderate), clarity (high/low), physiological states (stuttering/trembling/gasping)
- Emotion **if audible**

Speech Segment Merging Rule: If the same speaker's silent gap ≤2 s, merge into one segment. Split immediately if acoustic state or emotion shifts dramatically.

(b) Short Sound Effects
Only describe clearly perceivable, concretely identifiable short non-speech sounds.
- Categories: Human Non-Speech (coughing, claps, gasps) / Foley or Daily Objects (glass clinking, desk thumps) / Sports or Action Impacts (ball kicks, bat cracks) / In-Game Effects (gunshots, explosions) / Audience (short collective cheers) / Nature (wind gusts, thunder)
- [start, end] with offset
- Concise factual description (≤30 words), including dynamic changes in rhythm or intensity
- Do NOT describe abstract, synthetic, or non-perceivable sounds
- If none: output "none"

(c) Continuous Background Sounds (incl. background music)
- One sentence ≤30 words describing perceivable, identifiable background

- For music: describe style/genre and specific instruments
- If unidentifiable: output "none"
- Ambiguous, synthetic, or non-perceivable background noises are strictly prohibited

Final Format Checklist:
- Overall Summary: one sentence, $\leq$60 words
- Timestamps anchored to absolute range (first/last VAD time)
- Speaker Identity Consistency: timbre over emotion; same voiceprint = same ID
- Speaker ID Minimisation: identify minimum distinct speakers
- Paralinguistic Traits: all required dimensions explicitly defined
- Strict Formatting: section names, ordering, punctuation mirror the example

Q Overall Caption Synthesis Prompt

All clip-level captioning prompts with the origin audio are fed to **Qwen3.5Omni** (Team, 2025b) to get the overall caption with global context.

Overall Caption Synthesis Prompt (Verbatim)

Role: You are an expert in audio content summarisation. Your task is to perform a two-level processing on the holistic audio input and all chronological audio clip captions provided.

Two-Level Processing:
1. **Global Overview Summary:** Use $\leq$300 words to concisely summarise the core genre, semantic content, thematic thread, conceptual diversity, conclusion, character portrayal, and key acoustic features of the entire long audio sequence.
2. **Paired Summary Sequence:** Group and merge the clips to generate a complete, end-to-end concatenated summary. Each item must fuse vocal delivery (content + attribute), key sound effect, and background atmosphere into a unified, logically connected scene description. Ensure smooth transitions between the two clips.

Global Summary Rules:
Write a highly integrated narrative summary within 300 words: Organically fuse the core semantic thread (genre, content, thematic thread, conceptual diversity, conclusion), character dynamics (NO IDs; use names if mentioned, or specific attributes/identities inferred from context), and key acoustic features (intonation shifts, background noise, special sound effects) into a logically coherent, deeply insightful holistic perspective. Lists or chronological logs are strictly prohibited. Extract the "soul" and "narrative atmosphere" of the audio from an expert's point of view.

Grouping Protocol:
1. **Strict Non-Overlapping Pairing:** 1st Clip + 2nd Clip = 1st Summary; 3rd Clip + 4th Clip = 2nd Summary; and so on $(2k-1,\ 2k)$.
2. **Residual Handling:** If $N \bmod 2 \neq 0$, the final remaining clips constitute an independent final group.
3. **Input Order:** Maintain chronological order; reordering or skipping is strictly prohibited.

Timestamp Rules:
1. **Range Calculation:** Start time = earliest start of the first clip in the group; end time = latest end of the second clip.
2. **Output Format:** [Second 1, Second 2].

Speaker Reference Guidelines:
1. **No Anonymous IDs:** Identifiers like SPEAKER_XX or Speaker0 must not appear in output.

2. Cross-Clip ID: An identical speaker ID across different clips does not guarantee the same person. Infer based on attributes and transcript.
3. Reference Guidance: Infer a reasonable identity based on acoustic attributes (gender/age/timbre/tone) and contextual content, then describe using natural language.
- Example: David broadcasting news → David
- Example: Male, middle-aged, bass voice → News Anchor (middle-aged male, deep voice, authoritative tone)

4. Brief References: When the same speaker appears again within the same clip, use simplified references to avoid repeating full attribute descriptions.

Semantic Content Granularity:
1. Extract, Don't Copy: Direct copying of semantic content is prohibited. Retain core information density including main topics, central arguments, key temporal actions, and topic transition nodes.
2. Compression: Most content via indirect narration. Small amount of direct quotation allowed as evidence for key propositions.
3. Temporal Locality: Each clip summary must only focus on content within its current paired time interval. Do not bridge across paragraphs in the future.

Sound Effects & Environment Handling:
1. Narrative Integration: Weave significant sound effects directly into the narrative (e.g., "Thunder crashed as the argument escalated"). Do not redundantly list sounds.
2. Atmosphere Only: Only mention background noise when there is an atmospheric shift. If both clips are clearly marked "none," omit any sound effect commentary.

Final Checklist:
- First, output a summary of up to 300 words, starting with SUMMARY:
- Each group (two merged captions) occupies its own line: [$2k-1$'s start, $2k$'s end] Refined summary
- Language Style: Academic written language, rigorous and concise, natural expression, no literary exaggeration
- Language Requirement: Output language follows semantic content of the audio (Chinese semantics → Chinese output; English → English). Do not translate mixed phrases

Output Format:

```
SUMMARY: [Your narrative summary here]
[1st Start, 2nd End] Refined semantic summary
[3rd Start, 4th End] Refined semantic summary
.....
```

R QA Generation Templates

We employ **Gemini-3.1-Pro** (Team, 2025a) as the QA generator for both tiers. Below we present the complete set of prompts for all six tasks.

R.1 QA Template

1. Tier 1: Temporal Grounding:

Tier 1: QA Generation Prompt (Temporal Grounding)

Role: You are an expert question-answer (QA) pair curator specializing in Audio Understanding for long-form audio.

Input Format:
You will receive:

1. An **overall caption** summarizing the entire long-form audio file (domain, topic, key participants, and general narrative arc).
2. A sequence of **clip captions** segmented from the same audio.
- Format: [Start_Sec, End_Sec] Clip Caption X: ...
- Content: Includes speaker attributes, speech transcripts, sound effects, and background sounds.

Critical Rule:
No cross-clip leakage: When generating a question about a specific anchor, you must NOT reveal information contained in clip captions from other time ranges. The question and answer options must be answerable solely from the target clip's audio content.
Anchor-aware contextualization: Use the overall caption to determine the anchor's global position (e.g., "the second goal", "the first substitution", "the turning point after halftime") and its narrative role. Combine this global awareness with the clip caption's local detail to formulate the question--but never expose details that only appear in adjacent or distant clips.

Task: Temporal Grounding (TimeStamp Perception)
Generate a multiple-choice question that tests the model's ability to precisely locate the continuous second-based time segment [Start_Sec, End_Sec] where a specific multimodal event (semantic or acoustic) occurs.

1. Question Generation & Semantic Anchor Strategy
Every question MUST implement a robust Semantic Anchor-to-Target logic in a single cohesive sentence ($\leq$100 words). CRITICAL: Strictly FORBIDDEN from including any explicit timestamps in the question text.

Semantic Anchor (High Information Entropy): Use objective anchors: **1)** Ordinals/Statistics (e.g., "the third dragon contest", "the second news item"), **2)** Extremes/Specific Conditions (e.g., "the largest score gap", "the loudest crowd reaction"), or **3)** Structural Event Shifts (e.g., "when the substitute goalie entered", "when the background music suddenly stopped"). Absolutely no vague pronouns or relative temporal markers.

2. Option Design & Anti-Cheating (Critical)
Provide exactly 4 options (A-D). EVERY option MUST be strictly a timestamp interval string using absolute seconds-based INT-point format, i.e., "[Start_Sec, End_Sec]" with absolutely NO other words.

1. Max Boundary Limit: Strictly FORBIDDEN from generating any timestamps that exceed the maximum duration (end time) of the provided captions.
2. Zero-Intersection: All 4 intervals MUST be strictly mutually exclusive with absolutely zero overlap.
3. Deceptive Pairing & Macro-Temporal Gap: Structure the 4 options into two distinct pairs:
- *Pair 1 (Local):* The correct option and one nearby, non-overlapping distractor.
- *Pair 2 (Global):* Two proximally close distractors located far away from Pair 1. CRITICAL: Pair 2 MUST be separated from Pair 1 by $\geq$120 seconds.

4. Flexible Temporal Margin: Options do not need to strictly match the exact start/end boundaries. Apply a reasonable second-level margin (padding start and/or end) to fully encapsulate the target event and create a natural listening window.

3. Content & Attribution Rules
Dominant Language Alignment: The generated QA pair MUST strictly align with the dominant spoken language of the audio.
De-identification: NEVER reference speakers using raw annotation IDs. Identify subjects dynamically using physical/acoustic attributes or roles.

Output Format:

```
{
  "question": "<anchored question, <= 100 words>",
  "options": {
    "A": "[Start_Sec, End_Sec]",
    "B": "[Start_Sec, End_Sec]",
    "C": "[Start_Sec, End_Sec]",
    "D": "[Start_Sec, End_Sec]"
  },
  "answer": "<A/B/C/D>",
  "start_time": "INT",
  "end_time": "INT",
  "rationale": "<= 50 words, evidence-based>"
}
```

If Invalidated: If the clip caption lacks sufficient correlation to construct a valid QA pair, output exactly: NULL

2. Tier 1: Semantic Content Understanding:

Tier 1: QA Generation Prompt (Semantic Content)

Role: You are an expert question-answer (QA) pair curator specializing in Audio Understanding for long-form audio.

Input Format:
You will receive:
1. An **overall caption** summarizing the entire long-form audio file (domain, topic, key participants, and general narrative arc).
2. A sequence of **clip captions** segmented from the same audio.
   - Format: [Start_Sec, End_Sec] Clip Caption X: ...
   - Content: Includes speaker attributes, speech transcripts, sound effects, and background sounds.

Critical Rule:
No cross-clip leakage: When generating a question about a specific anchor, you must NOT reveal information contained in clip captions from other time ranges. The question and answer options must be answerable solely from the target clip's audio content.
Anchor-aware contextualization: Use the overall caption to determine the anchor's global position (e.g., "the second goal", "the first substitution", "the turning point after halftime") and its narrative role. Combine this global awareness with the clip caption's local detail to formulate the question--but never expose details that only appear in adjacent or distant clips.

Task: Deep Pragmatic & Intent Perception
Generate a multiple-choice question that evaluates a model's ability to transcend surface-level ASR, forcing it to infer unstated implications, synthesize conversational dynamics, and decode complex speaker intents within a noisy or complex auditory context.

1. Question Generation & Semantic Anchor Strategy
Every question MUST implement a robust Semantic Anchor-to-Target logic in a single cohesive sentence ($\leq$100 words). CRITICAL: Strictly FORBIDDEN from including any explicit timestamps in the question text.

Semantic Anchor (High Information Entropy): Use objective anchors: **1)** Ordinals/Statistics (e.g., "the third dragon contest"), **2)** Extremes/Specific Conditions (e.g., "the largest score gap"), or **3)** Structural Event Shifts (e.g., "when the substitute goalie entered"). Absolutely no vague pronouns or relative temporal markers.

Time Range Selection: Do not default to targeting only the most information-dense time range. Dynamically sample meaningful content across the entire provided audio caption.

The Semantic Target: Explicitly query the meaning, intent, or core extraction. Randomly alternate between these inquiry types:

- *Type A (Stance/Attitude Decoding):* What is the speaker's true underlying attitude, especially when disguised by politeness, sarcasm, or diplomatic language?
- *Type B (Event Comprehension):* Ask for the comprehension of high-level semantic concepts or extraction of specific factual outcomes.
- *Type C (Macro Summarization):* Tests the ability to synthesize and abstract narrative, thematic, or strategic information surrounding the anchored event.

Causal & Thematic Cohesion: The target Semantic Content MUST organically extend the logical or narrative trajectory established by the anchor, strictly prohibiting jarring shifts to unrelated trivia.

2. Option Design & Anti-Cheating (Critical)
Generate exactly 4 options (A-D). Dynamically shuffle the correct answer position.

"True to World, False to Audio" (Critical): All 3 incorrect options MUST be 100% factually correct according to external domain knowledge. The ONLY reason an option is incorrect is that the speaker did not explicitly articulate that point in the anchored temporal window. A domain expert MUST NOT be able to eliminate any option WITHOUT listening to the audio.

Homogeneity & Parity: All options MUST revolve around the EXACT SAME core entities and relationships. Identical sentence structure, emotional valence, length, and level of mechanistic detail.

Structured Hard Negatives (Mutation Recipes):
1. *Unspoken Domain Truth (Plausible Hallucination):* State a 100% accurate, highly specific domain fact about the target entities that was completely ABSENT from the audio segment.
2. *Causal Decoupling (Right Outcome, Wrong Valid Reason):* Keep the exact outcome but justify it using a different valid domain mechanic the speaker did NOT cite.
3. *Cross-Temporal Contamination (Right Fact, Wrong Anchor):* Use a genuine fact expressed by the speaker, but from a COMPLETELY DIFFERENT timestamp that does not align with the Semantic Anchor.

3. Content & Attribution Rules
Dominant Language Alignment: The generated QA pair MUST strictly align with the dominant spoken language of the audio.
De-identification: NEVER reference speakers using raw annotation IDs. Identify subjects dynamically using physical/acoustic attributes or roles.

Output Format (Type B shown as example):

```
{
  "qa_type": "TYPE B",
  "question": "<anchored question, <= 90 words>",
  "options": {
    "A": "<extracted content>",
    "B": "<extracted content>",
    "C": "<extracted content>",
    "D": "<extracted content>"
  },
  "answer": "<A/B/C/D>",
  "start_time": "float",
  "end_time": "float",
  "rationale": "<= 50 words justification>"
}
```

If Invalidated: If the clip caption lacks sufficient correlation to construct a valid QA pair, output exactly: NULL

3. Tier 1: Acoustic Event Understanding:

Tier 1: QA Generation Prompt (Acoustic Event)

Role: You are an expert question-answer (QA) pair curator specializing in Audio Understanding for long-form audio.

Input Format:
You will receive:
1. An **overall caption** summarizing the entire long-form audio file (domain, topic, key participants, and general narrative arc).
2. A sequence of **clip captions** segmented from the same audio.
   - Format: [Start_Sec, End_Sec] Clip Caption X: ...
   - Content: Includes speaker attributes, speech transcripts, sound effects, and background sounds.

Critical Rule:
No cross-clip leakage: When generating a question about a specific anchor, you must NOT reveal information contained in clip captions from other time ranges. The question and answer options must be answerable solely from the target clip's audio content.
Anchor-aware contextualization: Use the overall caption to determine the anchor's global position and its narrative role. Combine this global awareness with the clip caption's local detail to formulate the question--but never expose details that only appear in adjacent or distant clips.

Task: Acoustic Event Detection (AED) & Environmental Perception
Generate a multiple-choice question that tests the model's ability to identify non-verbal signals (sound effects, music, background noise, and acoustic transitions) within a targeted time window.

1. Question Generation & Cross-Modal Trigger Strategy
Generate a single cohesive sentence ($\leq$100 words) querying a specific non-verbal audio event. CRITICAL: Strictly FORBIDDEN from including any explicit timestamps.

Gatekeeper (Salience Prerequisite): ABORT generation if the target segment lacks a distinct, noticeable non-verbal sound. Do not force questions about "silence" in a naturally quiet room.

Acoustic-Centric Anchor: Contextualize the scene using objective markers (Ordinals, Extremes, or Structural Event Shifts). Forbidden: vague pronouns (e.g., "this time", "that event").

Cross-Modal Target & Trigger (Critical): The queried sound MUST organically extend the anchor's narrative. Randomly apply ONE of three psychological triggers:
   - *Trigger A (Epistemic Gap):* Query a specific **Foreground SFX** that acts as blind-spot info to confirm an unseen action or unspoken outcome.
   - *Trigger B (Anomaly/Intrusion):* Query a **Dynamic Transition or disruptive sound** that suddenly breaks the expected semantic scene.
   - *Trigger C (Emotional Resonance):* Query a shift in **Background Music or Atmosphere** that amplifies the narrative climax or emotional tension.

Causal & Thematic Cohesion: The target AED event MUST organically extend the logical, causal, or narrative trajectory established by the semantic anchor, strictly prohibiting jarring shifts to unrelated acoustic trivia.

2. Option Design & Anti-Cheating (Critical)
Provide 4 options (A-D) that are indistinguishable without audio-evidence.

Domain Relevance: All distractors must be contextually plausible for the setting (e.g., stadium cheers for sports, keyboard clicks for gaming). Prohibit out-of-domain sounds.

Plausible Hallucinations:
• *Temporal Displacement:* Use sounds appearing in other parts of the audio but not here.
• *Logical Consequence:* Use sounds naturally expected from the transcript/action (e.g., "rapid clicking" during a fight) even if not present in the audio.

Anti-Guessing: Defeat common-sense deduction. If the scene is a "high-energy fight," all distractors must also describe "high-energy" sounds. Do not allow "quiet" distractors to make the "loud" ground truth obvious.
No Trivial Rejections: Do not use absurd or out-of-context sounds.
Format Parity: Keep all options structurally identical.

3. Content & Attribution Rules
Dominant Language Alignment: The generated QA pair MUST strictly align with the dominant spoken language of the audio.
De-identification: NEVER reference speakers using raw annotation IDs. Identify subjects dynamically using physical/acoustic attributes or roles.

Output Format:

```
{
  "question": "<anchored question, <= 80 words>",
  "options": {
    "A": "<sound/music/environmental event>",
    "B": "<sound/music/environmental event>",
    "C": "<sound/music/environmental event>",
    "D": "<sound/music/environmental event>"
  },
  "answer": "<A/B/C/D>",
  "start_time": "float",
  "end_time": "float",
  "rationale": "<= 50 words>"
}
```

If Invalidated: If the clip caption lacks sufficient non-verbal acoustic events to construct a valid QA pair, output exactly: NULL

4. Tier 1: Paralinguistic Understanding:

Tier 1: QA Generation Prompt (Paralinguistic Perception)

Role: You are an expert question-answer (QA) pair curator specializing in Audio Understanding for long-form audio.

Input Format:
You will receive a sequence of **clip captions** from a long-form audio.
• Format: [Start_Sec, End_Sec] Clip Caption X: ...
• Content: Includes speaker attributes, speech transcripts, sound effects, and background sounds.

Task: Paralinguistic Perception
Generate a multiple-choice QA pair that rigorously tests a native audio model's ability to directly process paralinguistic features (age, gender, pitch, emotion, timbre, and speech rate). The QA pair MUST be impossible to answer using only an ASR transcript.

1. Question Generation & Semantic Anchor Strategy
Every question MUST implement a robust Semantic Anchor-to-Target logic in a single cohesive sentence ($\leq$100 words). CRITICAL: Strictly FORBIDDEN from including any explicit timestamps.

Semantic Anchor: Smoothly embed a macro-level, descriptive reference to a clear contextual hook (e.g., a major thematic shift, a specific debate resolution, or a significant acoustic event). The anchor must

read like a natural human inquiry, not a robotic search filter. Avoid mechanically quoting highly specific sentences just to set an anchor.

The Paralinguistic Target: Formulate the question to evaluate the speaker's acoustic delivery while completely ignoring the literal text transcript. Randomly select ONE dimension:
- *Age / Gender:* Identify the speaker's demographic profile based strictly on vocal markers.
- *Emotion:* Pinpoint the speaker's true emotional state as conveyed through vocal tone and energy.
- *Pitch & Speed:* Describe dynamic acoustic features such as pitch variations, intonation curves, or speech rate.
- *Timbre:* Classify the speaker's unique voice texture or vocal quality (e.g., breathy, hoarse, resonant, clear).

2. Option Design & Anti-Cheating (Critical)
Generate exactly 4 options (A-D). Dynamically shuffle the correct answer position.

Rule 1: Contextual Plausibility (The "Common Sense" Trap -- Critical): All 4 options MUST be highly plausible given the broader acoustic environment, speaker role, and domain. Forbidden: simple antonyms or out-of-context distractors (e.g., if context is intense esports broadcast, do NOT use "calm and slow" as a distractor).

Rule 2: Strict Category Consistency: All 4 options MUST belong exclusively to the queried paralinguistic dimension (e.g., all describe timbre, or all describe pitch). Do not mix dimensions.

Rule 3: The "Text-Blind" Trap (For Emotion/Intent): When querying emotion, include at least one distractor that aligns perfectly with the literal semantic text but completely fails to capture the actual paralinguistic tone (e.g., describing a sarcastic "great job" as "genuine praise").

Rule 4: Homogeneity & Zero-Leakage: All options must be structurally identical, roughly the same length, and devoid of arbitrary entities not mentioned in the audio. Neither options nor rationales may contain or hint at timestamps.

3. Content & Attribution Rules
Dominant Language Alignment: The generated QA pair MUST strictly align with the dominant spoken language of the audio.
De-identification: NEVER reference speakers using raw annotation IDs. Identify subjects dynamically using physical/acoustic attributes or roles.

Output Format:

```
{
  "question": "<anchored question, <= 100 words>",
  "options": {
    "A": "<paralinguistic descriptor>",
    "B": "<paralinguistic descriptor>",
    "C": "<paralinguistic descriptor>",
    "D": "<paralinguistic descriptor>"
  },
  "answer": "<A/B/C/D>",
  "start_time": "float",
  "end_time": "float",
  "rationale": "<= 40 words justification>"
}
```

If Invalidated: If the clip captions lack any distinct paralinguistic information to construct a valid QA pair, output exactly: NULL

4. Tier 2: Causal Alignment:

Tier 2: QA Generation Prompt (Causal Alignment)

Role: You are an expert question-answer pair curator specializing in Multi-hop Contextual Aggregation for super long-form audio caption (e.g., esports broadcasts, sports commentary, multi-hour podcasts).

Input Format:
You will receive:
1. An **overall caption** summarizing the entire long-form audio file (domain, topic, key participants, and general narrative arc).
2. A sequence of **clip captions** segmented from the same audio.
   • Format: [Start_Sec, End_Sec] Clip Caption X: ...
   • Content: Includes speaker attributes, speech transcripts, sound effects, and background sounds.

Critical Rule:
No cross-clip leakage: When generating a question about a specific anchor, you must NOT reveal information contained in clip captions from other time ranges.
Anchor-aware contextualization: Use the overall caption to determine the anchor's global position and narrative role. Combine global awareness with local detail--but never expose details from adjacent or distant clips.

Task: Multimodal & Multi-Hop QA Design
Design a strict MCQ requiring true multi-hop reasoning across a super long-form audio timeline. The question MUST NOT be answerable by keyword search, recency bias, or single-clip extraction. Force the model to compute an answer by synthesizing implicit concepts derived from BOTH semantic content and paralinguistic clues (speaker emotion, intent). Ignore subtle background sound/music unless overwhelmingly obvious and critical to the narrative.

1. Target Outcome Categorization
The target Outcome MUST represent a long-term structural shift or culmination, never a single localized event. Select ONE category:
   • *Category 1 (Macro-Semantic Resolution):* A definitive conclusion deduced solely by aggregating multiple jointed factual premises across the timeline.
   • *Category 2 (Relational Dynamics):* A hidden shift in conversational alignment or dominance, overriding literal agreements through tone shifts or interaction patterns across turns.
   • *Category 3 (Thematic Evolution):* A structural pivot in the underlying pragmatic focus. The core abstract topic transforms completely despite consistent surface-level domain vocabulary.
   • *Category 4 (Latent Intent):* An unspoken strategic goal executed, abandoned, or concealed, revealed only when cross-referencing explicit claims with contradictory tone/actions.

2. Question Design & Trajectory Extraction (Diversity Enforced)
The "Anti-Frankenstein" Rule (Strict Causality vs. Chronology):
Distinguish true CAUSATION from mere CHRONOLOGY. Do NOT stitch together unrelated sequential events. Every hop must explicitly and undeniably trigger the next. If a natural, organically strong causal chain ($\geq$2 hops) does not exist, REJECT and output NULL.

Trajectory Extraction (Quality > Quantity): Scan backwards from the outcome to extract necessary antecedent events. A rock-solid 2-hop chain is infinitely better than a fabricated 4-hop chain. Do not force hops if the audio only supports fewer.

High-Entropy Anchor Selection: Locate the final Outcome within the assigned Temporal Anchor (TA). Introduce this outcome using an objective, high-information anchor: 1) Ordinals/Statistics, 2) Extremes/Specific Conditions, or 3) Structural Event Shifts. Absolutely NO vague pronouns.

The Causal Inquiry (Natural Vibe): Smoothly embed the anchor into a natural human inquiry asking HOW or WHY this outcome occurred.

Zero AI Narrative Flavor & Forbidden Words: Strip all robotic meta-language. Do NOT use "causal chain," "sequence of events," "historical trajectory," or "multiple hops." Strictly FORBIDDEN from including explicit timestamps or dashes in the question text.

Temporal Anchor (TA) Enforcement:
TA 1: 0%-40% | TA 2: 40%-50% | TA 3: 50%-60% | TA 4: 60%-70% | TA 5: 70%-80% | TA 6: 80%-90% | TA 7: 90%-100%.

3. The "Event Bank" & Option Design (Critical)
To eliminate redundant text and force pure logical reasoning, construct a shared "Event Bank" followed by numeric sequence options.

Step 1 -- Atomic Event Bank (6-8 Items): Generate concise, single-action events labelled [1] to [8].
- Length: Strictly $\leq$15 words per item.
- Content: 3-4 True Causal Nodes, 2 Temporal Decoys (real but unrelated), 2 Plausible Hallucinations (logical but fake).
- Constraint: Do NOT include the final outcome in the bank.

Step 2 -- Numeric Sequence Options (A, B, C, D):
- Format: [x] $\rightarrow$ [y] $\rightarrow$ [z]. (No text allowed in options).
- Orthogonal Starting Points: The first number of ALL 4 options MUST be unique. Zero prefix sharing.

Step 3 -- The 3 Logical Traps:
- *Distractor 1 (Chronology Trap):* Correct numbers, but swapped order.
- *Distractor 2 (Trojan Horse):* Replaces one mid-chain node with a plausible hallucination number.
- *Distractor 3 (Misaligned Reality):* A logical chain built from Decoy numbers that did not actually cause the outcome.

Step 4 -- Orthogonal Starting Points (Critical): ALL 4 numeric options MUST begin with a COMPLETELY DISTINCT starting number. Zero tolerance for prefix sharing.

Step 5 -- Language Alignment: The Event Bank MUST strictly match the dominant spoken language of the provided audio captions.

Output Format (Category 3, TA 4 shown as example):

```
{
  "OUTCOME_type": "Category 3",
  "TQ_type": "TA 4",
  "question": "<natural HOW/WHY question,
    <= 100 words>",
  "event_bank": {
    "1": "<event, <= 20 words>",
    "2": "...", "3": "...",
    "4": "...", "5": "...", "6": "..."
  },
  "options": {
    "A": "[x] -> [y] -> [z]",
    "B": "[a] -> [b] -> [c]",
    "C": "[d] -> [e] -> [f]",
    "D": "[g] -> [h] -> [i]"
  },
  "answer": "<A/B/C/D>",
  "trajectory_evidence": [
    {"hop_index": 1, "start_time": "...",
     "end_time": "...",
     "event_description": "..."},
    {"hop_index": 2, ...},
    {"hop_index": k, ...}
  ],
  "outcome_start_time": "float",
```

```
  "outcome_end_time": "float",
  "rationale": "<verify correct chain +
    debunk Trojan Horse, <= 200 words>"
}
```

If Invalidated: If the provided captions lack sufficient causal correlation to construct a valid multi-hop QA pair, output exactly: NULL

6. Tier 2: Event Tracking:

Tier 2: QA Generation Prompt (Event Tracking)

Role: You are an expert question-answer pair curator specializing in Multi-hop Contextual Aggregation for super long-form audio caption (e.g., esports broadcasts, sports commentary, multi-hour podcasts).

Input Format:
You will receive:
1. An **overall caption** summarizing the entire long-form audio file (domain, topic, key participants, and general narrative arc).
2. A sequence of **clip captions** segmented from the same audio.
   - Format: [Start_Sec, End_Sec] Clip Caption X: ...
   - Content: Includes speaker attributes, speech transcripts, sound effects, and background sounds.

Critical Rule:
No cross-clip leakage: When generating a question about a specific anchor, you must NOT reveal information contained in clip captions from other time ranges. The question and answer options must be answerable solely from the target clip's audio content.
Anchor-aware contextualization: Use the overall caption to determine the anchor's global position (e.g., "the second goal", "the first substitution", "the turning point after halftime") and its narrative role (e.g., momentum shift, climax, resolution). Combine this global awareness with the clip caption's local detail to formulate the question--but never expose details that only appear in adjacent or distant clips.

Task: Event Tracking
Design a multi-hop MCQ for long-form audio that mandates global synthesis over local retrieval. Select a structural outcome and execute either **Pattern A: Event Counting** (tracking 2-5 occurrences of an event) or **Pattern B: Causal Logic Audit** (identifying a logically plausible but factually absent "Trojan Horse" among 2-4 real causal factors). All designs must follow the **Blind Question Protocol:** use high-entropy **semantic anchors** to lock the context while strictly omitting timestamps or evidence leaks.

1. Question Design: Dual-Task Protocols
The "Human-Pointer" Anchor: Lock the start/end of the inquiry using high-entropy, objective markers that any listener can confidently pinpoint (e.g., "After the second goal," "Once the host changed the subject," "When the game paused for a technical timeout").
Conversational Phrasing: Frame questions as natural human curiosity. Do NOT use robotic framing. Use intuitive verbs (e.g., "hear," "notice," "call," "mention"). Absolutely no vague pronouns or relative temporal markers.

Select ONE Pattern to apply:

Pattern A: Domain-Aware Event Counting
- The "Non-Explicit" Rule: The count MUST NOT be directly stated in a single caption or global summary. The model must scan the full timeline to aggregate sparse cues.
- Domain Specificity: Focus on events that are implied by caption but not explicitly numbered.

• Option Rule: MUST be very concise with a specific counting number OR ONLY counting number.

Pattern B: Multi-hop Logical Audit (Error Identification)
• The "3-True-1-False" Structure: Generate 4 options. Three MUST be verified, non-contiguous causal fragments. The fourth (Target Answer) MUST be a "Trojan Horse"--a description that uses real domain vocabulary and sounds plausible, but is irrelevant or logically disjointed from the actual causal chain.

2. Logical Rigor & Temporal Anchors
The "Occam's Razor" Principle (Strict Entailment):
• No Logical Leaps: The correct path must be an inescapable necessity.
• The "Bridge" Rule: Distant hops must be linked by concrete audible evidence (specific tone shifts, ambient changes, or repetitive phrases), not hallucinated causality.

Temporal Anchor (TA) Enforcement:
The Outcome or Final Count must strictly occur within the assigned window:
TA 1: 0%-40% | TA 2: 40%-50% | TA 3: 50%-60% | TA 4: 60%-70% | TA 5: 70%-80% | TA 6: 80%-90% | TA 7: 90%-100%.

3. Option Design & Anti-Cheating (Critical)
Rule 1: Vocabulary Mirroring (Psychological Camouflage)
All options must utilise entities, terminology, and settings extracted directly from the transcript.

Rule 2: Structural Symmetry (The "Clone" Rule)
All four options must maintain identical syntax, complexity, and length.

Rule 3: Pattern-Specific Implementation
• **Pattern A (Event Counting):** Options must combine a unique numerical value with a situational qualifier. All numerical distractors must be distinct, tightly clustered near the correct total (±1 increments).
• **Pattern B (Logical Audit):** Construct a "3-True-1-False" set. Each option must be an internally logical description forcing the model to synthesise 2-4 non-contiguous events for verification. The False Option must be professionally plausible but factually disjointed from the actual causal progression.

Rule 4: Language Alignment
The entire QA pair must strictly align with the dominant language of the source captions.

Output Format (Pattern B, TA 4 shown as example):

```
{
  "qa_type": "Pattern B",
  "question": "<natural question, <= 100 words>",
  "options": {
    "A": "<factual state>",
    "B": "<factual state>",
    "C": "<factual state>",
    "D": "<factual state>"
  },
  "answer": "<A/B/C/D>",
  "trajectory_evidence": [
    {"hop_index": 1, "start_time": "...",
     "end_time": "...",
     "event_description": "..."},
    {"hop_index": 2, ...},
    {"hop_index": k, ...}
  ],
```

```
  "distractor_rationale": "<explain traps,
    <= 200 words>"
}

If Invalidated:  If the provided captions lack sufficient causal
correlation to construct a valid multi-hop QA pair, output exactly:
NULL
```