LONGAUDIOSPAN: SPANNING THE DURATION AND DEPTH OF AUDIO COMPREHENSION

Anonymous authors
Paper under double-blind review

ABSTRACT

General audio comprehension now covers speech, sound, and music over durations from seconds to hours, driven by large audio-language models (LALMs) that are increasingly omni-modal. Yet the benchmarks that test them still rely on clips of seconds, where scores saturate and models converge; recent long-form efforts extend duration but evaluate long audio much as short clips are. We introduce **LongAudioSpan**, a benchmark that spans both duration and depth: it pairs audio from 10 minutes to over 2 hours with 3,240 questions across three cognitive levels, namely perception, understanding, and reasoning. Two paths supply the questions, differing in how question content is sourced and how ground truth is obtained. Native QA extracts questions from the audio's content, posing each as a multiple-choice item and an open-ended one graded by detailed rubrics. Anchor QA instead injects ground truth, planting acoustic anchors into the audio and building a perception-to-reasoning chain scored only to the first error. A fully automated pipeline constructs every item through structured captioning, QA generation, and adversarial critic feedback. Evaluating 12 LALMs on LongAudioSpan, we find the hard part comes before reasoning: distilling a few relevant facts from a long, redundant signal. This difficulty grows with audio length and falls hardest on perception, especially temporal grounding. LongAudioSpan is available at https://huggingface.co/datasets/audiospan/LongAudioSpan.

1 INTRODUCTION

Audio intelligence has expanded from specialized recognition tasks to encompass general audio-language comprehension. Early work focused on independent tasks such as speech recognition (Radford et al., 2023), sound classification (Chen et al., 2023), and audio-text alignment (Wu et al., 2023), each relying on specialized architectures with fixed output formats. More recently, large audio-language models (LALMs), from dedicated audio specialists to omni-modal systems that also handle vision and text, unify diverse audio capabilities through open-ended text generation. Over the past three years, these models have evolved along several axes: from speech-only (Zhang et al., 2023) to unified audio covering speech, sound, and music (Chu et al., 2023), from simple recognition to open-ended understanding (Chu et al., 2024), and from short clips (Deshmukh et al., 2023; Tang et al., 2024) to multi-hour audio (Qwen Team, 2026).

Benchmarks for evaluating LALMs have emerged (Sakshi et al., 2025; Ma et al., 2025; Wang et al., 2026), covering multidimensional assessment across different audio modalities, with task types from basic perception to deeper reasoning. As model capabilities improve, top accuracy on these benchmarks has risen from around 50% to over 80% within a year of publication, and the leading LALMs now cluster within a few points of one another. These benchmarks, however, operate on audio clips measured in seconds: the relevant evidence sits in a short span, and answering does not call for sustained attention across a long context.

Recent efforts have begun to fill this gap by extending audio duration to the minute level (Yang et al., 2026b; Luo et al., 2026; Ghosh et al., 2025b; He et al., 2025; Shao et al., 2026) or beyond (Ahia et al., 2025; Ye et al., 2026) (Table 1). Yet even with extended duration, these benchmarks share limitations across four dimensions: **1) Task framework**: tasks follow fixed schemes over a predetermined set of types, and cognitive depth is mostly flat. **2) Question format**: most rely on multiple-choice or fixed-answer questions, where models can achieve inflated scores through shortcut techniques rather than

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.

Table 1: Comparison of LongAudioSpan with existing audio comprehension benchmarks.

		Task Framework		Question Format		Data Construction		Quality Assurance	
Benchmark	**Duration**	**Depth**	**Scheme**	**Choice**	**Open**	**Source**	**Pipeline**	**Filter**	**Refine**
MMSU (Wang et al., 2026)	∼7 s	Tiered	Fixed	✓	✗	Wild	Manual	Manual	None
MMAU (Sakshi et al., 2025)	∼10 s	Tiered	Fixed	✓	✗	Corpus	Manual	Manual	None
MMAR (Ma et al., 2025)	∼20 s	Tiered	Fixed	✓	✗	Wild	Manual	Manual	None
MMAU-Pro (Kumar et al., 2026)	∼2 min	Tiered	Fixed	✓	✓	Wild	Manual	Manual	None
LongAudioBench (Ghosh et al., 2025b)	∼2 min	Flat	Fixed	✓	✓	Corpus	Auto	Multi	None
AudioMarathon (He et al., 2025)	∼3 min	Flat	Fixed	✓	◗	Corpus	Auto	Manual	None
ChronosAudio (Luo et al., 2026)	∼6 min	Tiered	Fixed	✗	◗	Synthetic	Auto	None	None
LongSpeech (Yang et al., 2026b)	∼10 min	Flat	Fixed	✗	◗	Corpus	Auto	Manual	None
LAT-Bench (Shao et al., 2026)	≤30 min	Flat	Fixed	✗	◗	Wild	Auto	None	None
BLAB (Ahia et al., 2025)	∼51 min	Flat	Fixed	✓	◗	Wild	Manual	Manual	None
VoiceGiraffe (Ye et al., 2026)	∼55 min	Tiered	Fixed	✓	✗	Wild	Auto	Multi	None
LongAudioSpan (Ours)	**10 min–2 h$^+$** ‖ **∼50 min**	**Tiered**	**Free**	✓	✓	**Wild**	**Auto**	**Multi**	**Adaptive**

In **Open**: ✓ open-ended question; ◗ fixed-answer question (e.g., automatic speech recognition, ASR).

genuine comprehension. **3) Data construction**: synthetic or corpus-derived audio yields reliable ground truth but lacks real-world variation; wild audio either requires costly manual annotation or depends on caption quality through caption-then-QA pipelines. **4) Quality assurance**: existing pipelines typically rely on post-hoc human review as a single filtering layer, with no mechanism to adaptively refine questions as the benchmark evolves.

To address these gaps, we introduce ***LongAudioSpan***, a benchmark spanning both duration, from minutes to hours, and depth of audio comprehension through a three-level cognitive framework of *perception*, *understanding*, and *reasoning*. To construct LongAudioSpan, we collect long-form audio in the wild and design a fully automated, scalable pipeline in two phases: Phase 1 generates a structured caption from audio; Phase 2 generates QA pairs with multi-level quality control and adaptive refinement. It follows two complementary paths: **Native QA**, which derives questions natively from the audio's inherent content, freely composing diverse question types within the cognitive framework, in both multiple-choice and open-ended formats; and **Anchor QA**, which introduces acoustic anchors into the audio and builds a question chain centered on the anchor and its surrounding context, with chain scoring to suppress short-cuts.

We evaluate LongAudioSpan on 12 recent LALMs, seven open-source and five proprietary. Even the strongest models are bottlenecked by distilling a few relevant facts from a long, redundant signal before any reasoning begins, a difficulty that grows with audio length.

In summary, we make the following contributions:

- **A comprehensive long-form audio benchmark with diverse evaluation paradigms** (§3). LongAudioSpan pairs bilingual in-the-wild audio spanning minutes to hours with a dual-path evaluation across three cognitive levels and three question formats.
- **A scalable automated construction pipeline with multi-level quality assurance** (§4). We automate the full loop from question generation to filtering, intercepting low-quality items at generation time and feeding adversarial critic feedback into later rounds.
- **A systematic empirical study of long-form audio comprehension** (§5). Across a broad set of recent LALMs, we analyze long-form comprehension along duration, cognitive levels, and scoring modes, characterizing where and why it breaks down.

2 Related Work

2.1 Large Audio-Language Models

Audio-language modeling has progressed from modality-specific systems toward unified, general-purpose models. Early systems targeted individual modalities and tasks: speech-centric language models (Zhang et al., 2023; Rubenstein et al., 2023) and general audio encoders coupled to LLMs for captioning and question answering over short clips (Deshmukh et al., 2023; Tang et al., 2024; Gong et al., 2024; Ghosh et al., 2024). The Qwen-Audio line (Chu et al., 2023; 2024) brought speech, environmental sound, and music into a single model, and open releases have since prolif-

erated along two lines. Audio specialists keep the input interface audio-only: the Audio Flamingo series (Ghosh et al., 2025a; 2026), Voxtral (Liu et al., 2025), Kimi-Audio (Kimi Team, 2025), MiMo-Audio (Xiaomi LLM-Core Team, 2025), StepAudio (Lin et al., 2026), Eureka-Audio (Zhang et al., 2026), and MOSS-Audio (Yang et al., 2026a). Omni-modal models pair audio with vision and text: the Qwen-Omni series (Xu et al., 2025a;b), Phi-4-Multimodal (Abouelenin et al., 2025), Baichuan-Omni (Li et al., 2025), and MiniCPM-o (Cui et al., 2026). Recent releases also push toward longer effective context (Xu et al., 2025b; Qwen Team, 2026), stronger multi-step reasoning (Xie et al., 2025; Tian et al., 2025), and finer temporal grounding (Sun et al., 2026; Hegde et al., 2026). Alongside open models, proprietary systems such as the Gemini series (Gemini Team, 2024; 2025; 2026) and Qwen3.5-Omni (Qwen Team, 2026) provide native audio understanding over long contexts.

2.2 Audio Comprehension Benchmarks

Benchmarks for LALMs initially centered on short clips under one minute. MMAU (Sakshi et al., 2025), AudioBench (Wang et al., 2025), and AIR-Bench (Yang et al., 2024) assess perception and understanding across speech, sound, and music, while MMAU-Pro (Kumar et al., 2026) broadens task coverage and adds open-ended items. MMAR (Ma et al., 2025), MMSU (Wang et al., 2026), and STAR-Bench (Liu et al., 2026) instead target deeper reasoning, multi-step or spatio-temporal, but still on clips measured in seconds. A more recent line pushes audio from a few minutes to about an hour (Table 1), drawing on corpora or synthetic narration (LongAudioBench (Ghosh et al., 2025b), AudioMarathon (He et al., 2025), ChronosAudio (Luo et al., 2026), LongSpeech (Yang et al., 2026b)) or in-the-wild recordings (LAT-Bench (Shao et al., 2026), BLAB (Ahia et al., 2025), VoiceGiraffe (Ye et al., 2026)). Yet evaluation still treats long audio much as it does short clips, with fixed task schemes and mostly flat depth, and construction remains largely automated over synthetic or caption-derived ground truth and checked by one-off filters.

2.3 QA Generation Pipelines

Constructing audio QA at scale requires generating questions whose answers are grounded in the audio. Early benchmarks rely on expert annotation (Sakshi et al., 2025; Ma et al., 2025; Wang et al., 2026), which yields reliable ground truth but scales poorly. To reduce annotation cost, many recent pipelines adopt a caption-then-QA design, deriving questions from a textual description of the audio rather than from the audio itself (Ye et al., 2026; Ghosh et al., 2025b; Wang et al., 2025). This design scales readily, but the resulting questions are bounded by the fidelity of the intermediate caption. Pipelines also differ in how they guard quality, from full human verification (Sakshi et al., 2025) to LLM-based filtering (Ye et al., 2026) and post-hoc text-only audits that flag items answerable without the audio (Kumar et al., 2026). These checks are one-off filters applied after generation.

2.4 Evaluation Methods

Audio benchmarks predominantly score with multiple-choice questions, which afford deterministic grading (a 25% random baseline for four options) but can be solved by option elimination rather than genuine comprehension. Open-ended items fall into two kinds: fixed-answer questions keep grading automatic but constrain what can be asked, while free-form questions better reflect real usage and rely on LLM-as-judge scoring (Kumar et al., 2026; Ghosh et al., 2025b; Luo et al., 2026). Such judging is typically holistic, a single overall rating that does not show which aspects of a response are correct; finer criteria have only graded reasoning traces behind multiple-choice answers (Ma et al., 2025). Chain-structured evaluation, explored in video understanding, links successive questions into a dependent sequence (Fu et al., 2026).

3 Benchmark Design

3.1 Overview

LongAudioSpan's design centers on two complementary question paths that differ in how question content is sourced and how ground truth is obtained. The base path, **Native QA**, draws questions from the audio's original content via a structured caption of the audio (§4). Its ground truth, however, is bounded by caption completeness: content omitted from the caption cannot be turned into

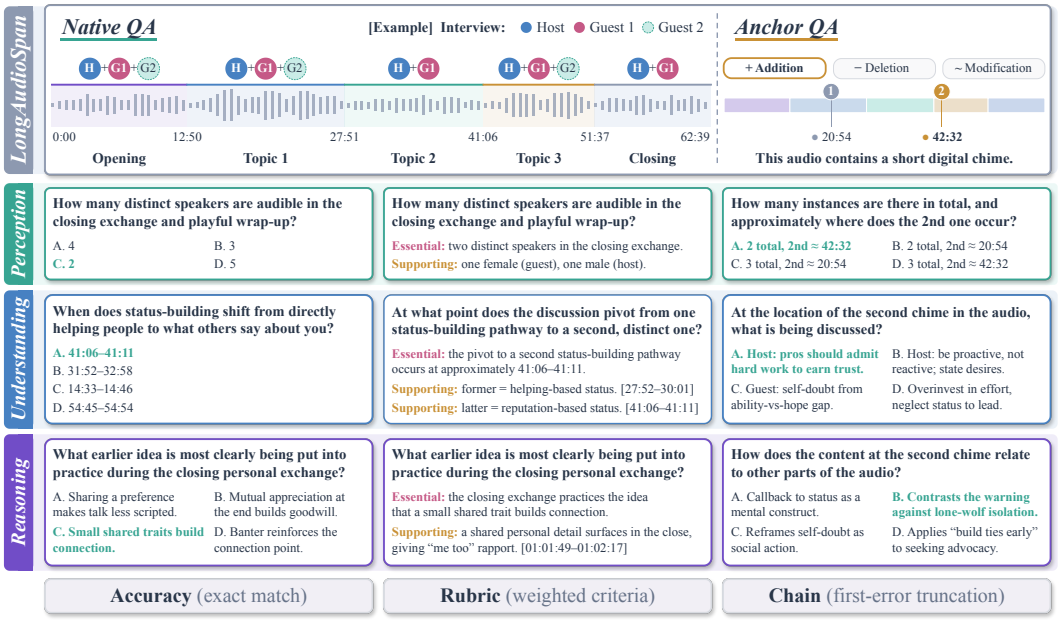


Figure 1: Overview of LongAudioSpan. Top: the two question paths over one audio sample, Native QA (left) and Anchor QA (right). Bottom: abridged example question for each taxonomy level (rows) and scoring scheme (columns).

questions, and on hour-long audio such omissions are hard to avoid. To sidestep this dependency, the second path, **Anchor QA**, injects ground truth rather than extracting it: it plants acoustic anchors into the audio and builds questions around their surrounding context, so the operation itself defines the correct answer. Together, the two paths combine broad coverage of natural content with verifiable ground truth at targeted positions.

The two paths share a unified cognitive taxonomy but differ in question format and scoring (§3.2). Instantiated over 360 in-the-wild audio samples, they yield 3,240 QA items balanced across languages (English and Chinese) and duration tiers (10 min to 2 h^{+}) (§3.3).

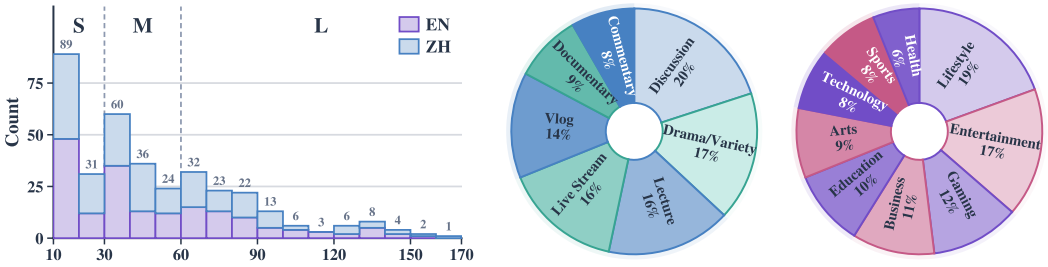
3.2 Taxonomy and Evaluation

Figure 1 presents each path's questions and scoring built on a shared taxonomy. The taxonomy follows a cognitive progression from information acquisition to inferential judgment. At the base, *Perception* (P) involves identifying and acquiring explicit information present in the audio. *Understanding* (U) constructs meaning from that explicit information, building coherent interpretations of the content. *Reasoning* (R) draws inferences and judgments that go beyond the given information.

Native QA. Native QA realizes the taxonomy as a free scheme. Each level is organized into three general dimensions, each opening a question space (Appendix B.2). For each audio, the generator draws on the question space for inspiration and grounds each question in the audio's content, yielding questions specific to that audio. Each question is posed in both formats, as an MCQ and an OEQ. Multiple-choice questions (MCQs) with four options are scored by exact match, the *Accuracy* over an MCQ set Q:

$$\text{Accuracy} = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[\hat{a}_q = a_q]. \tag{1}$$

Open-ended questions (OEQs) are graded by an ensemble of rubric-based LLM judges against item-specific criteria. Each criterion c takes a *semantic* score $\sigma_c \in \{0, 0.5, 1\}$ for content correctness and, when the criterion is time-bound, a *temporal* score $\tau_c \in \{0, 1\}$ that grounds the content at the right moment, treating a timestamp as a citation. The *Rubric* score is the importance-weighted mean of



(a) Duration (min) (b) Genre (c) Topic

Figure 2: Distribution of source audio across (a) duration tiers, (b) genres, and (c) topics.

the per-criterion scores s_c:

$$\text{Rubric} = \frac{\sum_{c \in C} w_c \, s_c}{\sum_{c \in C} w_c}, \qquad s_c = \begin{cases} \sigma_c \text{ or } \tau_c, & \text{essential,} \\ \mathbf{1}[\sigma_c > 0]\,(\sigma_c + \tau_c)/2, & \text{supporting.} \end{cases} \tag{2}$$

An *essential* criterion checks the direct answer to the question and is scored on a single dimension, σ_c for semantic content or τ_c for temporal localization, at weight $w_c = 1$. A *supporting* criterion rewards a detail that substantiates the answer at weight $w_c = 0.5$, scored on semantic content with temporal grounding as an additional trait. Grounding is conditional on content: a timestamp acts as a citation, so it earns no credit for a false statement, and the indicator $\mathbf{1}[\sigma_c > 0]$ zeros a supporting criterion whose content is wrong (Appendix C.2). The OEQ withholds the options the MCQ provides, removing the elimination shortcut and measuring a model's ability more realistically.

Anchor QA. Anchor QA uses a fixed chain that maps directly onto P→U→R. Each audio sample undergoes one of three anchor operations: *Addition* inserts a sound event into the audio, *Deletion* applies a silence or noise mask to remove information at chosen positions, and *Modification* alters acoustic attributes such as volume or playback speed (Appendix B.3). Around each anchor we construct three MCQs, one per level: P grounds the anchor by counting its occurrences and locating the Nth instance; U then infers the content being discussed there; and R in turn relates that content to other parts of the audio. The three questions form an implicit cognitive dependency, so an answer that follows an earlier error rests on a wrong premise. The *Chain* score therefore credits answers only up to the first error,

$$\text{Chain} = \frac{1}{3} \sum_{k=1}^{3} \prod_{j=1}^{k} \mathbf{1}[\hat{a}_j = a_j], \tag{3}$$

where $\hat{a}_j$ is the answer to the j-th question, so one early mistake zeros every later term. Under random guessing the expected score is 10.9%, well below the 25% baseline of a single MCQ.

3.3 Statistics

LongAudioSpan draws on 360 in-the-wild audio samples, split evenly across three duration tiers, S (10–30 min), M (30–60 min), and L (60+ min), and between English and Chinese; the samples further span 7 genres and 9 topics (Figure 2). Every sample also has an anchor-modified version under one of three operations, for 720 audio files totaling roughly 600 hours (avg. ∼50 min). The tiers form a deliberate duration gradient: S, M, and L average roughly 16, 42, and 90 minutes, with the longest sample reaching 2.7 hours.

Each sample is annotated with one question per cognitive level under each evaluation format, yielding 3,240 items: 1,080 Native MCQs, 1,080 Native OEQs, and 1,080 Anchor MCQs organized as 360 three-question chains (Table 2).

Table 2: Benchmark statistics.

Statistic	Value
Audio Sources	720
Original	360
Per tier (S/M/L)	120
Per lang. (EN/ZH)	180
With anchors	360
Per mode (A/D/M)	120
QA Items	3,240
Per audio (1P/1U/1R)	3
Native (MCQ)	1,080
Native (OEQ)	1,080
Anchor (MCQ)	1,080

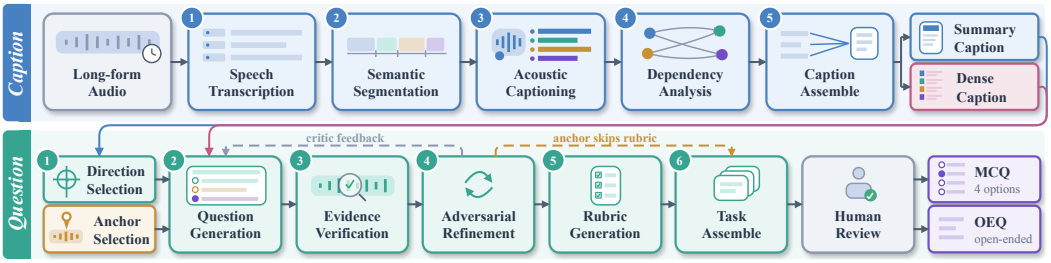


Figure 3: LongAudioSpan construction pipeline. Phase 1 (top) converts each long-form audio into a structured caption via a five-step pipeline. Phase 2 (bottom) generates questions under a shared six-step framework; Native QA runs all steps, while Anchor QA skips rubric generation.

4 Data Construction

4.1 Overview

Figure 3 shows the automated two-phase pipeline that constructs LongAudioSpan. **Phase 1** converts each audio into a structured caption (§4.2). **Phase 2** generates questions from that caption through two pipelines that share a common framework: the Native QA pipeline (§4.3) derives questions from the caption and verifies their ground truth against the audio, while the Anchor QA pipeline (§4.4) plants anchors into the audio and reads ground truth from the anchor settings. Quality assurance (§4.5) runs during and after generation. Models used at each stage are listed in Appendix C.4.

4.2 Caption Pipeline

The caption pipeline converts each audio into a structured caption in five steps.

1) Speech Transcription. An ASR model produces sentence-level transcripts with timestamps.

2) Semantic Segmentation. A language model splits the transcript into contiguous, semantically coherent segments using cues such as topic shifts, speaker changes, and narrative transitions, each capped at about 5 minutes and annotated with a title, summary, and key topics.

3) Acoustic Captioning. Each segment's audio is captioned by a multi-modal model, adding time-stamped acoustic details beyond the transcript: speaker identities and voice characteristics, paralinguistic cues, environment descriptions, and non-speech audio events.

4) Dependency Analysis. A language model extracts a speaker registry and cross-segment dependencies (e.g., causal links, contrasts, references), and produces global context (overall topic, genre, speaker roles).

5) Caption Assemble. Outputs are merged into the final caption and indexed as *evidence records*, each a verifiable fact with a segment, time range, and modality tag, serving as the grounding units for question generation. Appendix B.1 details the assembled caption's two views.

4.3 Native QA Pipeline

The native QA pipeline derives questions from the audio's natural content, covering all three cognitive levels. A language model creates the questions across the first two steps.

1) Direction Selection. For each cognitive level, the language model chooses a free question direction and the segments relevant to it, drawing on the summary caption and the question space (Appendix B.2). The dense captions of those segments supply detail for question generation.

2) Question Generation. From the dense captions of the relevant context, the language model concretizes the direction into a multiple-choice candidate: a question stem, four options, the correct option, and supporting evidence records. Lessons from the critic feedback loop (§4.5) are fed back into the prompt, so the model progressively avoids previously observed failure modes.

3) Evidence Verification. Cited evidence records are verified against the audio (§4.5).

4) Adversarial Refinement. Surviving candidates pass through a three-level solver and a critic feedback loop that retain only audio-dependent items (§4.5).

5) Rubric Generation. Each retained MCQ is first recast as an OEQ. From the solver's multiple reasoning paths, common and necessary points are distilled into essential/supporting rubric criteria, and a reference answer is synthesized.

6) Task Assemble. Items are unified to a common format, and options are shuffled to keep the correct answer balanced across positions, preventing position bias.

4.4 Anchor QA Pipeline

The anchor pipeline shares the native framework but differs in its first two steps. The other four are reused, with verification and refinement on U/R only and no rubric generation.

1) Anchor Selection. Guided by the summary caption, the pipeline selects 1–3 anchor positions that carry rich semantic context and relate to other segments. The anchor operation (Appendix B.3) is then applied, producing the modified audio and ground truth by construction.

2) Question Generation. From the anchor operations and their ground truth, together with the dense captions of the anchor context and its related segments and their known relations, the three-level question chain is generated.

4.5 Quality Assurance

Quality assurance combines automatic gates and a critic feedback loop during generation with a human review afterward. Appendix B.4 details the gate settings, reports each gate's pass and rejection rates on a construction run, and describes the review procedure.

Evidence verification. Every question must cite evidence records from the caption. Each cited fragment is validated against the audio by a multi-modal model, and a candidate is kept only if all its records are confirmed, guarding the benchmark against caption errors.

Three-level solver. The solver is a language model that answers each candidate at three levels of access. The *blind* level sees only the stem and options and rejects guessable items. The *text-only* level adds the transcript and rejects items answerable from the spoken words alone. The *full* level adds the dense caption, from which the model draws multiple reasoning paths and verifies solvability. A candidate is kept only if it fails the first two levels and the full level solves it.

Critic feedback loop. A critic agent analyzes how the blind and text-only solvers reached their answers, summarizing a root cause for each failure mode and the lessons shared across them. These lessons are kept in a memory bank and fed into later generation, so subsequent questions avoid the same shortcuts. The Native and Anchor pipelines keep separate memory banks.

Human review. Every item that clears the automatic gates receives a final human review against the audio, where annotators verify the question and its answer and, for anchor items, that the edit is audible and distinct from existing events, then accept it, apply a minor fix, or reject it.

5 Experiments

5.1 Experimental Setup

Audio models. We evaluate 12 models under a single native-audio protocol: each receives the audio itself, truncated to its input limit when it cannot ingest the whole clip, never a transcript in place of audio (App. C.1). Seven are open-source, spanning audio specialists (Audio Flamingo Next, Voxtral-Mini and Voxtral-Small, MOSS-Audio) and omni-modal models (Phi-4-Multimodal, Baichuan-Omni-1.5, Qwen3-Omni); five are closed and all omni-modal, three Gemini versions (2.5-Pro, 3-Flash, 3.1-Pro) and two Qwen3.5-Omni tiers (Flash and Plus). Models that expose a thinking mode are run in both standard and thinking configurations. Per-model input limits and the specs behind them are given in App. C.1 (Table 6).

Text-only baselines. Three text-only baselines, all GPT-5.4 with no audio, bound the task at rising levels of textual access. *Question-only* sees the question and its options alone, a floor set by answer-choice priors. *Transcript* adds the ASR transcript, a rung reading only the spoken words. *Caption* adds the dense authoring caption the questions were written from (for Chain, its anchor-augmented form), a text reference over the material the benchmark was built on.

Table 3: Main results across three evaluation modes, *Accuracy* (Native-MCQ), *Rubric* (Native-OEQ), and *Chain* (Anchor-MCQ), and three duration tiers (S/M/L); *Avg* is the unweighted mean over the tiers. **Limit** is each model's audio-input cutoff in minutes (–: untruncated or not applicable) and ✓ marks a thinking variant; gray cells are scored on audio truncated to that limit (protocol and per-model limits in §5.1). Best per column in **bold**, second-best underlined (audio models only).

				Accuracy (%)				Rubric (%)				Chain (%)			
Model	**Size**	**Think**	**Limit**	**S**	**M**	**L**	**Avg**	**S**	**M**	**L**	**Avg**	**S**	**M**	**L**	**Avg**
Text-only Baselines															
GPT-5.4 (question-only)	–	✓	–	41.1	41.1	40.6	40.9	–	–	–	–	8.9	11.9	10.3	10.4
GPT-5.4 (transcript)	–	✓	–	64.8	60.7	53.4	59.7	41.0	41.5	37.2	39.9	9.4	14.2	14.2	12.6
GPT-5.4 (caption)	–	✓	–	83.2	80.0	77.3	80.2	59.1	57.4	53.0	56.5	66.7	64.4	53.1	61.4
Open-source Models															
Phi-4-Multimodal	5.6B	✗	30m	33.7	34.6	34.2	34.2	18.4	13.6	10.1	14.0	12.9	7.2	8.8	9.6
AudioFlamingoNext	8B	✗	30m	46.1	43.6	37.2	42.3	16.1	12.2	13.7	14.0	5.3	5.6	6.1	5.7
		✓	30m	43.3	38.9	35.0	39.1	26.9	20.3	16.3	21.2	7.5	11.1	13.1	10.6
Voxtral-Mini	3B	✗	40m	45.3	35.8	37.2	39.4	25.6	21.8	13.6	20.3	5.6	9.4	18.3	11.1
Voxtral-Small	24B	✗	40m	42.5	40.0	31.9	38.1	28.9	24.2	20.7	24.6	12.5	4.7	4.7	7.3
MOSS-Audio	8B	✗	50m	45.0	43.1	35.6	41.2	23.7	17.7	10.7	17.4	8.3	11.4	9.2	9.6
		✓	50m	52.2	52.8	38.6	47.9	30.1	15.4	11.6	19.0	13.9	15.3	11.4	13.5
Baichuan-Omni-1.5	11B	✗	80m	46.9	46.9	36.4	43.4	19.0	12.5	9.7	13.7	11.9	14.4	14.7	13.7
Qwen3-Omni	30B (A3B)	✗	80m	53.9	54.4	46.2	51.5	31.3	24.3	20.6	25.4	20.5	19.2	13.2	17.6
		✓	80m	56.9	57.5	49.2	54.5	38.4	29.2	22.8	30.1	7.8	10.6	15.3	11.2
Closed-source Models															
Gemini-2.5-Pro	–	✓	–	67.5	64.2	54.7	62.1	46.9	40.7	36.4	41.3	24.2	25.3	21.9	23.8
Gemini-3-Flash	–	✗	–	65.6	69.2	64.2	66.3	54.9	**55.0**	**53.8**	**54.6**	16.7	20.6	11.9	16.4
Gemini-3.1-Pro	–	✓	–	76.7	**74.4**	65.8	72.3	56.2	48.8	43.3	49.4	17.5	25.3	20.0	20.9
Qwen3.5-Omni-Flash	–	✗	–	73.9	71.4	58.9	68.1	50.3	40.9	35.9	42.4	23.1	19.2	18.3	20.2
Qwen3.5-Omni-Plus	–	✗	–	**79.2**	**74.4**	**70.3**	**74.6**	**56.9**	48.6	47.0	50.8	**30.8**	**33.1**	**23.1**	**29.0**

5.2 Overall Performance

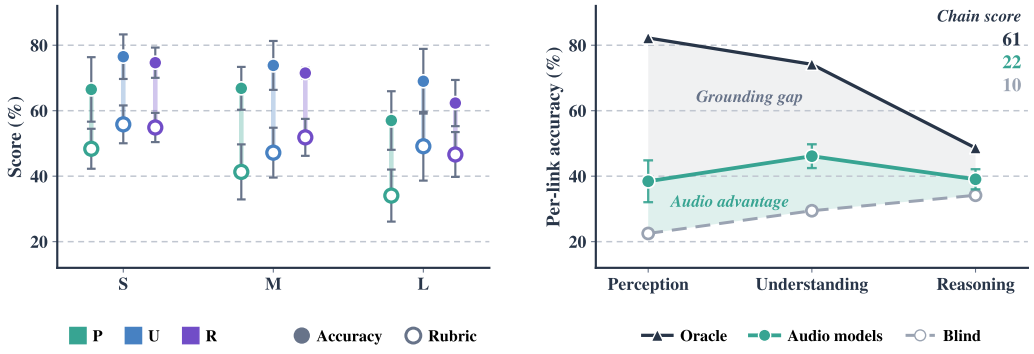
We read Table 3 along three axes: the overall level of performance, the gap between models, and the variation within each model.

Current models are burdened by the audio itself, a long and redundant signal they must distill. Handed the same content as text, the dense authoring caption the questions were written from rather than the audio, a text-only solver (GPT-5.4, caption baseline) leads every audio model on Accuracy and Rubric. What separates the two is modality: one side reads the facts off text, the other must recover them from the waveform. The text side winning places the difficulty before reasoning, in gathering usable information from the audio: an audio encoder emits tens of frames per second, so the model must sift a long and largely redundant stream for the few facts a question turns on.

Long-form audio exposes a clear gap between closed- and open-source models. The open models are also limited in reach: each is capped at a 30–80 min input (grey cells) and sees only part of the Medium and Long audio, while the closed models run untruncated. The gap holds even where reach is equal: on the Short tier, which every model ingests in full, every closed model outscores every open model on Accuracy, and the same separation holds on Rubric. The exception is Chain, whose scoring is bottom-loaded (Figure 7): with every model compressed toward the floor, the closed and open ranges overlap, though the highest scores stay closed.

Performance declines steadily with duration, worst on hour-long inputs. The fall spans all three modes: Accuracy, Rubric, and Chain each drop from S to L, and the L tier (avg. $\sim$90 min) is the lowest for nearly every configuration. Part of this is length alone: the text-only transcript and dense-caption baselines lose 6–11 accuracy points from S to L, while the content-free question-only baseline holds flat. Longer inputs thus tax comprehension even as clean text, a long-context penalty the audio models carry on top of distilling the signal.

Two cross-cutting factors round out the picture. Enabling *thinking* generally helps, most clearly on open-ended Rubric. *Language* does not drive the ordering: the strongest models are language-balanced, and a sizable gap appears only on the weaker open models (App. D.1).



(a) Native QA (b) Anchor QA

Figure 4: Scores by cognitive level (Perception, Understanding, Reasoning). (a) Native QA per duration tier: filled dots Accuracy, open Rubric, ticks ± 1 SD. (b) Anchor QA per-link accuracy for a blind solver, the audio models (± 1 SD), and a text oracle given the anchors; corner ladder is the truncated chain score (Eq. 3).

5.3 Cognitive levels

Figure 4 splits both question paths by cognitive level.

Perception is the weakest cognitive level. Across the closed-source models it is the lowest level on both paths, and its weakness shows most once scoring moves past a multiple-choice pick. Under both metrics perception ranks lowest, but the drop is far steeper under Rubric and widens with duration: on the M and L tiers, perception sits well below understanding and reasoning (Figure 4a). The deficit is specific: recognition of content keeps pace with the higher levels, while time-sensitive perception lags behind. Within perception, localization and counting are the weak dimensions, and they fail differently (App. D.2): counting is low under both formats alike, a capacity limit no scoring change lifts, while localization collapses only once the options are withheld. Anchor QA isolates the same cost: hand a text oracle the anchor positions and its perceptual link clears the audio models by the widest margin anywhere on the chain, a gap that closes toward reasoning only because the oracle itself falls there (Figure 4b).

Grounding an answer in the audio is the core failure. It surfaces once scoring makes a model produce a grounded answer, on both question paths. On the native path, a correct multiple-choice pick only weakly predicts a good open-ended answer: matched item by item on the closed-source models, localization keeps only about half its accuracy once the options are withheld (App. D.3), because four options let a model confirm a timestamp it cannot itself produce. In the open-ended answers, models describe what happens far more reliably than when it happens, even on reasoning questions (App. D.4). On the anchor path, grounding becomes a requirement: a link earns credit only when the model has named the anchor, so the truncated chain score (Eq. 3) falls far below per-link accuracy, for audio models and a blind, no-audio solver alike. Even on anchors it never located, a model answers understanding nearly as often as when it did (App. D.5): the higher links are answered about as much by guessing as by grounding. The effect is sharpest where the anchor is hardest to find: reasoning peaks on silenced spans, the operation whose perception link is weakest, while their chain score stays the lowest of the three.

6 Conclusion

LongAudioSpan is a benchmark for grounded, long-form audio comprehension, pairing durations from minutes to hours with three levels of cognitive depth, each scored through matched multiple-choice and open-ended formats. Across 12 LALMs, the failure is consistent: before reasoning is even tested, models cannot reliably ground what they describe in the audio itself, and the gap widens as audio lengthens. LongAudioSpan offers a baseline for tracking progress as models learn to sustain comprehension across longer spans and close this gap. The relevant measure is less the amount of audio a model can accept than the amount it can use.

AI USAGE STATEMENT

In this work, we used generative AI tools as components of our method and for manuscript preparation. Large language model APIs are part of LongAudioSpan's automated construction pipeline, generating structured captions, questions, and adversarial critic feedback (§4, Appendix B), and an LLM judge scores open-ended responses (§3.2, Appendix C). All pipeline output passes the verification and human review steps described there before entering the benchmark. We additionally used AI writing assistants to polish the manuscript's prose. We did not use generative AI to develop theoretical models, formulate or prove mathematical claims, propose or refine the research hypotheses, or design the research methodology. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The construction pipeline (captioning, QA generation, adversarial critic feedback) is described in Appendix B. The evaluation protocol, scoring rubrics, and judge prompts are in Appendix C, and the models and API versions used are listed in Appendix C.4. LongAudioSpan is available at the anonymized link in the abstract during review; the real repository will be made public on acceptance.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Orevaoghene Ahia, Martijn Bartelds, Kabir Ahuja, Hila Gonen, Valentin Hofmann, et al. BLAB: Brutally long audio bench. *arXiv preprint arXiv:2505.03054*, 2025.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning*, 2023.

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

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, et al. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. In *Advances in Neural Information Processing Systems*, 2023.

Zach Evans, Julian D. Parker, Matthew Rice, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable Audio 3, 2026. URL https://arxiv.org/abs/2605.17991.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, et al. Video-MME-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Gemini Team. Gemini 3.1 pro model card. Google DeepMind, 2026. https://deepmind.google/models/model-cards/gemini-3-1-pro/.

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 *Annual Meeting of the Association for Computational Linguistics*, 2024.

Sreyan Ghosh, Arushi Goel, 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. In *Advances in Neural Information Processing Systems*, 2025a.

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 *International Conference on Machine Learning*, 2025b.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, et al. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. In *International Conference on Learning Representations*, 2024.

Peize He, Zichen Wen, Yubo Wang, Yuxuan Wang, Xiaoqian Liu, et al. AudioMarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio LLMs. *arXiv preprint arXiv:2510.07293*, 2025.

K. Hegde, A. Sridhar, Naveen Vakada, Yinyi Guo, and Erik Visser. Event-grounded question answering over long audio via structured retrieval. *arXiv preprint arXiv:2602.14612*, 2026.

Kimi Team. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, et al. MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. In *AAAI Conference on Artificial Intelligence*, 2026.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, et al. Baichuan-Omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

Bin Lin, Bo Zhao, Boyong Wu, Chao Yan, Chen Wu, et al. StepAudio 2.5 technical report. *arXiv preprint arXiv:2605.23463*, 2026.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, et al. Voxtral. *arXiv preprint arXiv:2507.13264*, 2025.

Zihan Liu, Zhikang Niu, Qiuyang Xiao, Zhisheng Zheng, Ruoqi Yuan, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Jianze Liang, Xie Chen, Leilei Sun, Dahua Lin, and Jiaqi Wang. STAR-Bench: Probing deep spatio-temporal reasoning as audio 4D intelligence. In *International Conference on Learning Representations*, 2026.

Kaiwen Luo, Liang Lin, Yibo Zhang, Moayad Aloqaily, Jialiang Tao, 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. *arXiv preprint arXiv:2601.04876*, 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. In *Advances in Neural Information Processing Systems*, 2025.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 2023.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, et al. AudioPaLM: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

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 *International Conference on Learning Representations*, 2025.

Mingchen Shao, Hang Su, Wenjie Tian, Bingshen Mu, Zhennan Lin, Lichun Fan, Zhenbo Luo, Jian Luan, and Lei Xie. Listening with time: Precise temporal awareness for long-form audio understanding. *arXiv preprint arXiv:2604.22245*, 2026.

Luoyi Sun, Xiao Zhou, Zeqian Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. SpotSound: Enhancing large audio-language models with fine-grained temporal grounding. *arXiv preprint arXiv:2604.13023*, 2026.

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 *International Conference on Learning Representations*, 2024.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, et al. Step-Audio-R1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

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 North American Chapter of the Association for Computational Linguistics*, 2025.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, 2026.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2023.

Xiaomi LLM-Core Team. MiMo-Audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-Reasoner: Improving reasoning capability in large audio language models. In *Annual Meeting of the Association for Computational Linguistics*, 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. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, et al. MOSS-Audio technical report. *arXiv preprint arXiv:2606.01802*, 2026a.

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. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2026b.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Jashin Ye, Dongxiao Wang, Yixuan Ye, Sashuai Zhou, Weihuang Lin, Mingyang Han, Kunpeng Wang, Zeyu Yuan, Boyu Li, Haoxiang Shi, Jingchen Shu, Jun Song, and Bo Zheng. VoiceGiraffe: A benchmark for extreme long-context audio-language understanding. *arXiv preprint arXiv:2605.27976*, 2026.

Dan Zhang, Yishu Lei, Jing Hu, Shuwei He, Songhe Deng, Xianlong Luo, Danxiang Zhu, Shikun Feng, Rui Liu, Jingzhou He, Yu Sun, Hua Wu, and Haifeng Wang. Eureka-Audio: Triggering audio intelligence in compact language models. *arXiv preprint arXiv:2602.13954*, 2026.

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

APPENDIX CONTENTS

A Data Ethics **14**

A.1 Licensing and Release 14

A.2 Ethics and Responsible Use 14

B Construction Process **15**

B.1 Caption Pipeline 15

B.2 Native QA Pipeline 15

B.3 Anchor QA Pipeline 15

B.4 Quality Assurance 17

C Evaluation Protocol **18**

C.1 Evaluation Setup 18

C.2 Scoring Details 19

C.3 Judge Prompt 20

C.4 Models and APIs 22

D Additional Results **22**

D.1 Performance Across Languages 22

D.2 Performance Across Dimensions 22

D.3 Question Formats: Multiple-Choice vs. Open-Ended 23

D.4 Rubric Axes: Semantic vs. Temporal 24

D.5 Chain Scoring: Grounding vs. Guessing 26

E Limitations **27**

A DATA ETHICS

A.1 LICENSING AND RELEASE

LongAudioSpan is released for non-commercial research and evaluation. The artifacts we produce, QA items, rubric criteria, anchor edit manifests, sound events, and evaluation scripts, are distributed under a CC BY-NC-SA 4.0 license. The source recordings are publicly available media collected from the Internet, and copyright of each recording remains with its original creator. The recordings are included only for evaluation, and downloading the dataset constitutes agreement not to redistribute the audio or use it in commercial products.

A.2 ETHICS AND RESPONSIBLE USE

All audio comes from publicly available media; speakers are referenced by role rather than identity, and no personal information in the audio is used as an answer key. LongAudioSpan is an evaluation resource and is not intended for training speaker-identification or surveillance systems; the anchor perturbations exist only to probe grounding. Human reviewers were compensated at or above local market rates and worked only with publicly available media.

Summary Caption

```
GLOBAL
  topic:  how status and power operate in professional life, and how respect and
          influence are built.
  arc:    defines status as perception-based, distinct from power as resource control;
          reframes both as legitimate, then moves into practical frameworks. ...
  tone:   insightful, pragmatic, encouraging; warm.

SPEAKERS
  spk_1  male host     guides conversation, asks questions
  spk_2  female guest  primary expert on status and power
  spk_3  male guest    secondary participant

SEGMENT  (id . time . event_type . title)
  seg_000  0-26s      monologue   Status as a construct shaped by others' judgments
  seg_001  26-157s    interview   The guest's path into power and status research
  seg_002  157-310s   discussion  Defining power and status and their difference
  ...
  seg_031  3731-3759s transition  Final acknowledgments and goodbye

DEPENDENCIES  (source -> target . type . relation)
  seg_000 -> seg_002   definition refinement
     the opening claim that status lives in others' minds is formalized in seg_002 via
     the contrast with power as control over resources.
  seg_002 -> seg_004   causal
     after both are defined, seg_004 explains why they matter. ...
```

Figure 5: Summary caption: an audio-level overview (global topic, narrative arc, and tone; a speaker registry; a segment index; and cross-segment dependencies). It drives direction selection for Native QA and anchor selection for Anchor QA.

B Construction Process

B.1 Caption Pipeline

Every question is written from a structured caption, not from the raw audio (§4). The caption is read in two views. The *summary* view is an audio-level overview: global topic, narrative arc, and tone; a speaker registry; a segment index; and the cross-segment dependencies between them. It drives direction selection for Native QA and anchor selection for Anchor QA, the stages that reason over the whole audio at once. The *dense* view holds per-segment detail: the acoustic environment, a timestamped transcript with paralinguistic annotation, and discrete audio events. Generation reads the global context together with the dense captions of only the segments a question targets.

Both views are grounded in *evidence records*. Assembling the caption flattens it into one record per groundable unit, each carrying a stable ID, its segment, a time range, a modality (speech, sound, environment, cross-segment, or global), and the grounded content. The generator cites the IDs it relied on, and verification re-checks exactly those IDs against the audio before a candidate is kept (§4.5). These records are the grounding unit referenced throughout §4. Figures 5 and 6 render one audio sample through both views.

B.2 Native QA Pipeline

Table 4 lists question directions for each (cognitive level, dimension) pair introduced in §3.2. Following the free scheme of §3.2, the generator treats these directions as inspiration, deciding the actual direction of each question for the audio at hand and combining several where useful (§4.3).

B.3 Anchor QA Pipeline

Anchor QA uses one operation of each type, applied to the audio before any question is written (§4.4): addition, deletion, and modification. Each leaves a trace that is localizable to a point in time and layered across the perception, understanding, and reasoning levels, and each is applied at one to three positions so the perception question can ask how many times the trace occurs and where a given one falls.

Dense Caption

```
SEGMENT  seg_001 . 25.7-156.7s . interview

ENVIRONMENT
  Quiet professional studio; no reverb or background noise.

SPEECH
  30.0-31.2s    spk_1: "All right welcome to the show"
                (warm, welcoming, slight rising intonation)
  32.3-33.0s    spk_2: "Happy to be here"  (brief, cheerful, upbeat)
  34.0-41.4s    spk_1: "I'm really curious what drew you to organizational psychology,
                and to power and status in the workplace"
                (measured, thoughtful, curious)
  ...           (continues to 156.6s)

AUDIO EVENTS
  93.6-93.8s    quiet mouth click / lip smack
```

Figure 6: Dense caption: per-segment detail (acoustic environment, a timestamped transcript with paralinguistic annotation, and discrete audio events), shown for one segment. Every segment carries the same fields, and generation reads the global context together with the dense captions of the segments a question targets.

Table 4: Native QA question directions per cognitive level and dimension.

Level	Dimension	Question Space
Perception	Recognition	Linguistic recognition; Speaker identification; Sound scene classification; Subtle audio discrimination; Prosodic and paralinguistic recognition; Audio source discrimination; Overlapping speech resolution; Sound event recognition; Music recognition
	Localization	Keyword/phrase localization; Speaker turn localization; Sound event localization; Temporal ordering; Relative temporal localization; Duration estimation
	Counting	Linguistic counting; Speaker counting; Sound event counting; Conditional counting; Comparative counting; Frequency/density judgment
Understanding	Interpretation	Role/relationship recognition; Emotional atmosphere; Epistemic stance recognition; Subtext and implicit intent; Rhetorical device recognition; Social function recognition; Silence and non-response semantics; Content structure recognition; Quotation and paraphrase recognition; Humor and comedic intent recognition
	Tracking	Topic transition detection; Emotional/stance evolution; Narrative progression; Engagement/dominance shift; Discourse function shift; Information density change; Structural segment transition; Prosodic consistency tracking
	Summarization	Holistic summarization; Key point extraction; Speaker contribution analysis; Surface vs. actual consensus; Selective summarization; Omission detection; Bias detection; Segment importance judgment
Reasoning	Causal	Attribution; Foreshadowing; Causal chain reconstruction; Competing explanations; Correlation vs. causation; Prosodic causation
	Relational	Long-range reference resolution; Cross-segment information linking; Retrospective reinterpretation; Structural segment contextual relevance
	Logical	Contradiction detection; Stance reversal detection; Argument validity assessment; Information completeness check; Audio atmosphere vs. stated content divergence

Addition mixes an external sound event into the audio at the chosen positions. Across the anchor set these events span 62 distinct categories, from animal calls and alarms to mechanical impacts and ambient nature. The event clips are synthesized with Stable Audio 3 (Evans et al., 2026): each category is rendered from a short text prompt, so the clips stay consistent within a category and carry no stock-audio fingerprint a model could have memorized. The event is laid over the existing audio at a controlled signal-to-noise ratio, roughly 6–10 dB, and for each sample any category that could already occur in it is excluded, so the added event cannot be mistaken for one that was there to begin with. A single event is reused across positions, so the chain can ask how often it occurs and where.

Deletion obscures a chosen span so its content can no longer be made out, while the span keeps its place and length on the timeline. The span is either silenced outright or smeared into unintelligibility by a low-pass filter near 300 Hz mixed with a modulated noise bed. What the listener notices is that something at that point has dropped out, which is the cue the perception question tests.

Modification alters an acoustic attribute over a span while leaving the words spoken there unchanged: either loudness, raised or lowered by 8–12 dB, or playback speed, sped up or slowed by a factor of 1.5–3 through time-stretching. The delivery changes, so the trace is a shift in how a passage sounds.

Anchor selection reads the summary view in two passes. The first shortlists a few candidate segments, each picked for its semantic weight and for the relations it holds to other segments, which later give the reasoning question somewhere to point. The second fixes a precise window inside a chosen segment, held within that segment, kept clear of the audio's opening and close, floored at a few seconds and capped to a small fraction of the total so the change stays local. The window is placed over active speech, so the operation lands on audible content instead of being lost in a pause. When an operation is placed at several positions, they are spread across distinct segments, keeping each occurrence separately locatable. These placement constraints, an added event that is not already present and a position clear of silence, are verified once more during human review (§B.4).

Chain construction builds the three questions so that their dependency lives in the options: each reads on its own, yet the wrong options are laid out so an error at one level leads to a consistent one at the next (§3.2). The perception question crosses two axes in a single item, pairing a correct or off-by-one count of the trace with either its true position or a far-off point; asked together, count and position cannot be settled one without the other. The understanding question offers the real content found at each position, so a miscount that reindexes "the Nth" occurrence lands on another position's genuine material. The reasoning question offers cross-segment relations drawn from those same positions, plus a "no clear relation" foil, so a position carried over wrong selects a true but misattributed relation. The perception options are assembled directly from the recorded ground truth; the understanding and reasoning options are written by the generator from the captioned anchor context (§4.4).

Random baseline. The three chained questions are each four-option multiple-choice items, so an uninformed guess is correct with probability $1/4$ at every level, independently. Since the chain credits only the longest correct prefix (Eq. 3), a random responder scores 0 unless it first clears the perception question, and earns $k/3$ for the first k levels it happens to clear. The induced distribution is bottom-loaded: it lands on 0 three quarters of the time and reaches a full 1 only once in sixty-four attempts (Figure 7), for an expected score of 10.9%. That floor is below the 25% of a single four-option MCQ precisely because a late correct answer earns nothing once an earlier one is wrong, which is the property that makes the chain resistant to guessing.

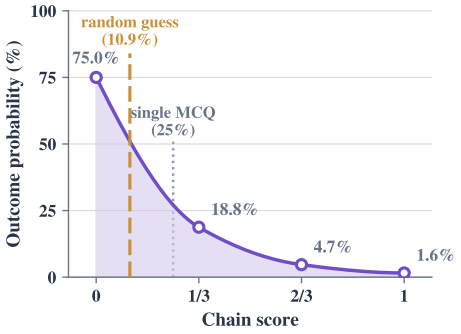


Figure 7: Anchor *Chain* score under random guessing. The four possible outcomes $\{0, 1/3, 2/3, 1\}$ are marked; the curve interpolates their probabilities.

B.4 Quality Assurance

A few settings behind the automatic gates of §4.5 are worth stating. Each solver level answers the candidate over three independent rollouts, and the gate compares how often it succeeds against a threshold, so a pass reflects a real shortcut rather than a single lucky draw: an item is discarded if the blind solver is correct in more than $1/3$ of its rollouts, or the text-only solver in more than $2/3$. Blind is the stricter bar because the stem and options alone should almost never suffice, whereas the transcript legitimately carries content and is disqualifying only when it settles the answer on nearly every rollout. The full solver must itself solve the item, confirming that it is answerable once the audio is in hand. When a gate flags an item, the critic names the shortcut it exploited from a fixed catalog, among them specificity bias, semantic leakage, elimination by absurdity, common-sense giveaway, and transcript sufficiency, and stores the lesson under that name so

Table 5: Automatic-gate yields (%) by cognitive level (P/U/R) and overall; *Solver rejection* rows are shares of solver inputs, the rest pass.

(%)	P	U	R	All
Evidence pass	69	90	88	88
Solver rejection				
Blind	9	32	41	37
Text-only	21	48	37	41
Unsolvable	19	6	16	12
Survival	29	11	5	8

Critic Memory Bank

PERCEPTION

Make questions genuinely audio-dependent: if the transcript already states the answer, paraphrases the intent, or directly explains the cause, the item is text-solvable and invalid. Target judgments that require tone, delivery, timing, emphasis, or cross-segment integration, and use distractors that are all plausible, transcript-grounded alternatives rather than obvious mismatches or unsupported negatives.

UNDERSTANDING

For shift- or transition-tracking questions, do not make the correct answer stand out by mirroring the stem, explicitly naming the target state, or using a different type of cue than the distractors. All options should be parallel, neutral timeline anchors from the same local context and level of abstraction, so the solver must identify the transition from the source's actual sequence rather than from lexical or structural giveaways.

REASONING

In causal multiple-choice questions, avoid making the correct answer the only option that echoes the stem's core contrast or wording. Instead, make all options plausible competing causes for the same outcome, matched in semantic family, relevance, and abstraction level, so the item tests inference rather than wording alignment.

Figure 8: Examples from the critic's memory bank, grouped by cognitive level. Text is quoted verbatim from the stored lessons.

later generation is steered away from the specific pattern rather than by a generic warning. Figure 8 shows a few examples from the resulting bank, grouped by cognitive level.

The gates are the dominant filter on candidate quality. Table 5 reports their pass and rejection rates on a construction run: evidence verification and the solver together retain roughly 8% of candidates, and most of what the solver rejects is shortcut-solvable, removed by the blind or text-only level. The three cognitive levels fail at different gates: perception items are most often caught at evidence verification for citing events absent from the audio, understanding items at the text-only level for being answerable from the transcript, and reasoning items at the blind level for being guessable from priors, which leaves reasoning the lowest survival rate. The memory bank's effect is visible within the run: candidates generated after the bank has accumulated lessons clear the solver gate more often than early ones, with the solver pass rate rising from roughly 8% to 12% and end-to-end survival from roughly 7% to 9%, while the evidence-verification rate barely moves. The gain sits at the solver gate the bank targets, consistent with its lessons steering generation away from the shortcuts that gate catches.

Human review runs item by item against the audio. The reviewer sees the open-ended question with its reference answer and rubric criteria, and the multiple-choice stem with its options and marked answer; timestamps in the reference answer and criteria point to the stretch of audio the item is about. The reviewer plays that stretch, checks the open-ended item first and then the multiple-choice one, and either accepts the item, applies a minor fix such as tightening a stem, an option, or a criterion, or rejects it when a factual error leaves it beyond repair. For anchor items the reviewer also plays the edited audio to confirm the anchor is audible, matches its stated operation, and cannot be confused with another event already in the audio.

C Evaluation Protocol

C.1 Evaluation Setup

LongAudioSpan is run under a single input protocol: native audio only. A model that cannot ingest the whole audio still receives audio, truncated to its limit, never a text transcript as a fallback, so every score measures audio understanding rather than a cascade. Each model carries one *max processing duration*, the **Limit** column of Table 3, and for an audio sample it receives $\min(\text{limit}, \text{length})$ as the leading prefix, taken content-agnostically rather than centered on where the answer lies.

Table 6: Audio-input limit per evaluated model and the specs behind it, expanding the **Limit** column of Table 3. **Context** is the trained position range; **Tokens/s** is positions per second into the language model after audio pooling; **Limit** is the applied truncation cutoff, with – for a model that runs untruncated on every tier.

Model	Context	Tokens/s	Basis for limit	Limit
Open-source Models				
Phi-4-Multimodal (Abouelenin et al., 2025)	131,072	–	Reported ceiling	30 min
AudioFlamingoNext (Ghosh et al., 2026)	131,072	–	Reported ceiling	30 min
Voxtral-Mini (Liu et al., 2025)	32,768	12.5	Reported ceiling	40 min
Voxtral-Small (Liu et al., 2025)	32,768	12.5	Reported ceiling	40 min
MOSS-Audio (Yang et al., 2026a)	40,960	13	Position budget	50 min
Baichuan-Omni-1.5 (Li et al., 2025)	65,536	12.5	Position budget	80 min
Qwen3-Omni (Xu et al., 2025b)	65,536	13	Position budget	80 min
Closed-source Models				
Gemini-2.5-Pro	API	–	Per-request cap 9.5 h	–
Gemini-3-Flash	API	–	Per-request cap 9.5 h	–
Gemini-3.1-Pro	API	–	Per-request cap 8.4 h	–
Qwen3.5-Omni-Flash	API	–	Per-request cap 3 h	–
Qwen3.5-Omni-Plus	API	–	Per-request cap 3 h	–

Alongside the audio, each item carries a fixed one-line instruction prepended to the question, set by answer format. Multiple-choice items (Accuracy and Chain) constrain the reply to a single option letter, matched exactly against the key; open-ended items (Rubric) instead ask for audio timestamps, the citations that temporal% scores, though omitting one is not penalized (App. C.2). The two templates:

Model Prompt

```
MCQ
Listen carefully to the audio and answer the following question. Answer with ONLY the
option letter.

OEQ
Listen carefully to the audio and answer the following question. Cite audio timestamps in
your analysis.
```

The limit follows a fixed priority order: a stated audio ceiling in the model card or technical report is taken directly; otherwise the limit is the largest audio length that fits the model's trained context, computed from the context window, the reserved text budget, and the audio encoding rate. The context is the trained range alone. The applied value is the highest length that runs cleanly under this ceiling, dropping where output degrades earlier (MOSS-Audio produces gibberish past roughly 50 min). Closed models run through APIs whose per-request duration caps clear the longest tier, so those models run untruncated. Table 6 collects the per-model figures.

C.2 Scoring Details

Each criterion c carries an importance weight w_c (1 for essential, 0.5 for supporting) and is graded by the judge ensemble on two dimensions. The *semantic* score $\sigma_c \in \{0, 0.5, 1\}$ marks the required content as wrong, partial, or correct. The *temporal* score $\tau_c \in \{0, 1\}$ marks whether a predicted timestamp falls within the criterion's reference time range under a ± 5 s tolerance; timestamps in any format are normalized to seconds before comparison.

Table 7: Supporting-criterion score $s_c = \mathbf{1}[\sigma_c{>}0]\,(\sigma_c{+}\tau_c)/2$ by semantic verdict σ_c and temporal verdict τ_c.

	Temporal τ_c		
Semantic σ_c	correct	wrong	none
correct (1)	1.0	0.5	1.0
partial (0.5)	0.75	0.25	0.5
wrong (0)	0	0	0

Which dimensions apply follows Eq. 2. An essential criterion is graded on one dimension: τ_c for a localization criterion, which asks when something occurs, and σ_c otherwise. A supporting criterion takes the gated mean of σ_c and τ_c when it carries a reference time range, and σ_c alone otherwise (Table 7). Two conventions complete the rule: a semantically wrong criterion scores $s_c = 0$ regardless of timing, since a timestamp cannot ground a false statement; and a missing timestamp falls back to σ_c rather than incurring a penalty, because answers are not required to cite time.

Because σ_c and τ_c are logged separately, we read the rubric along each axis for the analysis of App. D.4 as two per-criterion sub-scores, *semantic%* and *temporal%*. Write $N_\sigma(\cdot)$ and $N_\tau(\cdot)$ for counts of semantic and temporal verdicts: a semantic verdict is *correct*, *partial*, or *wrong* when a claim is demanded and *none* on a pure-localization criterion; a temporal verdict is *correct* or *wrong* when a timestamp is given, *missing* when one is due but omitted, and *none* when no timing is demanded. Semantic% is the mean semantic score over claim-demanding criteria, and temporal% is the accuracy over attempted timestamps:

$$\text{semantic\%} = \frac{N_\sigma(\text{cor}) + \frac{1}{2}N_\sigma(\text{par})}{N_\sigma(\text{cor}) + N_\sigma(\text{par}) + N_\sigma(\text{wro})}, \qquad \text{temporal\%} = \frac{N_\tau(\text{cor})}{N_\tau(\text{cor}) + N_\tau(\text{wro})}. \quad (4)$$

Semantic% excludes *none* (no claim demanded); temporal% excludes both *none* (no timing demanded) and *missing* (a timestamp due but omitted), so it scores localization conditional on an attempt and separates it from both content and silence.

Temporal% is conditioned to match how each criterion type is scored. On an essential localization criterion the timestamp *is* the answer (σ_c is *none*), so temporal% is unconditioned. On a supporting criterion the timestamp is a citation and the gate $\mathbf{1}[\sigma_c > 0]$ zeros the whole criterion when the claim is wrong, so timing bears on the score only when the claim is right; supporting temporal% therefore counts N_τ over the rows where σ_c is *correct* alone.

The two sub-scores compose back into the rubric differently by importance. An essential criterion is graded on one axis, so its mean score is the share-weighted average $p_\sigma\,\text{semantic\%} + p_\tau\,\text{temporal\%}$, where p_σ, p_τ are the fractions of essential criteria graded on the semantic and temporal axis (about 0.73 and 0.27); this is a partition up to the $\approx$0.2% of essential criteria the judge grades on both. A supporting criterion is gated and dual-axis: semantic% drives its score and temporal% enters only as the halving penalty of Eq. 2 on the claims it grades correct, so the essential–supporting gap of App. D.4 is the semantic–temporal gap re-expressed through criterion importance. App. D.4 reads this semantic vs. temporal split model by model and by importance.

C.3 Judge Prompt

Each judge in the ensemble receives the system prompt below, followed by the question, the criterion list (each with `text`, `importance`, and an optional `time_range`), and the model's response, and returns one `semantic`/`temporal` verdict pair per criterion.

Judge System Prompt

```
You are an audio-understanding evaluation expert. For each evaluation criterion
provided, judge the quality of the model's response independently.

## Evaluation Dimensions

### Dimension Selection Rules

Choose which dimension(s) to evaluate based on the criterion's `importance`
and its content:

**essential criterion: evaluate exactly one dimension**
- If the criterion requires a specific time point or time span (localization),
  evaluate **temporal** only; set semantic to `"none"`.
- If the criterion requires content, meaning, causality, etc. (semantic), evaluate
  **semantic** only; set temporal to `"none"`.

**supporting criterion: evaluate both dimensions when applicable**
- If a `time_range` is provided: evaluate both **semantic** and **temporal**.
- If no `time_range` is provided: evaluate **semantic** only; set temporal to
  `"none"`.
```

````
### semantic

- **correct** = the response fully captures what the criterion requires; the key
  information is complete and accurate.
- **partial** = the response touches on part of the content, but key information is
  incomplete or contains errors.
- **wrong** = the response does not address what the criterion requires, or the core
  content is incorrect.
- **none** = the criterion is a localization task; semantic is not evaluated.

### temporal

- **correct** = the timestamp given by the model falls within the valid range (+/-5
  seconds tolerance around the target), or the time span given by the model overlaps
  the reference `time_range`.
- **wrong** = the model produced a timestamp or span, but it is outside the valid
  range and does not overlap.
- **missing** = the model did not produce a timestamp (the model is not required to
  do so).
- **none** = the criterion is semantic; temporal is not evaluated.

The model may use different timestamp formats. Normalize before comparing:
- `MM:SS` (e.g. `12:30`) = 12 minutes 30 seconds.
- Raw seconds (e.g. `522.640s`) = 8 min 42 sec = `08:42`.
- Out-of-range minutes (e.g. `64:00`) = `01:04:00`.

Format differences alone do not affect the judgment.

## Judgment Principles

- Focus on whether the core semantics are correct; different wording that expresses
  the same meaning is acceptable.
- Evaluate **semantic** and **temporal** independently: semantic judges only whether
  the content itself is correct, regardless of whether a timestamp is cited.
  Timestamp citation is judged under temporal.
- Each criterion is judged on its own requirements only. Details mentioned in other
  criteria apply to those criteria alone and do not affect this one.
- A criterion is satisfied as long as the core content is present and correct.
- Features listed in a criterion are examples; citing the key feature is sufficient.
- Keep `reason` and `summary` concise: 1-2 sentences each.

## Input

You will receive:
1. The question (open-ended QA).
2. A list of evaluation criteria (each with `text`, `importance`, and optionally
   `time_range`).
3. The model's response.

## Output Format

```json
{
  "criteria": [
    {
      "index": 0,
      "semantic": "none",
      "temporal": "correct",
      "reason": "The model timestamp 14:02 is within +/-5s of 14:03."
    },
    {
      "index": 1,
      "semantic": "correct",
      "temporal": "none",
      "reason": "The model correctly identifies the motive as loyalty formed from early
allegiance to Sun Yat-sen."
    },
    {
      "index": 2,
      "semantic": "partial",
      "temporal": "wrong",
      "reason": "Semantic: the model mentions the vocal features but incompletely. Temporal
: the cited timestamp is outside 00:41-00:45."
    }
  ],
  "summary": "A one-sentence summary of the overall quality of the response."
}
```
````

Table 9: Per-language scores (English vs. Chinese), each averaged over the three duration tiers. Formatting as in Table 3.

			Accuracy (%)			Rubric (%)			Chain (%)		
Model	**Size**	**Think**	**EN**	**ZH**	$\Delta_{\text{EN}-\text{ZH}}$	**EN**	**ZH**	$\Delta_{\text{EN}-\text{ZH}}$	**EN**	**ZH**	$\Delta_{\text{EN}-\text{ZH}}$
Open-source Models											
Phi-4-Multimodal	5.6B	✗	38.3	30.0	+8.4	23.3	4.9	+18.4	10.0	9.2	+0.8
AudioFlamingoNext	8B	✗	47.8	36.9	+10.9	19.3	8.7	+10.6	5.9	5.4	+0.6
		✓	39.8	38.3	+1.5	25.5	16.8	+8.8	10.6	10.6	0.0
Voxtral-Mini	3B	✗	41.7	37.2	+4.4	26.4	14.3	+12.2	13.5	8.7	+4.8
Voxtral-Small	24B	✗	40.2	36.1	+4.1	29.5	19.7	+9.8	6.5	8.2	−1.7
MOSS-Audio	8B	✗	43.1	39.3	+3.9	20.1	14.6	+5.5	8.3	10.9	−2.6
		✓	49.4	46.3	+3.2	19.5	18.6	+0.9	14.3	12.8	+1.5
Baichuan-Omni-1.5	11B	✗	45.9	40.9	+5.0	15.5	12.0	+3.5	12.4	15.0	−2.6
Qwen3-Omni	30B (A3B)	✗*	50.5	52.5	−2.0	22.4	28.4	−6.0	12.9	22.4	−9.5
		✓	48.1	31.1	+17.0	30.2	30.0	+0.2	9.8	5.9	+3.9
Closed-source Models											
Gemini-2.5-Pro	–	✓	63.3	60.9	+2.4	43.2	39.4	+3.9	22.0	25.6	−3.5
Gemini-3-Flash	–	✗	69.4	63.1	+6.3	**55.2**	**53.9**	+1.3	18.3	14.4	+3.9
Gemini-3.1-Pro	–	✓	**75.9**	68.7	+7.2	49.7	49.1	+0.5	22.8	19.1	+3.7
Qwen3.5-Omni-Flash	–	✗	70.4	65.7	+4.6	43.2	41.5	+1.7	19.4	20.9	−1.5
Qwen3.5-Omni-Plus	–	✗	74.4	**74.8**	−0.4	49.3	52.4	−3.0	**29.8**	**28.1**	+1.7

C.4 Models and APIs

Generation quality is enforced by the downstream verification and review stages (§4.5), and the roles split by modality. The audio-facing stages (speech transcription, acoustic captioning, and audio verification) run on audio-capable models, while question generation and judging run on text-only language models. Table 8 lists the models used in this release.

Table 8: Construction-pipeline models.

Role	**Model**
Speech transcription (ASR)	`qwen3-asr-flash`
Acoustic captioning	`gemini-3.1-pro`
Language model (generation)	`qwen3.7-plus`
Audio verification	`qwen3-omni-flash`
Language model (judge)	`gpt-5.4`

D Additional Results

D.1 Performance Across Languages

The direction and magnitude of the language gap are model-dependent. Splitting the leaderboard by language (Table 9), most models attain higher multiple-choice accuracy in English, though the magnitude varies substantially and the direction is not uniform. The gap is largest in open-ended generation for several open models that answer markedly better in English than in Chinese; Phi-4-Multimodal is the extreme case, scoring near zero on Chinese Rubric. The Chinese-origin Qwen-Omni family exhibits the opposite tendency: Qwen3.5-Omni Plus, the strongest model overall, is balanced on multiple-choice accuracy and slightly favours Chinese under Rubric. The gap can also be format-dependent within a single model: one Qwen3-Omni variant favours English strongly on multiple choice yet remains balanced under Rubric. Chain gaps are the smallest and direction-less: sustaining the full P→U→R chain is hard in either language, so scores stay low and neither pulls far ahead.

D.2 Performance Across Dimensions

Perception is the weakest level under both metrics, with localization and counting its low dimensions. Averaged over the five models, perception scores below understanding and reasoning on

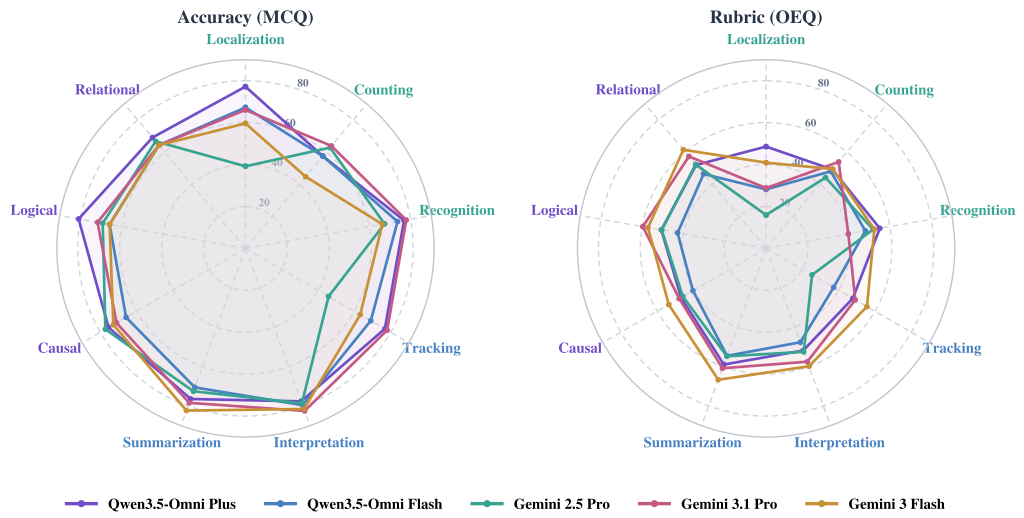


Figure 9: Accuracy (left) and Rubric (right) across the nine dimensions for the five closed-source models. Axis labels are coloured by cognitive level: **Perception**, **Understanding**, **Reasoning**. These three label colours coincide with three of the model line colours, but the labels sit on the rim and the lines in the field, so the two do not compete.

both panels (Accuracy 63.8%, Rubric 43.6%), the same ordering as §5.3. Within it, localization and counting are the two low dimensions and recognition keeps pace with the stronger levels.

Localization is the dimension multiple-choice most overstates, and counting the one it cannot lift. The two lowest dimensions under Rubric, localization and tracking, both turn on placing events in time. Localization keeps only 53% of its Accuracy score under open-ended grading, by far the steepest drop of any dimension, because multiple-choice lets a model pick a timestamp it cannot itself produce. Counting is the opposite: already low under Accuracy and barely lower under Rubric, a genuine capacity limit the easier format cannot lift. App. D.4 traces localization's collapse to the rubric's temporal axis. The same signature separates models: Gemini 2.5 Pro dips at localization and tracking while matching the field elsewhere, and Gemini 3.1 Pro closes that notch.

D.3 Question Formats: Multiple-Choice vs. Open-Ended

On item-matched native questions, multiple-choice success is a weak proxy for grounded comprehension. Because every native item is posed in both formats, we can measure this directly. For the five closed-source models we form the matched set M of items scored in both formats and, for each item i, record its MCQ correctness $a_i \in \{0, 1\}$ (Eq. 1) and its open-ended Rubric score r_i (Eq. 2). Splitting M by the multiple-choice outcome into $M^+ = \{i : a_i = 1\}$ and $M^- = \{i : a_i = 0\}$, we report the mean Rubric within each group,

$$\text{Rubric}^{\pm} = \frac{1}{|M^{\pm}|} \sum_{i \in M^{\pm}} r_i. \tag{5}$$

The overall Rubric is the weighted average of the two: with $p = |M^+|/|M|$ the MCQ-correct rate, $\text{Rubric} = p\,\text{Rubric}^+ + (1-p)\,\text{Rubric}^-$. We then summarize how far the multiple-choice outcome separates open-ended quality by the probability that a correct item outscores a wrong one on the rubric, i.e. the area under the ROC curve,

$$\text{AUC} = \frac{1}{|M^+|\,|M^-|} \sum_{i \in M^+} \sum_{j \in M^-} \Big(\mathbf{1}[r_i > r_j] + \tfrac{1}{2}\,\mathbf{1}[r_i = r_j] \Big), \tag{6}$$

where 0.5 means the multiple-choice outcome is uninformative about the rubric score and 1 that every correct item outscores every wrong one.

Table 10: Per-model average scores on the item-matched native set for the five closed-source models (matched items carry at least one essential criterion). $\mathrm{Rubric}^{\pm}$ and AUC are defined in Eq. 5 and Eq. 6. Accuracy, Rubric, and $\mathrm{Rubric}^{\pm}$ are in percent; AUC is a probability in $[0, 1]$.

Model	Accuracy	Rubric	Rubric^{+}	Rubric^{-}	AUC
Gemini 2.5 Pro	62.0	41.2	51.1	25.1	0.71
Gemini 3 Flash	66.2	54.4	59.9	43.7	0.62
Gemini 3.1 Pro	72.4	49.3	54.9	34.4	0.65
Qwen3.5-Omni Flash	68.1	42.3	46.7	32.9	0.61
Qwen3.5-Omni Plus	74.5	50.8	56.0	35.8	0.66
Pooled	68.6	47.6	53.8	34.1	0.65

Pooled over the five closed-source models, open-ended answers to the very items answered correctly in multiple choice average only $\mathrm{Rubric}^{+} = 53.8$, far short of full credit, while the items answered incorrectly still average $\mathrm{Rubric}^{-} = 34.1$; the multiple-choice outcome separates the two by an AUC of just 0.65, barely above the 0.5 floor, and stays within 0.62–0.66 across cognitive levels. Every closed-source model shows the pattern (Table 10), though how it fails varies: Gemini 2.5 Pro separates its correct and wrong items most cleanly (AUC 0.71) yet still reaches only $\mathrm{Rubric}^{+} = 51.1$ on the items it answers correctly, while Gemini 3 Flash pairs the highest Rubric^{+} (59.9) with the highest Rubric^{-} (43.7), making its multiple-choice outcome the least informative about open-ended quality (AUC 0.62). No model both produces strong open-ended answers on the items it gets right and cleanly separates them from the ones it gets wrong.

The clear cases divide into four cells, and disagreement runs mostly one way. Scoring each open-ended answer as passing (Rubric ≥ 0.8) or failing (Rubric ≤ 0.2) and setting the partial middle aside (about 39% of pairs), the matched (item, model) pairs split into four cells: both pass (21%), both fail (15%), correct only in multiple choice (19%), and correct only open-ended (6%). The multiple-choice-only cell outnumbers its mirror by roughly three to one. Figures 10 through 12 give one clean item from three of the cells, each evaluated on the same five models.

The disagreement is a gap between recognizing an answer and producing it. A perception localization item makes this concrete (Figure 11). With four candidate timestamps on offer, all five models select the right one; with the options withheld, none produces the true time and every open-ended answer scores zero on the localization rubric. Two models even quote the exact sentence that opens the queried event, locating the right moment in content while failing to place it in time.

The other three cells show the effect is specific, and can even reverse. When a model genuinely understands an item, the formats agree: on the acoustic-environment item of Figure 10 all five pick the correct option and all five give the same correct open-ended description, scoring full credit. The reverse cell runs the other way (Figure 12): on a summarization item all five name the right answer open-ended, yet four of them choose a more dramatic distractor over that same answer in multiple choice, so the options mislead a model that can produce the answer. The both-failing cell holds items too hard for either format, such as an event no model finds in either form, where the shortcut simply does not arise.

D.4 Rubric Axes: Semantic vs. Temporal

Open-ended performance is bounded by temporal grounding. Each criterion carries a *semantic* verdict (what happened) and a *temporal* verdict (when), scored as in §C.2. Figure 13 plots both across the duration tiers, one line per model. Semantic accuracy is high and clustered (43–65%) and drifts down only gently with length. Temporal accuracy is the weak, length-fragile axis: it starts lower and falls as the audio grows. The per-model shapes carry the story: Gemini 2.5 Pro is a pure localization failure, placing events in time correctly in single digits at M/L while still describing them above 54%; Qwen3.5-Omni Plus stays the most grounded, its temporal tracking its semantic; Gemini 3.1 Pro holds temporal through M then collapses at L; and Gemini 3 Flash alone keeps temporal flat across tiers, which is why its overall rubric score does not fall. The deficit is confident misplacement, not silence: pooled over the other four closed-source models, correct timing drops

Both correct: item L_EN_057, perception recognition

```
AUDIO      1:05:51 total; a whole-recording judgment, no timestamp needed
QUESTION   What best describes the acoustic environment of the recording?
OPTIONS    A. outdoor, wind and traffic     B. dry close-mic studio (key)
           C. live stadium, distant crowd   D. reverberant hall with echo
CRITERIA   essential: identifies a dry, clean, close-mic studio setting
           supporting: notes the absence of reverb, echo, and background noise

MULTIPLE-CHOICE   scored against the key: all five models pick B (correct)
OPEN-ENDED    options withheld, scored against the rubric: all five score full credit
  Gemini 2.5 Pro      dry studio, a podcast or treated room
  Gemini 3 Flash      a professional, dry studio setting
  Gemini 3.1 Pro      a small room or studio, clear voices, no echo
  Qwen3.5-Omni Flash  quiet indoor, minimal ambience, no reverb
  Qwen3.5-Omni Plus   quiet indoor space, minimal reverberation
  all five describe the dry studio unaided: recognition here needs no options
```

Figure 10: A perception recognition item where the two formats agree: all five models pick the correct option, and all five give the same open-ended description for full credit. Model descriptions are abridged.

Multiple-choice only: item S_EN_038, perception localization

```
AUDIO      10:27 total; the 08:12 answer sits late in the recording
QUESTION   At what time does the first customer testimonial begin speaking
           after the main presenter?
OPTIONS    A. 08:24      B. 09:10      C. 08:12 (key)      D. 06:42
CRITERIA   essential: names the switch to a distinct testimonial speaker at 08:12
           supporting: locates the presenter's speech just before the switch

MULTIPLE-CHOICE   scored against the key: all five models pick C (correct)
OPEN-ENDED    options withheld, scored against the rubric: all five score 0
  Gemini 2.5 Pro      04:55  "'We presently have four of these units ...'"
  Gemini 3 Flash      08:44  (quotes the same opening line)
  Gemini 3.1 Pro      01:14
  Qwen3.5-Omni Flash  08:31
  Qwen3.5-Omni Plus   05:28
  none states 08:12: each locates the event but cannot place it in time
```

Figure 11: A perception localization item where the two formats disagree: all five models pick the correct timestamp, none produces it open-ended, and two quote the right utterance at the wrong time. Model text is quoted verbatim and abridged.

from 42% at S to 23% at L while wrong timing rises from 46% to 67%, and outright omissions stay rare and flat (10–13%).

The same gap explains the importance ordering. Every criterion is tagged *essential* (the must-have answer, weight 1) or *supporting* (a corroborating detail, weight 0.5), and models score higher on essential criteria (50.7, same pool) than supporting ones (42.5). That reads like models getting the core right and missing the extras, but it is not: semantic accuracy is nearly identical on the two levels (58.8 vs. 52.9). What differs is timing load. Almost every supporting criterion demands a timestamp (92%), whereas most essential criteria do not (only 27% are graded on time), so the timing-heavy level inherits the weak temporal axis and scores lower. It is also where Gemini 3 Flash separates: its supporting score holds flat with duration, so its essential-supporting gap stays near zero.

Underneath these means, the per-question rubric scores are close to all-or-nothing (Figure 14). Every model's distribution is U-shaped: the two largest bins are the extremes, $[0, 0.1)$ and $[0.9, 1.0]$, and the middle bins are the emptiest. Models differ mainly in how the mass splits between the poles. The strongest place the most at the top: Gemini 3 Flash scores 26% of questions in the top bin against 18% in the bottom, whereas the confident misplacer Gemini 2.5 Pro reaches the top bin on only 9% and piles 25% at the bottom.

Open-ended only: item L_EN_028, understanding summarization

```
AUDIO      1:01:10 total; the narrator's verdict sits in the closing segment
QUESTION   In the concluding segment, what does the narrator call the primary
           missed opportunity that could have altered Marty Graham's path?
OPTIONS    A. no mental-health treatment while employed (key)
           B. courts rejected the personality defense earlier
           C. no break in the foster-care abuse cycle before homelessness
           D. community missed signs of drug addiction
CRITERIA   essential: names the failure to provide mental-health treatment
           supporting: ties it to thin support for parents of ill children

MULTIPLE-CHOICE   scored against the key: all five miss it; four pick C, one picks D
OPEN-ENDED    options withheld, scored against the rubric: all five score near full
  Gemini 2.5 Pro      failure to provide accessible mental-health care
  Gemini 3 Flash      lack of resources and early intervention
  Gemini 3.1 Pro      lack of care and support for the parents
  Qwen3.5-Omni Flash  failure to give early intervention and support
  Qwen3.5-Omni Plus   lack of resources for parents raising such children
  all five name the treatment gap unaided; the options steer four of them to C
```

Figure 12: A summarization item where the disagreement reverses: all five miss the correct option, four choosing a distractor, yet all five name the right answer open-ended. Model answers are abridged.

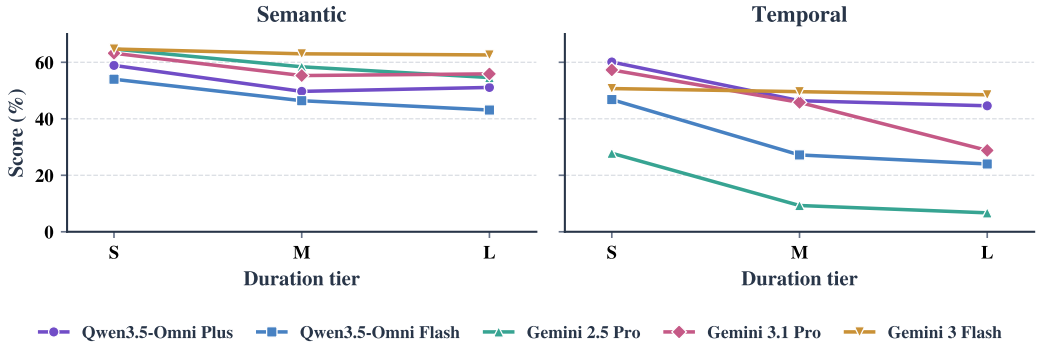


Figure 13: Semantic vs. temporal accuracy by duration tier, per closed-source model, en+zh merged. *Semantic%* is the mean per-criterion content score; *temporal%* is timestamp accuracy.

Table 11: Per-config Anchor-QA link rates and the grounding split, pooled over tiers and languages. P/U/R are per-link accuracy; Chain the truncated score (Eq. 3). $U\,|\text{loc.}$ and $U\,|\text{miss.}$ are understanding accuracy with the anchor located or missed (read as located − missed).

Config	P	U	R	Chain	$U\,\|$loc.	$U\,\|$miss.
Gemini 2.5 Pro	38.3	48.6	42.5	23.8	54.3	45.0
Gemini 3 Flash	29.7	46.9	36.1	16.4	45.8	47.4
Gemini 3.1 Pro	38.3	48.1	39.2	20.9	42.8	51.4
Qwen3.5-Omni Flash	38.1	39.7	35.8	20.2	41.6	38.6
Qwen3.5-Omni Plus	47.8	47.2	41.7	29.0	54.1	41.0

D.5 Chain Scoring: Grounding vs. Guessing

Understanding and reasoning are answered without locating the anchor. On each chain we split understanding and reasoning accuracy by whether the perception link was correct, that is, whether the anchor was located or missed on that same chain; when the missed-anchor rate matches the located rate, the credit does not come from grounding. Pooled over the five closed models (Table 11), understanding is answered 44.9% of the time on chains whose anchor was never located, against 48.1% when it was, a three-point grounding premium; reasoning is similarly flat (37.5 vs. 41.6). The split does separate models: understanding tracks grounding most for Qwen3.5-Omni Plus (54.1

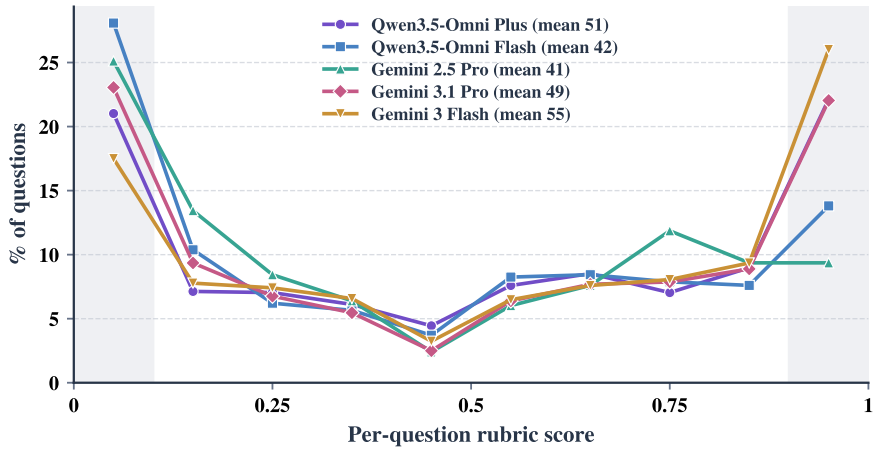


Figure 14: Per-question rubric score distributions, pooled over the S/M/L tiers, per closed-source model (en+zh merged): the % of questions in ten equal-width bins over $[0, 1]$, pole bins $[0, 0.1)$ and $[0.9, 1.0]$ shaded. The legend gives each model's mean.

located vs. 41.0 missed, +13.1) and inverts for Gemini 3.1 Pro thinking (-8.6), which answers understanding *more* often when it missed the anchor, the signature of a guesser.

First-error truncation floors a guesser and credits only grounding. The text-only solvers (§5.1) trace what the chain score responds to. A blind solver clears each link 22–34% of the time on its own, yet scores only 10.4 on the chain, because ungrounded guesses cannot be collected into a correct prefix. Adding the ASR transcript barely moves it (12.6). Only injecting the anchors into the caption (the oracle of Figure 4b) recovers the chain, to 61.4. The score therefore tracks temporal grounding specifically, not verbal content, and resists the per-link guessing that inflates the link rates.

The guessing signature is strongest for the operation that is hardest to locate. The three anchor operations differ in how findable they leave the target (Table 12). Pooled over the five closed models, perception is highest for Modification (43.8), then Addition (38.7), then Deletion (32.8): a silenced span is the hardest anchor to find and an altered but still-present one the easiest, an ordering that holds for every model and every duration tier. Yet reasoning peaks under Deletion (42.5) despite its worst perception, the one operation where a higher layer scores above its own perception link, and its chain score stays lowest (19.2): the operation hardest to ground is exactly where the higher layers answer without it.

Table 12: Anchor-QA link rates and chain score by anchor operation, pooled over the five closed models, tiers, and languages. Addition inserts a sound event, Deletion silences a span, Modification alters acoustic attributes. P/U/R are per-link accuracy; Chain the truncated score (Eq. 3).

Operation	P	U	R	Chain
Addition	38.7	49.3	34.0	22.6
Deletion	32.8	45.0	42.5	19.2
Modification	43.8	44.0	40.7	24.3

E LIMITATIONS

The two question paths are complementary, yet each carries its own constraint. For Native QA, coverage is limited by caption fidelity: items can only draw on what the caption records, and the bound tightens on the longest audio, where omissions are hardest to catch and human review cannot recover what was never written. For Anchor QA, synthetic anchors provide deterministic ground truth, but confine questions to the grounding-to-reasoning chain template, sacrificing the free-form diversity that Native QA offers. Beyond these design tradeoffs, the benchmark covers two languages (English and Chinese) and in-the-wild speech; low-resource languages, cross-lingual questions, and long music or environmental recordings lie outside its scope.