AUDIOTRACE: EVIDENCE-GROUNDED QUESTION ANSWERING OVER LONG AUDIO

Anonymous authors
Paper under double-blind review

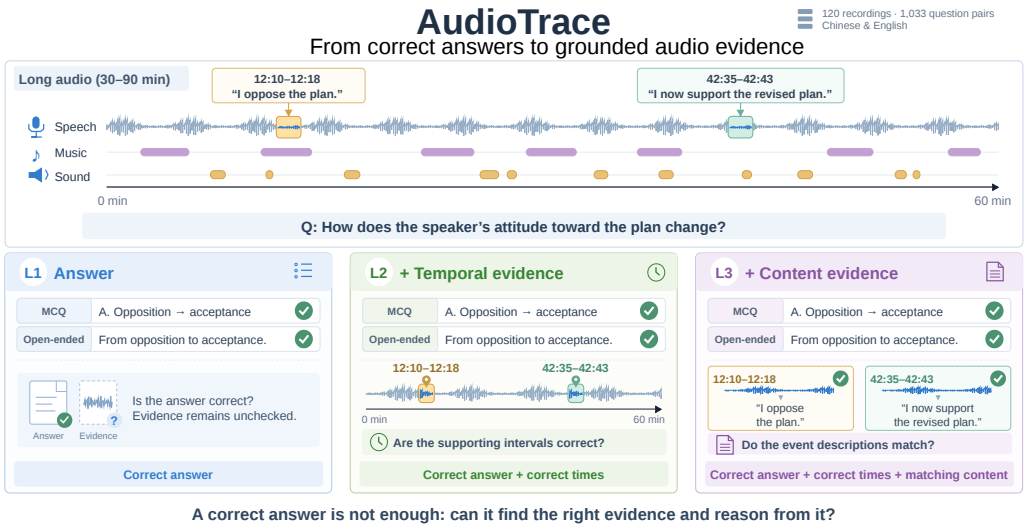


Figure 1: Overview of AudioTrace for answer–evidence disentanglement. Three levels assess answer-only correctness (L1), answer correctness with temporal evidence localization (L2), and evidence-supported correctness through temporal localization and semantic evidence grounding (L3). Each level includes both multiple-choice and open-ended QA.

ABSTRACT

Evaluations of large audio-language models (LALMs) have expanded beyond minute-scale clips to include recordings lasting several hours. Yet current long-audio QA evaluations face two key limitations. (1) Inflated scores can mask weaknesses in fine-grained audio understanding and reasoning. (2) Answer correctness is often assessed without verifying whether models identify the precise temporal spans and factual content supporting their predictions. To address these limitations, we introduce AudioTrace, a benchmark that disentangles answer correctness from evidence grounding in long-audio QA. The dataset comprises 120 Chinese and English recordings lasting 30–90 minutes each (99.8 hours total), with 1,033 pairs of multiple-choice and open-ended questions across five question categories. Our construction pipeline combines automated annotation with human review, yielding event-level temporal and content evidence for each answerable question. The evaluation protocol has three levels with independently generated responses: L1 assesses answer correctness alone; L2 additionally requires temporal localization of supporting evidence; and L3 further requires semantic grounding of that evidence. Experiments reveal a substantial gap between answer correctness and evidence-supported correctness. Across 17 model–paradigm configurations, mean L1 answer accuracy is 84.8% for multiple-choice QA and 43.8% for open-ended QA, whereas mean L3 joint accuracy is 22.1% and 10.9%, respectively. These results show that a correct answer alone does not establish accurate understanding of the supporting audio or reliable reasoning from it. Evaluating trustworthy long-audio QA therefore requires finer-grained checks of the temporal location and factual content of supporting evidence. The code and dataset are publicly available at `https://anonymous.4open.science/r/open_audiotrace`.

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 AudioTrace with representative audio understanding benchmarks.

		QA format		Construction		Quality control		Task & evaluation	
Benchmark	Duration	MCQ	Open	Source	Pipeline	Filter	Refine	Task	Eval.
MMAU (Sakshi et al., 2025)	∼10 s	✓	×	Corpus	Manual	Manual	×	Multi	Answer
MMAU-Pro (Kumar et al., 2025)	∼2.1 min	✓	✓	Wild	Manual	Manual	×	Multi	Answer
LongAudioBench (Ghosh et al., 2025)	∼2 min	✓	✓	Corpus	Auto	Multi	×	Multi	Answer
AudioMarathon (He et al., 2025)	∼3.5 min	✓	○	Corpus	Auto	Manual	×	Multi	Task
ChronosAudio (Luo et al., 2026)	∼6 min	×	○	Synthetic	Auto	×	×	Multi	Task
LongSpeech-Eval (Guo et al., 2025)	∼133 s	×	✓	Synthetic	Auto	Auto	×	Single	Answer
LongSpeech (Yang et al., 2026b)	∼10 min	×	○	Corpus	Auto	Manual	×	Multi	Task
LAT-Bench (Shao et al., 2026)	≤30 min	×	○	Wild	Auto	×	×	Multi	Task
BLAB (Ahia et al., 2025)	∼51 min	✓	○	Wild	Manual	Manual	×	Multi	Task
VoiceGiraffe (Ye et al., 2026)	∼55.2 min	✓	×	Wild	Auto	Multi	×	Multi	Answer
AudioSpan (Huang et al., 2026b)	∼50 min	✓	✓	Wild	Auto	Multi	✓	Multi	Answer
Vox-Infinity (Cheng et al., 2026)	∼5.2 min	✓	✓	Mixed	Auto	Multi	✓	Multi	Answer
MuLA-Bench (Yang et al., 2026c)	∼46.7 min	×	✓	Wild	Auto	Multi	✓	Multi	Answer
AudioTrace (ours)	**∼49.9 min**	✓	✓	Wild	Auto	Multi	✓	Multi	**Levels**

○: fixed-target or structured generation (e.g., ASR, counting, or timestamp prediction), rather than free-form QA.

1 INTRODUCTION

Large audio-language models (LALMs) have developed rapidly. Their capabilities have expanded from recognition and description to open-ended question answering and multi-step reasoning (Chu et al., 2024; Goel et al., 2025; Xu et al., 2025b). As these models expand from processing short clips to recordings lasting several hours, research is increasingly focusing on long-audio understanding and reasoning (Gemini Team et al., 2024; Ghosh et al., 2026).

Audio benchmarks have evolved alongside these models. One line broadens the skills under evaluation, assessing comprehension and reasoning across speech, music, and sound, primarily using short clips (Yang et al., 2024; Sakshi et al., 2025; Ma et al., 2025; Kumar et al., 2025). A complementary line extends evaluation from short clips to long recordings, testing audio understanding over substantially longer temporal contexts (Hegde et al., 2026; Huang et al., 2026b; Dai et al., 2026; Shao et al., 2026; Yang et al., 2026c; Cheng et al., 2026).

Despite this progress, long-audio QA still faces an answer–evidence disentanglement problem: a correct answer does not establish that a model found the relevant audio or used it faithfully. A model may arrive at the right answer from option cues or question-level expectations without locating the underlying audio facts, yielding shortcut-mediated correctness. Conversely, identifying a plausible time span does not establish that the model understood its content, especially when speech overlaps or multiple sounds occur together. A reliable evaluation must therefore distinguish answer-only correctness from evidence-supported correctness by checking both the temporal location and the semantic content of the events cited for an answer.

To address this gap, we introduce **AudioTrace**, a benchmark for evidence-grounded QA over long audio. AudioTrace covers long recordings in Chinese and English and provides paired multiple-choice and open-ended questions. The benchmark has two defining features. **Fine-grained understanding in long contexts:** questions cover event localization, detail understanding, context linking, information aggregation, and answerability judgment, examining how models find, interpret, and combine information within extended recordings. **Explicit supporting evidence:** each answerable question is annotated with supporting event intervals and their content across speech, music, and sound, allowing individual predictions to be checked against identifiable audio events. Automated review and independent listening by three human reviewers provide quality control. Together, these features connect the demands of long-context understanding to evidence that can be inspected at the event level. Table 1 contrasts AudioTrace with existing benchmarks.

As illustrated in Figure 1, AudioTrace introduces **three levels of evaluation, each covering both multiple-choice and open-ended QA**. **L1** measures answer-only correctness. **L2** requires a correct answer together with accurately localized temporal evidence, and **L3** additionally requires the supplied event content to support the answer. This design decouples answer format from evidence requirements, with independent inference at each level. It enables fine-grained evaluation of whether

models can answer correctly, localize the supporting temporal evidence, and ground their answers in the factual content of the supporting events.

Extensive experiments show that current LALMs remain far from reliable on long-audio QA. Even the best-performing model in our evaluation, Qwen3.8 Omni Flash (E2E), sees its multiple-choice and open-ended accuracies fall from 97.0% and 62.1% under answer-only evaluation (L1) to 60.8% and 37.9% when temporal and semantic evidence grounding are also required (L3). These results show that high answer accuracy does not imply reliable temporal evidence localization or semantic evidence grounding. Although models can answer questions correctly, they still struggle to consistently locate and integrate the supporting evidence needed for trustworthy QA.

Our contributions include:

- **The first multi-level benchmark for evidence-grounded long-audio QA.** AudioTrace jointly evaluates answer correctness, temporal evidence localization, and semantic evidence grounding in MCQ and open-ended QA, distinguishing answer-only success from success under explicit supporting-evidence checks.
- **An automated pipeline with fine-grained evidence linking.** We develop a scalable construction pipeline that links each question to timestamped events. Automated validation and independent human review ensure annotation quality.
- **A comprehensive evaluation of current LALMs.** We conduct extensive experiments on current LALMs, revealing a substantial gap between answering correctly and accurately retrieving and interpreting supporting evidence and exposing persistent challenges for trustworthy long-audio QA.

2 Related Work

2.1 Large Audio-Language Models

LALMs connect audio representations to language models for description, instruction following, and question answering across speech, sound, and music (Gong et al., 2023; Tang et al., 2023; Chu et al., 2023; 2024; Kong et al., 2024; Ghosh et al., 2024). Recent developments extend long-audio understanding and reasoning in audio-focused models (Ghosh et al., 2025; Goel et al., 2025; Ghosh et al., 2026; Shen et al., 2026), while multimodal systems incorporate audio alongside other modalities (Gemini Team et al., 2024; Xu et al., 2025a;b). AudioTrace evaluates trustworthy, evidence-grounded QA, examining these models' ability to correctly interpret supporting evidence in long recordings and reason from it to answer questions.

2.2 Audio Understanding Benchmarks

Audio understanding benchmarks have developed along two complementary directions: broadening capability coverage and extending audio duration. One line of work, primarily using short clips, evaluates capabilities ranging from captioning and question answering (Kim et al., 2019; Drossos et al., 2020; Lipping et al., 2022) to comprehension and reasoning across speech, sound, and music (Yang et al., 2024; Wang et al., 2025; Sakshi et al., 2025; Ma et al., 2025; Kumar et al., 2025).

Another line extends evaluation to long audio, covering long-audio QA and speech understanding (Ghosh et al., 2025; Guo et al., 2025; He et al., 2025; Yang et al., 2026b), as well as localization, comprehension, and reasoning in extended contexts (Ahia et al., 2025; Luo et al., 2026; Ye et al., 2026). Within this direction, TAG-Bench evaluates query-based temporal localization, while LAT-Bench also combines temporal alignment and semantic quality in dense captioning (Dai et al., 2026; Shao et al., 2026). LongAudio-RAG explores structured event retrieval for long-audio QA, with multi-hour QA and short-audio temporal detection evaluated separately (Hegde et al., 2026). AudioSpan evaluates perception, understanding, and reasoning over long recordings using both multiple-choice and open-ended questions (Huang et al., 2026b). MuLA-Bench expands open-ended long-audio QA across languages and domains, covering factual understanding, reasoning, temporal localization, and long-range retrieval (Yang et al., 2026c). Vox-Infinity annotates answer provenance to analyze spoken QA by the amount of historical context required, evaluating whether models retain and use information from earlier in the audio (Cheng et al., 2026).

However, evaluating QA and temporal grounding separately cannot establish whether a model reasons from the correct audio evidence to arrive at its answer. To address this limitation, AudioTrace assesses answer correctness, temporal evidence localization, and semantic evidence grounding

through a multi-level protocol. It examines whether models can both answer correctly and accurately identify the supporting audio segments and their factual content, exposing limitations in fine-grained understanding and reasoning that answer accuracy alone cannot reveal.

2.3 Evaluating Evidence Localization and Support

Several video QA benchmarks link answer evaluation to temporal and spatial evidence (Lei et al., 2020; Xiao et al., 2024; Zhou et al., 2024; Wang et al., 2026). TVQA+ and NExT-GQA count a prediction as jointly successful only when the answer is correct and temporal overlap meets a specified threshold, while VideoZeroBench extends verification to answer correctness and temporal and spatial grounding (Meng et al., 2026). EG-VQA is particularly relevant to our event-level evidence representation: it evaluates temporal segments paired with textual descriptions, using Evidence-Grounded F1 (EG-F1), a one-to-one matching metric that requires both semantic and temporal consistency (Huang et al., 2026a). Complementary protocols use clue-conditioned inputs or reference keyframes to examine evidence acquisition and integration (Chen et al., 2024; Ben-Ami et al., 2025). AudioTrace extends this evidence-centered perspective to long-audio QA, assessing whether models can answer reliably through an accurate understanding of the correct supporting evidence.

3 Dataset Construction and Evaluation

3.1 Dataset Overview

AudioTrace contains 120 long recordings in Chinese and English, totaling 99.8 hours, with 1,033 question pairs: 876 answerable and 157 unanswerable. Each pair provides an open-ended and a four-choice version and is counted once. We collect recordings from the Internet using keywords and retain those lasting 30–90 minutes. Appendix A.1 details the audio sources and composition.

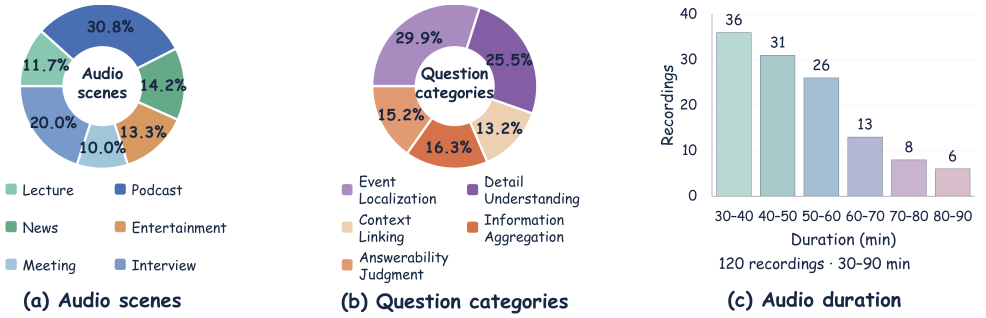


Figure 2: Composition of AudioTrace. (a) Audio scenes across 120 recordings. (b) Five categories across 1,033 question pairs. (c) Audio duration distribution with 10-minute bins; intervals include the lower endpoint and exclude the upper endpoint.

Audio distribution. Figure 2(a) shows six recording scenes. Lectures feature sustained explanations of a particular subject; podcasts develop topics through relatively free-form conversations; news focuses on reporting events and explaining their context. Entertainment covers media such as variety shows, featuring live interaction, expressive speech, and accompanying background sounds. Meetings involve discussion of an agenda through speaker turns and responses, while interviews explore personal experiences and viewpoints through questions and answers. These scenes cover monologues and dialogues, as well as continuous narration and alternating speaker turns, providing material for understanding spoken content, acoustic cues, and their contextual relations. Recordings are in Chinese and English and span 30–90 minutes, with the duration distribution shown in panel (c). Detailed recording statistics appear in Appendix A.1.

Question distribution. Figure 2(b) groups the 1,033 question pairs into five categories. *Event Localization* (309; 29.9%) asks where a specified utterance, topic, or acoustic event occurs. *Detail Understanding* (263; 25.5%) concerns the content or audible attributes of an individual event. *Context Linking* (136; 13.2%) resolves relations between events, such as references, question–response links, and temporal order. *Information Aggregation* (168; 16.3%) combines information across events to aggregate, compare, count, or track facts, and to summarize themes, viewpoints. *Answerability Judgment* (157; 15.2%) consists of questions for which the recording provides insufficient evidence, testing whether models recognize the information gap and abstain. This composition tests understanding of specific events and details as well as the ability to connect and integrate information across segments. Appendix A.2 provides three representative questions per category.

3.2 Annotation Pipeline and Quality Control

To turn these long recordings into question–answer pairs with explicit supporting evidence, we construct a pipeline comprising event annotation, question generation, and multi-stage review, as illustrated in Figure 3. Appendix A.3 details temporal alignment and event annotation.

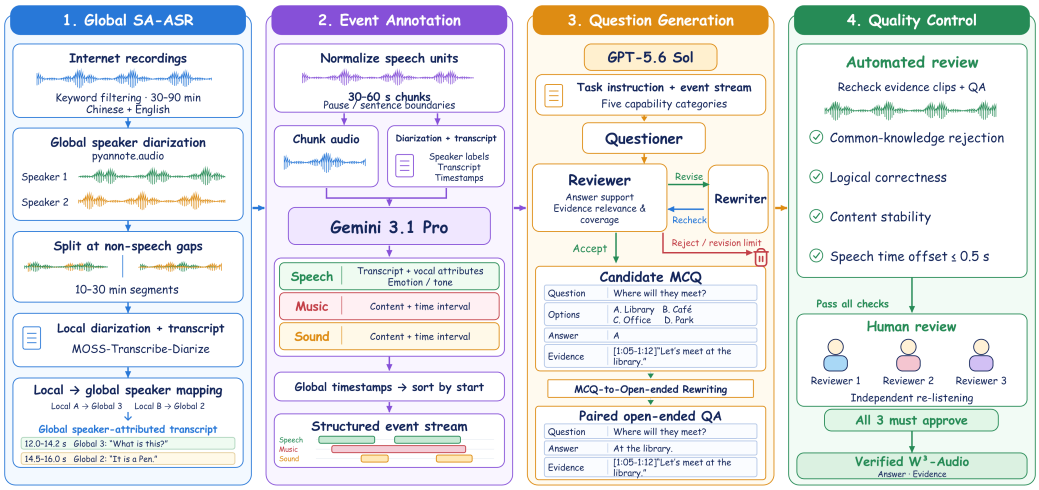


Figure 3: Overview of the AudioTrace construction pipeline. Long recordings are converted into a unified event stream to generate paired multiple-choice and open-ended questions, which are then screened through automated and human review.

Structured event annotation. We first use pyannote.audio to establish a global speaker diarization timeline and split the recording at non-speech gaps into 10–30 minute segments. MOSS-Transcribe-Diarize then produces local speaker labels and timestamped transcripts for each segment (Bredin, 2023; MOSI.AI, 2026). Local labels are mapped to global speakers by maximizing the harmonic mean of bidirectional temporal coverage ratios, maintaining speaker consistency across segments. We further partition the audio into 30–60 second chunks along pauses or sentence boundaries. Each chunk, together with its corresponding diarization and transcript, is provided to Gemini 3.1 Pro to add sound and music events and vocal attributes, such as speaker emotion, associated with each transcribed utterance. Finally, chunk-local event timestamps are mapped to the full-recording timeline, and events are sorted chronologically by their `start` field to form a unified stream.

Question generation. We use GPT-5.6 Sol for three roles: question generation, review, and revision. The questioner receives a capability-specific instruction and event context, and proposes a multiple-choice question with its answer and supporting events. A separate review stage checks whether the question is answerable from the available audio information, whether its answer is supported, and whether its evidence omits necessary events or includes irrelevant ones. Each candidate receives one of three decisions: accept, revise, or reject. Repairable candidates are returned to a rewriter with the review feedback and then reviewed again; rejected candidates and those still requiring revision after the configured limit are not accepted. Evidence must contribute directly to the answer: neighboring segments or distractors are not included merely to demonstrate their irrelevance. Each multiple-choice question is rewritten as a corresponding open-ended question.

Question verification. We combine automated checks with human review to control annotation quality. A common-knowledge answerability audit checks whether the answer can be determined from common knowledge alone. Event-level screening re-extracts the evidence audio: speech evidence must pass transcript and boundary consistency checks against an independent MOSS retranscription, followed by an utterance-completeness check. Sound and music evidence is independently described and checked for central-event agreement. After automated screening, three reviewers independently re-listen to the audio and inspect the question, answer, and evidence. A sample is retained only when all three approve it. Appendix A.4 specifies the automated filtering rules.

3.3 Task and Annotation Definition

Given a recording and a question, evidence-grounded QA requires an answer (`answer`) and an evidence set (`evidence`). The evidence set consists of audio intervals that directly support the answer.

Each interval has `start` and `end` fields specifying its boundaries in seconds from the beginning of the recording. An answer may be supported by one or more intervals. Each evidence entry also records the event type (`speech`, `music`, or `sound`) and its content. Speech entries additionally include a `speaker_attr` field describing paralinguistic cues such as tone and emotion.

For unanswerable questions, the reference answer indicates that the recording provides insufficient information, and the evidence set is empty. Appendix A.5 provides the field definitions and a complete question–answer annotation example.

3.4 Three-Level Evaluation

The three-level hierarchy separates answering, temporal localization, and evidence content matching by progressively increasing what a successful response must provide. Each level covers both multiple-choice and open-ended QA, yielding six independent inference settings. Models receive the recording or its model-generated event stream and the question, with options supplied only for MCQ. No reference answers or evidence are provided at any level. Let N be the number of evaluated questions and let $\mathbb{I}[\cdot]$ denote the indicator function.

Level-1: Answer correctness. Models return only an option or a direct answer. An MCQ answer is correct if its selected option matches the reference. For open-ended QA, following rubric-based LLM evaluation and evaluator-bias research (Zheng et al., 2023; Liu et al., 2023), two independent judges assign integer scores from 1 to 5 for correctness, completeness, and relevance. An answer is correct only when both judges assign 5 on all three dimensions. Each judge receives only the question, reference answer, and predicted answer. This correctness rule applies to answers at all three levels. Accuracy is

$$\mathrm{Acc}_{L1} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{y}_i = y_i]. \tag{1}$$

This level evaluates answer correctness only, without considering supporting evidence.

Level-2: Correct answers with temporal evidence. Models independently produce an answer and one or more supporting time intervals. We take the union of the predicted intervals and the union of the reference intervals, and compute their temporal intersection-over-union (tIoU). In addition to mean tIoU, we report

$$\mathrm{Acc}_{L2} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{y}_i = y_i]\,\mathbb{I}[\mathrm{tIoU}_i > 0.5]. \tag{2}$$

This level tests whether a correct answer is accompanied by accurately localized support.

Level-3: Correct answers with temporal and factual evidence. Models independently produce an answer, supporting intervals, and descriptions of the corresponding events. Following EG-VQA (Huang et al., 2026a), we use EG-F1 to jointly evaluate temporal alignment and semantic consistency. A predicted and reference event can match only if their event types agree, their pairwise tIoU is strictly greater than 0.5, and their content embedding cosine similarity is strictly greater than 0.5. We use Qwen3-Embedding-0.6B (Zhang et al., 2025) as the shared text encoder; speech inputs combine transcripts and speaker attributes, while music and sound use event descriptions. We find a maximum-cardinality one-to-one matching. With k matched pairs among m predicted and n reference events, EG-F1 $= 2k/(m+n)$. Alongside mean EG-F1, we report

$$\mathrm{Acc}_{L3} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{y}_i = y_i]\,\mathbb{I}[\mathrm{tIoU}_i > 0.5]\,\mathbb{I}[\text{EG-F1}_i > 0.5]. \tag{3}$$

This level requires the answer, temporal support, and event content to satisfy their respective criteria in the same response. The tIoU condition complements event-level matching by penalizing excessive duration outside the reference evidence.

All accuracy measures and continuous grounding means use the full set of 1,033 questions. For the 157 questions without reference evidence, an empty evidence prediction is treated as correct (tIoU=1 and EG-F1=1), while any predicted evidence is treated as a false positive (tIoU=0 and EG-F1=0). Thus, Level-2 and Level-3 joint accuracy requires both a correct answer and a valid evidence decision, including correctly abstaining when the recording contains no supporting evidence. Together, the three levels distinguish successful answering from successful answering with verifiable temporal and factual support; each level is scored using its own independently generated response. Reusing an L3 response to score L1 or L2 would instead measure answering or temporal localization under an L3 factual-grounding instruction, rather than under each level's intended task. Appendix B provides the answer-scoring rubric, metric definitions, matching procedure, and empty-evidence conventions.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluated systems. We evaluate AudioTrace under two complementary inference paradigms: end-to-end audio inference and caption-based cascaded inference. The end-to-end setting includes Gemini 3.1 Pro and Gemini 3.6 Flash (Google DeepMind, 2026a;b), Qwen3.5 Omni Plus (Qwen Team, 2026a), Qwen3.8 Omni Flash (Qwen Team, 2026b), and FireRedAudio (Shen et al., 2026); all receive complete recordings, without truncation or transcript substitution. The cascade setting additionally covers Qwen3 Omni (Xu et al., 2025b), Audio Flamingo (Ghosh et al., 2026), Kimi Audio (KimiTeam et al., 2025), MiMo Audio (Xiaomi LLM-Core Team, 2025), MiniCPM-o (Cui et al., 2026), MOSS-Audio (Yang et al., 2026a), and Step-Audio2 (Wu et al., 2025). Each audio captioner produces a timestamped event stream, which is passed to the same text reasoner, Claude Opus 4.8 (Anthropic, 2026). This setting accommodates models that cannot process full recordings directly and evaluates their caption ability under a shared reasoning backend. Both paradigms use the same question set, with independent MCQ and open-ended inference at each of the three levels in Section 3.4. Thinking is disabled for all benchmark inference models.

Answer judges. For open-ended answers, we use GPT-5.6 Sol (OpenAI, 2026) and DeepSeek-V4-Flash (DeepSeek-AI, 2026) as independent judges, both with thinking disabled, and apply the rubric rule and the correctness rule in Section 3.4. Appendix C validates agreement between the automated judges and human raters.

4.2 OVERALL PERFORMANCE

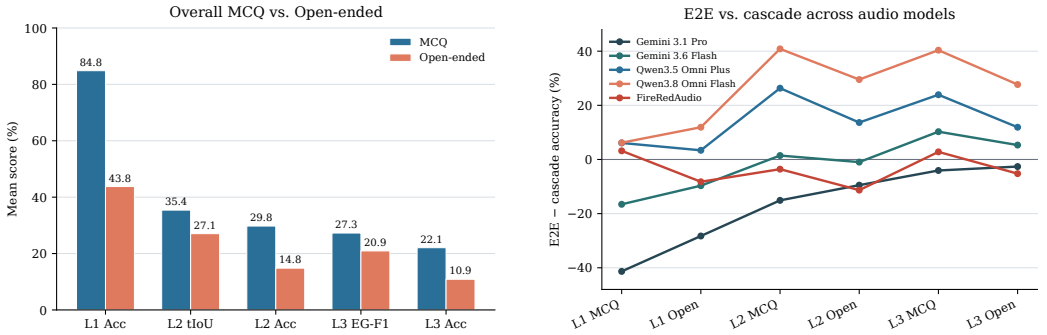
Table 2: Main results on AudioTrace. L1 reports answer accuracy; L2 reports mean tIoU and L2 joint Acc (blue); L3 reports mean EG-F1 and L3 joint Acc (red). Bold values indicate the best result in each column within each track. Acc is in percent; tIoU and EG-F1 are on $[0, 1]$.

Models	**Level-1**		**Level-2**				**Level-3**			
	MCQ	Open	MCQ		Open		MCQ		Open	
	Acc	Acc	tIoU	Acc	tIoU	Acc	EG-F1	Acc	EG-F1	Acc
End-to-end audio inference										
Gemini 3.1 Pro	50.5	23.2	0.233	21.7	0.154	9.3	0.209	19.3	0.139	9.3
Gemini 3.6 Flash	74.9	41.9	0.444	39.7	0.292	16.6	0.394	34.4	0.250	14.1
Qwen3.5 Omni Plus	95.8	51.9	0.485	49.5	0.351	25.4	0.435	39.1	0.315	19.5
Qwen3.8 Omni Flash	**97.0**	**62.1**	**0.679**	**72.0**	**0.563**	**45.5**	**0.645**	**60.8**	**0.555**	**37.9**
FireRedAudio	93.7	45.2	0.429	41.2	0.237	13.9	0.396	35.6	0.243	12.0
Caption-based cascaded inference										
Gemini 3.1 Pro	**91.9**	51.5	0.415	36.8	0.332	18.8	0.318	23.3	0.251	11.9
Gemini 3.6 Flash	91.5	51.6	0.416	38.2	0.332	17.5	0.330	24.1	0.256	8.8
Qwen3 Omni	79.8	35.8	0.183	10.8	0.152	5.7	0.095	6.8	0.072	3.8
Qwen3.5 Omni Plus	89.7	48.5	0.330	23.1	0.266	11.7	0.208	15.2	0.174	7.6
Qwen3.8 Omni Flash	90.8	50.1	0.395	31.2	0.310	16.0	0.287	20.4	0.233	10.3
Audio Flamingo	77.2	34.5	0.215	13.2	0.172	6.2	0.118	8.2	0.103	5.2
FireRedAudio	90.5	**53.4**	**0.482**	**44.8**	**0.378**	**25.3**	**0.434**	**32.8**	**0.354**	**17.2**
Kimi Audio	67.0	23.1	0.208	15.4	0.177	10.0	0.151	12.9	0.123	9.5
MiMo Audio	87.3	42.3	0.281	20.0	0.231	8.0	0.198	13.0	0.158	4.9
MiniCPM-o	87.5	43.7	0.276	16.4	0.216	6.7	0.145	9.7	0.117	4.3
MOSS-Audio	89.4	45.2	0.303	19.3	0.237	9.5	0.169	12.3	0.133	6.0
Step-Audio2	87.9	39.9	0.245	12.5	0.199	6.0	0.106	7.4	0.081	2.6

Table 2 reports the overall performance under the two paradigms and the six independent response settings induced by the three evaluation levels and two question formats.

Finding 1: Answer correctness alone does not show that a model retrieved the relevant audio evidence or reasoned from it reliably. Averaged across the 17 model–paradigm configurations in Table 2, MCQ accuracy falls from 84.8% at L1 to 22.1% at L3, while open-ended accuracy drops from 43.8% to 10.9%. Even the best-performing model, Qwen3.8 Omni Flash (E2E), drops from 97.0% to 60.8% on MCQ and from 62.1% to 37.9% on open-ended QA. These results reveal a substantial gap between answer correctness and evidence-supported correctness: high QA accuracy does not establish that a model can accurately locate the relevant audio segments and understand the facts supporting its answer. Evaluating trustworthy long-audio QA must therefore go beyond the final answer to verify temporal evidence localization and semantic evidence grounding, exposing limitations in fine-grained understanding and reasoning that answer scores alone fail to reveal.

Finding 2: Limited temporal evidence localization and semantic evidence grounding remain obstacles to trustworthy long-audio reasoning. Across the 17 model–paradigm configurations, mean L2 tIoU is only 0.354 for MCQ and 0.271 for open-ended QA, while mean L3 EG-F1 is 0.273 and 0.209, respectively. Even Qwen3.8 Omni Flash (E2E), the strongest model on all four measures, reaches tIoU scores of 0.679 and 0.563 and EG-F1 scores of 0.645 and 0.555. The low averages of these metrics expose weaknesses in the submitted evidence. The tIoU results indicate imprecise temporal localization, while EG-F1 further reveals difficulty recovering supporting events that agree with the reference in both time and factual content. Reliable long-audio reasoning therefore remains challenging at the level of identifying and accurately representing its supporting evidence.



(a) Overall MCQ and open-ended performance

(b) Model-specific end-to-end versus cascade accuracy

Figure 4: Overall performance and model-specific inference-paradigm differences. Panel (a) reports mean MCQ and open-ended performance across the main audio experiments. Panel (b) shows the end-to-end accuracy minus the cascade accuracy (%) for each of five audio models across six evaluation settings.

Finding 3: Answer options may assist both answering and evidence search, yet fine-grained evidence grounding remains challenging. As shown in Fig. 4(a), MCQ outperforms open-ended QA not only in mean answer accuracy but also in mean tIoU, EG-F1, and the corresponding joint accuracies. This advantage across metrics suggests that candidate options may both provide answer cues and narrow the search space for supporting evidence, guiding models toward temporal segments and factual content consistent with a candidate answer. MCQ performance may therefore benefit partly from option-guided evidence search, whereas open-ended QA places greater demands on independently recovering and articulating audio facts. Nevertheless, Fig. 4(a) also shows that MCQ joint accuracy drops substantially as evidence requirements tighten, despite the availability of options. Option cues do not eliminate the difficulty of accurately localizing and interpreting supporting evidence, further demonstrating that high multiple-choice answer accuracy alone is insufficient to establish reliable long-audio understanding.

Finding 4: The relative performance of end-to-end and cascaded QA depends on the audio model and evaluation setting. Figure 4(b) compares five audio models used directly for QA with cascades that pass each model's audio descriptions to Claude Opus 4.8. Even at the same evaluation level, the stronger configuration differs by model: on L1 MCQ, cascading Gemini 3.1 Pro yields 91.9% accuracy versus 50.5% end-to-end, whereas Qwen3.8 Omni Flash yields 97.0% end-to-end versus 90.8% cascaded. The contrast persists when evidence is required. On L3 open-ended QA, FireRedAudio reaches 17.2% joint accuracy in the cascade versus 12.0% end-to-end, while Qwen3.8 Omni Flash reaches 37.9% end-to-end versus 10.3% cascaded. Because cascading changes both the

audio representation and the answering model, these comparisons do not isolate the cause of each difference. They show that neither inference paradigm is uniformly preferable across audio models and evaluation settings, making both configurations relevant to assessing long-audio QA systems.

Further analyses of these findings appear in Appendix D.

4.3 Discussion: Shortcuts in Automated QA Construction and Evaluation

Automatically generated MCQs create a tension between evaluation validity and data yield: answer options may reveal an answer without audio, but rejecting every potentially vulnerable item can discard questions with valid audio evidence. We discuss what these shortcuts mean for evidence-grounded evaluation and for the efficiency of benchmark construction.

Observation 1: Option shortcuts can aid answering and evidence search, but grounding remains difficult. Without audio, Qwen3.5-27B reaches 65.8% L1 MCQ accuracy versus 1.0% open-ended accuracy (Table 3), revealing substantial answer-level vulnerability to supplied options. In the audio experiments, MCQ also exceeds open-ended QA on mean L2 tIoU (0.354 versus 0.271) and L3 EG-F1 (0.273 versus 0.209; Table 2). One possible explanation is that the options narrow the search for supporting events. Yet option cues do not resolve grounding: the question-only control has 0.0% joint accuracy at both L2 and L3, while audio-system mean MCQ accuracy is 84.8% at L1 but only 22.1% under L3's joint requirements. Together, the results show that correct option selection is insufficient and that temporal and factual grounding remain challenging even when options are available.

Table 3: Question-only Qwen3.5-27B on 876 answerable questions (accuracy, %). MCQ receives the question and options; open-ended QA receives only the question. Under the same evaluation protocol, the model could not fabricate evidence to pass L2/L3 and scored zero at both levels.

Format	L1 answer	L2 joint	L3 joint
MCQ	65.8	0.0	0.0
Open-ended	1.0	0.0	0.0

Observation 2: Exhaustive shortcut filtering sharply reduces usable data. Our automatic evidence screening retains 1,050 of 2,431 generated candidates (43.2%; Table 4). If every answerable MCQ solved by the question-only model were then rejected after construction, 576 of the 876 final answerable pairs (65.8%) would be removed, leaving 300. The two rates describe different stages and cannot be multiplied. Replacing the rejected, evidence-verified questions to reach a fixed dataset size would require further generation and review. This is a retrospective filtering calculation with one model, not a measured regeneration cost; it nevertheless exposes the yield penalty of treating any text-only success as grounds for rejection.

Table 4: Automatic candidate retention. AudioTrace rates use 2,431 initial candidates. AudioSpan's separately reported overall yield is shown for context (Huang et al., 2026b).

Stage	Retained	Rate
AudioTrace		
Initial candidates	2,431	100%
Answerability audit	2,400	98.7%
Speech consistency	1,783	73.3%
Speech completeness	1,392	57.3%
Sound/music authenticity	1,050	43.2%
AudioSpan		
All automatic gates	—	∼8%

We therefore prioritize open-ended data generation and evaluation, which avoid option-based cues. When MCQs are necessary, we recommend stringent option screening or joint checks of the answer, temporal location, and factual evidence. Further shortcut analyses appear in Appendix E.

5 Conclusion

We introduced AudioTrace, a benchmark for trustworthy long-audio QA that distinguishes answer correctness from evidence-supported correctness. Its automated construction pipeline links questions to fine-grained event evidence, while a three-level protocol evaluates both multiple-choice and open-ended QA. Across end-to-end and cascaded inference, our experiments reveal a substantial gap between answering correctly and accurately localizing and interpreting supporting evidence. Temporal evidence localization and semantic evidence grounding remain challenging even for the strongest models. These findings call for evaluation beyond final answers and greater attention to evidence retrieval and fine-grained understanding in the development of trustworthy long-audio QA systems.

AI USE STATEMENT

We used AI tools to assist with dataset construction and answer evaluation, code implementation and data analysis, literature and reference searches, and manuscript editing. The construction and evaluation procedures are described in the main text and appendices. The authors are responsible for verifying AI-assisted outputs and take responsibility for the final manuscript and research artifacts.

ETHICS STATEMENT

Licensing and release. AudioTrace is intended for non-commercial research and evaluation. The original benchmark question data, including paired multiple-choice and open-ended questions, are publicly available in the anonymous repository. Our original benchmark materials, including the question data, scoring rubrics, structured speech, sound, and music event annotations, and temporal and content evidence annotations, are covered by the CC BY-NC-SA 4.0 license. The data-construction and evaluation code, inference prompts, run configurations, and inference outputs needed to reproduce the paper's scores are also publicly available. The original recordings are not redistributed: they remain subject to their respective owners' rights and terms, and our release does not grant redistribution rights over them. The repository provides the original download links for all recordings in the benchmark, allowing the audio to be retrieved through its original channels.

Ethics and responsible use. All recordings are drawn from publicly accessible media. Speakers are referred to by their roles rather than their identities, and non-public personal information is not used as an answer key. AudioTrace is an evaluation resource for long-audio understanding and evidence-grounded question answering, not for training speaker-identification or surveillance systems. Its temporal and content evidence annotations support the verification of model answers against the recordings. Human reviewers are compensated at or above local market rates and review only publicly accessible media and the associated benchmark annotations.

REPRODUCIBILITY STATEMENT

The construction pipeline, including structured audio annotation, evidence-linked question generation, and iterative review and revision, is described in Section 3 and Appendix A. Appendix B specifies the three-level evaluation protocol, answer-scoring rubric, and tIoU and EG-F1 calculations; Appendix C provides the answer-judge prompt and human-calibration protocol. Evaluated models and inference paradigms are described in Section 4.1, model and API versions are listed in Appendix F, and threshold sensitivity is examined in Appendix D.3. The public repository contains the original benchmark question data, construction and evaluation code, six inference prompts, run configurations, inference outputs needed to reproduce the reported scores, and original download links for the benchmark recordings. The recordings themselves are not redistributed.

REFERENCES

Orevaoghene Ahia, Martijn Bartelds, Kabir Ahuja, Hila Gonen, Valentin Hofmann, Siddhant Arora, Shuyue Stella Li, Vishal Puttagunta, Mofetoluwa Adeyemi, Charishma Buchireddy, Ben Walls, Noah Bennett, Shinji Watanabe, Noah A. Smith, Yulia Tsvetkov, and Sachin Kumar. BLAB: Brutally Long Audio Bench, 2025. URL https://arxiv.org/abs/2505.03054.

Anthropic. Introducing Claude Opus 4.8, 2026. URL https://www.anthropic.com/news/claude-opus-4-8.

Dan Ben-Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. HERBench: A Benchmark for Multi-Evidence Integration in Video Question Answering, 2025. URL https://arxiv.org/abs/2512.14870.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Interspeech 2023*, pp. 1983–1987, 2023. doi: 10.21437/Interspeech.2023-105. URL https://www.isca-archive.org/interspeech_2023/bredin23_interspeech.html.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. CG-Bench: Clue-grounded Question Answering Benchmark for Long Video Understanding, 2024. URL https://arxiv.org/abs/2412.12075.

Xize Cheng, Wenxu Jia, Chenyuhao Wen, Dongjie Fu, Zehan Wang, Xinyu Zhang, and Tao Jin. Vox-Infinity: Benchmarking the Limits of Long-Context Spoken Language Models, 2026. URL https://arxiv.org/abs/2609.22452.

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, 2023. URL https://arxiv.org/abs/2311.07919.

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, 2024. URL https://arxiv.org/abs/2407.10759.

Junbo Cui, Bokai Xu, Chongyi Wang, et al. MiniCPM-o 4.5: Towards Real-Time Full-Duplex Omni-Modal Interaction, 2026. URL https://arxiv.org/abs/2604.27393.

Yuhang Dai, Xin Shu, Zengxi Li, Lei Xie, Xiangang Li, and Jianwei Yu. TAG-Bench: Benchmarking Temporal Audio Grounding in Large Audio Language Models, 2026. URL https://arxiv.org/abs/2609.01542.

DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An Audio Captioning Dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9052990. URL https://doi.org/10.1109/ICASSP40776.2020.9052990.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

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, 2024. URL https://arxiv.org/abs/2406.11768.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19358–19405. PMLR, 2025. URL https://proceedings.mlr.press/v267/ghosh25b.html.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. Audio Flamingo Next: Next-Generation Open Audio-Language Models for Speech, Sound, and Music, 2026. URL https://arxiv.org/abs/2604.10905.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models, 2025. URL https://arxiv.org/abs/2507.08128.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. Listen, Think, and Understand, 2023. URL https://arxiv.org/abs/2305.10790.

Google DeepMind. Gemini 3.1 Pro Model Card, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Google DeepMind. Gemini 3.6 Flash Model Card, 2026b. URL https://deepmind.google/models/model-cards/gemini-3-6-flash/.

Shoutao Guo, Shaolei Zhang, Qingkai Fang, Zhengrui Ma, Min Zhang, and Yang Feng. FastLongSpeech: Enhancing Large Speech-Language Models for Efficient Long-Speech Processing, 2025. URL https://arxiv.org/abs/2507.14815.

Peize He, Zichen Wen, Yubo Wang, Yuxuan Wang, Xiaoqian Liu, Jiajie Huang, Zehui Lei, Zhuangcheng Gu, Xiangqi Jin, Jiabing Yang, Kai Li, Zhifei Liu, Weijia Li, Cunxiang Wang, Conghui He, and Linfeng Zhang. AudioMarathon: A Comprehensive Benchmark for Long-Context Audio Understanding and Efficiency in Audio LLMs, 2025. URL https://arxiv.org/abs/2510.07293.

Kartik Hegde, Arvind Krishna Sridhar, Naveen Vakada, Yinyi Guo, and Erik Visser. Event-Grounded Question Answering over Long Audio via Structured Retrieval, 2026. URL https://arxiv.org/abs/2602.14612.

Linpeng Huang, Weixing Chen, Zexin Chen, Yang Liu, and Liang Lin. EG-VQA: Benchmarking Verifiable Video Question Answering with Grounded Temporal Evidence, 2026a. URL https://arxiv.org/abs/2606.24797.

Wen Huang, Yunfei Chu, Meng Gao, Haolin He, and Jin Xu. LongAudioSpan: Spanning the Duration and Depth of Audio Comprehension, 2026b. URL https://arxiv.org/abs/2608.26431.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in The Wild. In *Proceedings of NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pp. 119–132. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report, 2025. URL https://arxiv.org/abs/2504.18425.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25125–25148. PMLR, 2024. URL https://proceedings.mlr.press/v235/kong24a.html.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. MMAU-Pro: A Challenging and Comprehensive Benchmark for Holistic Evaluation of Audio General Intelligence, 2025. URL https://arxiv.org/abs/2508.13992.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8211–8225, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.730. URL https://aclanthology.org/2020.acl-main.730/.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-AQA: A Crowdsourced Dataset for Audio Question Answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 1140–1144. IEEE, 2022. doi: 10.23919/EUSIPCO55093.2022.9909680. URL https://doi.org/10.23919/EUSIPCO55093.2022.9909680.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

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, 2026. URL https://arxiv.org/abs/2601.04876.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix, 2025. URL https://arxiv.org/abs/2505.13032.

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. VideoZeroBench: Probing the Limits of Video MLLMs with Spatio-Temporal Evidence Verification, 2026. URL https://arxiv.org/abs/2604.01569.

MOSI.AI. MOSS Transcribe Diarize Technical Report, 2026. URL https://arxiv.org/abs/2601.01554.

OpenAI. GPT-5.6 System Card, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6.

Qwen Team. Qwen3.5-Omni Technical Report, 2026a. URL https://arxiv.org/abs/2604.15804.

Qwen Team. Qwen3.8-Omni: Towards Native Omni-Modal Agents, 2026b. URL https://arxiv.org/abs/2609.25611.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/29528.

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, 2026. URL https://arxiv.org/abs/2604.22245.

Feiyu Shen, Fenglong Xie, Junjie Li, Kun Xie, Lei Xie, Xu Tang, Xuelong Geng, Yan Jia, Yao Hu, Yichen Han, Yichen Wu, Ziqi Dai, Junjie Chen, Kai Huang, Manzhen Wei, and Yixuan Li. FireRedAudio: A General-Purpose Audio Language Model with Decoupled Continuous Representations for Understanding and Generation, 2026. URL https://arxiv.org/abs/2608.24168.

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, 2023. URL https://arxiv.org/abs/2310.13289.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.218. URL https://aclanthology.org/2025.naacl-long.218/.

Shijie Wang, Honglu Zhou, Ziyang Wang, Ran Xu, Caiming Xiong, Silvio Savarese, Chen Sun, and Juan Carlos Niebles. Evidence-Backed Video Question Answering, 2026. URL https://arxiv.org/abs/2607.11862.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, Yuxin Zhang, Zhao You, Brian Li, Changyi Wan, Hanpeng Hu, Jiangjie Zhen, Siyu Chen, Song Yuan, Xuelin Zhang, Yimin Jiang, Yu Zhou, Yuxiang Yang, Bingxin Li, Buyun Ma, Changhe Song, Dongqing Pang, Guoqiang Hu, Haiyang Sun, Kang An, Na Wang, Shuli Gao, Wei Ji, Wen Li, Wen Sun, Xuan Wen, Yong Ren, Yuankai

Ma, Yufan Lu, Bin Wang, Bo Li, Changxin Miao, Che Liu, Chen Xu, Dapeng Shi, Dingyuan Hu, Donghang Wu, Enle Liu, Guanzhe Huang, Gulin Yan, Han Zhang, Hao Nie, Haonan Jia, Hongyu Zhou, Jianjian Sun, Jiaoren Wu, Jie Wu, Jie Yang, Jin Yang, Junzhe Lin, Kaixiang Li, Lei Yang, Liying Shi, Li Zhou, Longlong Gu, Ming Li, Mingliang Li, Mingxiao Li, Nan Wu, Qi Han, Qinyuan Tan, Shaoliang Pang, Shengjie Fan, Siqi Liu, Tiancheng Cao, Wanying Lu, Wenqing He, Wuxun Xie, Xu Zhao, Xueqi Li, Yanbo Yu, Yang Yang, Yi Liu, Yifan Lu, Yilei Wang, Yuanhao Ding, Yuanwei Liang, Yuanwei Lu, Yuchu Luo, Yuhe Yin, Yumeng Zhan, Yuxiang Zhang, Zidong Yang, Zixin Zhang, Binxing Jiao, Daxin Jiang, Heung-Yeung Shum, Jiansheng Chen, Jing Li, Xiangyu Zhang, and Yibo Zhu. Step-audio 2 technical report, 2025. URL https://arxiv.org/abs/2507.16632.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xiao_Can_I_Trust_Your_Answer_Visually_Grounded_Video_Question_Answering_CVPR_2024_paper.html.

Xiaomi LLM-Core Team. Mimo-audio: Audio language models are few-shot learners, 2025. URL https://arxiv.org/abs/2512.23808.

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, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni Technical Report, 2025b. URL https://arxiv.org/abs/2509.17765.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, Zhengyuan Lin, Ziqi Chen, Zhaoye Fei, Chenghao Liu, Jun Zhan, Kang Yu, Kexin Huang, Mingshu Chen, Qinyuan Cheng, Ruixiao Li, Shimin Li, Songlin Wang, Yang Gao, Yiyang Zhang, and Xipeng Qiu. Moss-audio technical report, 2026a. URL https://arxiv.org/abs/2606.01802.

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, 2026b. URL https://arxiv.org/abs/2601.13539.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.109. URL https://aclanthology.org/2024.acl-long.109/.

Zeyu Yang, Xinyu Zhang, Zibo Bi, Pei Zhang, Xize Cheng, Jin Xu, Baosong Yang, and Satoshi Nakamura. MuLA-Bench: A Multilingual Long-Form Audio Understanding Benchmark via Multi-Tier Auditing, 2026c. URL https://arxiv.org/abs/2609.23416.

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, 2026. URL https://arxiv.org/abs/2605.27976.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica.

Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Sheng Zhou, Junbin Xiao, Xun Yang, Peipei Song, Dan Guo, Angela Yao, Meng Wang, and Tat-Seng Chua. Scene-Text Grounding for Text-Based Video Question Answering, 2024. URL https://arxiv.org/abs/2409.14319.

A Dataset Construction Details

This appendix documents how long recordings become question–answer pairs with event-level support: source composition and examples, temporal alignment, evidence review, and the released annotation schema. Automated checks screen candidate annotations; final human review verifies both evidence content and temporal boundaries.

A.1 Audio Sources and Composition

AudioTrace contains 120 recordings and 1,033 question pairs. Recordings are collected from the Internet using keywords and retained only if they last 30–90 minutes. Table 5 reports their scene and duration composition, and Table 6 reports the language distribution.

Table 5: Genre and duration composition of AudioTrace. SD is the sample standard deviation.

Genre	Audios	Share	Hours	Mean (SD) (min)	Range (min)
Lecture	14	11.7%	12.56	53.81 (19.55)	32.32–89.77
Podcast	37	30.8%	30.43	49.34 (10.59)	30.38–73.32
News	17	14.2%	10.86	38.35 (5.31)	30.12–48.19
Entertainment	16	13.3%	15.54	58.28 (10.32)	40.39–81.65
Meeting	12	10.0%	9.53	47.65 (13.70)	31.66–74.31
Interview	24	20.0%	20.83	52.08 (20.73)	30.00–89.96
Total	120	100.0%	99.75	49.88 (15.00)	30.00–89.96

Table 6: Audio language distribution in AudioTrace. Count share and duration share are computed over recordings and total audio duration, respectively.

Language	Audios	Count share	Hours	Duration share	Mean (min)
Chinese	63	52.5%	52.53	52.66%	50.03
English	57	47.5%	47.22	47.34%	49.71
Total	120	100.0%	99.75	100.00%	49.88

A.2 Representative Questions

Table 7 presents three questions from each category in AudioTrace. All examples are selected from original English questions; the English manuscript preserves the original wording, and the Chinese manuscript provides translations. The Answerability Judgment examples lack sufficient support in their source recordings and have abstention references.

A.3 Temporal Alignment and Event Annotation

Transcription and speaker alignment. We apply pyannote.audio to the full recording to obtain a global speaker diarization timeline. The recording is split at non-speech gaps into 10–30 minute segments, each processed by MOSS-Transcribe-Diarize to obtain local speaker labels and timestamped transcripts. After converting local timestamps to the full-recording timeline, we map local labels to global speakers within each transcription segment using the harmonic mean of bidirectional temporal coverage ratios.

Let C denote the current transcription segment, $L_{C,\ell}$ the union of all speech intervals assigned to local speaker ℓ in that segment, and $G_{C,g}$ the union of global speaker g's speech intervals clipped to C. With $|\cdot|$ denoting total interval duration, the two coverage ratios are

$$p_{\ell g} = \frac{|L_{C,\ell} \cap G_{C,g}|}{|L_{C,\ell}|}, \qquad r_{\ell g} = \frac{|L_{C,\ell} \cap G_{C,g}|}{|G_{C,g}|}. \tag{4}$$

The first measures how much of the local speaker's speech is covered by the global speaker; the second measures how much of the global speaker's speech within the current segment is covered by the local speaker. We select the global speaker with the highest harmonic-mean score:

$$H_{\ell g} = \frac{2 p_{\ell g} r_{\ell g}}{p_{\ell g} + r_{\ell g}} = \frac{2|L_{C,\ell} \cap G_{C,g}|}{|L_{C,\ell}| + |G_{C,g}|}, \qquad \hat{g}_C(\ell) = \arg\max_{g \in \mathcal{G}_C} H_{\ell g}, \tag{5}$$

where $\mathcal{G}_C$ contains the global speakers present in the current segment; a ratio or score is set to zero when its denominator is zero. Global-speaker duration is computed only within the current

Type	Representative Question Examples
Event Localization	– At approximately what time does a person say, "Come on"? – At approximately what time does the female audience member ask about appropriate ways to use AI in education? – At approximately what time does the interviewee say, "He is very unshy"?
Detail Understanding	– When the audience member says, "my last, next and last question," what does her delivery most clearly suggest? – Around 1688 seconds, what does the man mean when he says, "This is it"? – Around 8 minutes 31 seconds, how is the Wise app described?
Context Linking	– Near the end of the event, what is the temporal relationship between the audience's applause and Sam saying, "Thank you all very much"? – Around 10:08, when Kimber is asked, "You good?", how does she respond to the question? – In the lecturer's comparison between engineering and microeconomics, what does "something to push a ping pong ball off tables" refer to?
Information Aggregation	– Please state the instructions given during the initial cargo mission briefing. – Across the interview moments around 15:40 and 48:00, how many separate times does the woman suggest having a cup of coffee? – What best summarizes M's view of taking risks for her community and how she chose to help?
Answerability Judgment	– How much money did the researchers spend to train the meta-learning system mentioned in the talk? – What cash prize was awarded to the winner of the 2007 robot-building contest mentioned in the lecture? – According to the audio, exactly how much money did Amsterdam allocate to address high housing prices under its version of the donut plan?

Table 7: Representative questions from the five question categories.

transcription segment, excluding speech elsewhere in the recording. All speech units with the same local label in a segment share this mapping, and local labels are assigned independently.

Ties are resolved by exact overlap duration, a collar-weighted overlap score, and speaker identifier, in that order. For each pair of speech intervals, the collar score adds to the exact overlap 0.2 times the additional overlap obtained by expanding both ends of the local interval; these contributions are summed. The expansion is configurable and defaults to 1.5 seconds per end. Whenever the segment contains candidate speakers, the implementation selects the highest-ranked candidate without a minimum-score threshold.

Speech normalization and unified event stream construction. We normalize the transcribed speech units and partition the audio into 30–60 second chunks, placing boundaries at pauses or sentence endings. Each chunk is provided to Gemini 3.1 Pro together with its global speaker labels and timestamped transcripts. The model adds sound and music events and paralinguistic descriptions of each utterance, such as emotion and tone. We then convert event boundaries to global timestamps by adding the chunk's start time in the full recording to the local start and end times. Finally, we collect the speech, music, and sound events across chunks and sort them chronologically by their `start` field, forming a unified event stream for subsequent question generation.

A.4 Review Criteria

Review comprises generation-time question–answer review and common-knowledge auditing, subsequent event-level evidence screening, and final human verification.

Generation-time review and common-knowledge auditing. GPT-5.6 Sol reviews candidate questions, answers, and evidence against the event context; repairable candidates are revised and reviewed again. An independent common-knowledge audit receives only the question and answer options, without audio, transcripts, reference answers, or evidence. It determines whether common knowledge, established domain knowledge, or information stated in the question suffices to iden-

tify a unique answer. New questions are audited individually, and only explicit negative verdicts on such answerability enter the candidate pool. Rejected, uncertain, or failed candidates trigger regeneration, with at most 30 attempts per recording. Guessing from option style does not count as common-knowledge answerability. The historical question pool additionally undergoes a Qwen3.5-27B check for answers explicitly disclosed by the question stem.

Speech transcript and boundary consistency. Each annotated speech interval is extracted and independently retranscribed by MOSS-Transcribe-Diarize. The onset of the first detected speech segment and the offset of the last are compared with the corresponding annotated boundaries; both absolute deviations must be at most 0.5 seconds. The original and retranscribed text are normalized using Unicode NFKC and case folding, retaining only alphanumeric characters. Their character-sequence similarity, computed with Python's `SequenceMatcher.ratio`, must be at least 0.90. This is not WER and does not verify speaker identity across segments. Missing boundary timestamps or failure of either condition prevents acceptance.

Speech-segment completeness. GPT-5.6 Sol examines the original evidence text, with the MOSS retranscription as diagnostic context, for boundaries that cut through a sentence, clause, phrase, or required word sequence. The only verdicts are `complete` and `fragmented`; clearly truncated segments are rejected. Complete short responses such as "OK" or "yes" are acceptable, and informal phrasing or dependence on preceding context is not itself grounds for rejection. This step does not reassess transcript similarity or boundary thresholds.

Music and sound authenticity. Each music or sound evidence interval is extracted separately and independently described by Qwen3-Omni-30B-A3B-Captioner. GPT-5.6 Sol then compares this description with the original annotation to determine whether they refer to a sufficiently similar central event. The only verdicts are `present` and `reject`. Wording and fine-grained attributes may differ, but clearly different, contradictory, or entirely unsupported central events are rejected. Timestamps, event order, clip duration, and temporal coverage are excluded from this decision; no timing-discrepancy threshold is applied to music or sound evidence.

All evidence items associated with a question must pass their type-specific checks before human review; missing, failed, or invalid audit outputs do not count as passes. Unanswerable questions with empty evidence lists are exempt from event-level checks but still undergo question–answer review and human verification. Question naturalness is not a separate automated gate, and the three-stage evidence screening used for this question pool does not include an additional full-context audit. For exhaustive questions, evidence lists only events directly supporting the answer, rather than non-matching intervals. Review feedback and decisions are retained for traceability.

Human verification of content and boundaries. Three reviewers independently re-listen to the audio and inspect each question, answer, and evidence event. A sample passes only if all three approve the question and answer and verify that every event's factual description agrees with the audio and supports the answer, and that its start and end times correctly delimit the support. This requirement applies to speech, sound, and music, including sound and music boundaries not checked by the automated screening.

A.5 Annotation Fields and a Worked Example

`question` is the question, `answer` the reference answer, and `evidence` a list of audio intervals directly supporting the answer. Each evidence entry records its boundaries in seconds from the recording start using `start` and `end`; `stream` identifies speech, music, or sound, and `text` records the content. Speech events additionally use `speaker_attr` for paralinguistic cues such as tone and emotion. `evidence_id` identifies each evidence entry. `task_id` identifies the question's subtask, and `is_answerable` indicates answerability. Unanswerable questions have abstention answers and empty `evidence` lists. Multiple-choice versions additionally provide four options and the correct-option label.

Table 8 shows a real open-ended question and all its question–answer content fields: `additional_audio_000017_r001_0002` from `audio_000017`, in Detail Understanding. We use `evidence[0]` to denote the first entry in the evidence list and make the nesting explicit. The paper uses `start` and `end` for the source fields `start_sec` and `end_sec`, preserving their values. Internal event and speaker identifiers and file paths are omitted as provenance metadata. The original question, answer, and evidence content are in English; the English manuscript preserves the original text, and the Chinese manuscript provides translations.

Field	Example Value
question	Around 68 minutes into the lecture, which two industries does the lecturer give as examples?
answer	Oil and airlines
task_id	T2_a
is_answerable	true
evidence[0].evidence_id	evidence1
evidence[0].stream	speech
evidence[0].start	4081.927
evidence[0].end	4084.776
evidence[0].text	Oil is probably the classic example, airlines is a good example, okay?
evidence[0].speaker_attr	Middle-aged male speaker affirming with a reassuring and slightly conversational tone in British accent.

Table 8: A T2 detail-understanding question with its supporting speech interval and speaker attributes.

Table 9: Integer rubric for open-ended answers. Each dimension is scored independently.

Score	Correctness	Completeness	Relevance
5	Correct answer; no substantive factual errors.	All necessary information correctly covered.	Direct answer; all content relevant.
4	Core answer correct; localized error in secondary or extra content.	Main requirements covered; a secondary necessary component missing.	Direct answer with a minor tangent.
3	Some key content correct; important factual errors remain.	Some requirements covered; an important component missing.	Addresses the question with substantial irrelevant material.
2	Isolated correct details; core answer wrong.	Only isolated necessary details covered.	Mostly off-topic; weakly addresses the question.
1	No correct answer, wholly wrong or contradictory, or empty.	No necessary information correctly covered, or empty.	Unrelated, no meaningful answer attempt, or empty.

B Evaluation Protocols and Metrics

The protocol separates three questions: is the answer correct, does the predicted time span recover the supporting audio, and do the predicted events match its factual content? This appendix defines the corresponding answer, tIoU, and EG-F1 scores and their joint acceptance rules. MCQ and open-ended QA are generated independently at each of L1–L3; each level is scored using only its own response. Answer-judge prompts and human validation are provided in Appendix C.

B.1 Open-Ended Answer Evaluation

Judge inputs and scope. The answer judge receives only the question, the reference answer, and the predicted answer. It receives neither reference nor predicted evidence, audio, answer options, or the evaluated model's identity. Thus, an answer's score is independent of the quality of its accompanying evidence. The same rubric applies to L1, L2, and L3. The judge uses only the supplied text, accepts semantically equivalent paraphrases and synonyms, and does not reward verbosity or explanations.

Three-dimensional rubric. Correctness measures the factual accuracy of the expressed answer; completeness measures correct coverage of information required by the question; relevance measures how directly the response addresses the question, independently of factual correctness. Each dimension receives an integer score from 1 to 5 (Table 9). The judge first identifies necessary answer components from the question and reference, then records whether each is correctly covered and provides a brief justification. Optional reference details are not treated as requirements. For example, if a question asks only for a count and the reference lists three examples, answering "3" can receive 5/5/5 without listing those examples. An incorrect additional example affects correctness but need not reduce completeness of the requested count or topical relevance.

Scores and answer correctness. Two independent judges score each answer. Let $s_{ijd}^{(\ell)} \in \{1,2,3,4,5\}$ denote the score from judge $j \in \{1,2\}$ for question i, level ℓ, and dimension $d \in \{c,u,r\}$ (correctness, completeness, and relevance). We average across judges within each dimension:

$$\bar{s}_{id}^{(\ell)} = \frac{s_{i1d}^{(\ell)} + s_{i2d}^{(\ell)}}{2}. \tag{6}$$

The continuous answer score is the unweighted mean of these three dimension-wise means,

$$S_i^{(\ell)} = \frac{\bar{s}_{ic}^{(\ell)} + \bar{s}_{iu}^{(\ell)} + \bar{s}_{ir}^{(\ell)}}{3}. \tag{7}$$

The answer-correctness indicator requires both judges to assign the maximum score on every dimension:

$$A_i^{(\ell)} = \prod_{j=1}^{2} \prod_{d\in\{c,u,r\}} \mathbb{I}\left[s_{ijd}^{(\ell)} = 5\right]. \tag{8}$$

Thus, all six scores must equal 5; the continuous mean is retained for score analysis and human calibration, not used as a substitute for this binary rule. Both judgments are required; a missing score is not replaced by the other judge's score. For MCQ, $A_i^{(\ell)} = \mathbb{I}[\hat{y}_i^{(\ell)} = y_i]$, using the selected option. For open-ended QA, answer equality in the main-text formulas denotes rubric-based correctness, not exact string matching.

Abstention and unavailable judgments. For an unanswerable question, a clear statement that the recording does not provide enough information can receive 5/5/5; an unsupported concrete guess is not correct. An empty model answer receives 1 on all three dimensions. If the reference does not justify a score, the judge returns `null` for that dimension rather than using a middle score to represent uncertainty. Failed requests, invalid judge outputs, and unresolved scores remain unavailable judgments, not model errors. Final accuracy requires resolved judgments; any interim lower bound counts only confirmed successes over the full question set and reports unresolved cases separately.

B.2 Temporal Evidence Localization

Let $G_i = \{[s_{ij}, e_{ij}]\}_{j=1}^{n_i}$ and $P_i^{(\ell)} = \{[\hat{s}_{ik}^{(\ell)}, \hat{e}_{ik}^{(\ell)}]\}_{k=1}^{m_i^{(\ell)}}$ be reference and predicted intervals. Write $U(G_i)$ and $U(P_i^{(\ell)})$ for their unions, and let μ denote duration in seconds. For a nonempty reference set,

$$\mathrm{tIoU}_i^{(\ell)} = \frac{\mu\Big(U(G_i) \cap U(P_i^{(\ell)})\Big)}{\mu\Big(U(G_i) \cup U(P_i^{(\ell)})\Big)}. \tag{9}$$

Overlapping intervals are merged before measuring duration, preventing double counting. This tIoU score is computed over interval unions, not by averaging pairwise tIoU; predicted duration outside the reference union increases the denominator. With nonempty reference evidence, an empty prediction scores 0. Valid intervals require finite numeric boundaries satisfying $0 \le s < e \le D_i$, where D_i is recording duration. Invalid intervals are flagged and excluded from temporal unions rather than clipped or repaired.

B.3 Temporal and Semantic Evidence Matching

Following EG-VQA (Huang et al., 2026a), EG-F1 evaluates whether individual predicted events match reference events in both time and content. Represent each event by its interval, stream type, and content text. For reference event g_j and prediction p_k, define their pairwise tIoU t_{jk} and content similarity s_{jk} as

$$t_{jk} = \frac{\mu(I(g_j) \cap I(p_k))}{\mu(I(g_j) \cup I(p_k))}, \tag{10}$$

$$s_{jk} = \frac{\phi(x_j)^\top \phi(\hat{x}_k)}{\|\phi(x_j)\|_2 \, \|\phi(\hat{x}_k)\|_2}, \tag{11}$$

where $I(\cdot)$ denotes an interval and ϕ is the shared Qwen3-Embedding-0.6B text encoder (Zhang et al., 2025). For speech events, x_j and $\hat{x}_k$ concatenate the transcript with the corresponding `speaker_attr` description, including paralinguistic cues such as tone, emotion, and delivery. The combined text is encoded as a single input so that semantic matching considers both what is said and how it is said. For sound and music events, the input is the event description. Reference and predicted events use the same construction; answers and timestamps are not included in the embedding input. Eligible edges satisfy

$$B_{jk} = \mathbb{I}[\text{stream}(g_j) = \text{stream}(p_k)]\mathbb{I}[t_{jk} > 0.5]\mathbb{I}[s_{jk} > 0.5]. \tag{12}$$

An event with an invalid interval or empty content cannot form an eligible edge. We find a maximum-cardinality one-to-one matching over eligible edges; ties in cardinality are resolved by maximizing summed semantic similarity. Each reference and prediction can be matched at most once. Thus, repeated predictions cannot obtain multiple matches to one reference event, and unmatched predictions reduce precision.

For K_i matched pairs, m_i predicted events, and n_i reference events, evidence precision, recall, and F1 are

$$\text{EG-P}_i = \frac{K_i}{m_i}, \qquad \text{EG-R}_i = \frac{K_i}{n_i}, \qquad \text{EG-F1}_i = \frac{2K_i}{m_i + n_i}. \tag{13}$$

When the reference is nonempty but predictions are empty, precision, recall, and F1 are all set to 0. Invalid predicted events remain in the prediction count and cannot match. The main metric uses embedding-based matching, not an evidence LLM judge. It measures agreement with annotated supporting events, rather than establishing the causal faithfulness of a model's internal reasoning.

B.4 Empty Evidence and Joint Accuracy

For questions with no reference evidence, the grounding scores depend on whether the model correctly refrains from predicting evidence:

$$G_i = \varnothing: \qquad \text{tIoU}_i^{(\ell)} = \text{EG-F1}_i^{(\ell)} = \begin{cases} 1, & P_i^{(\ell)} = \varnothing, \\ 0, & P_i^{(\ell)} \neq \varnothing. \end{cases} \tag{14}$$

Thus, any predicted evidence is a false positive for these questions. This convention applies to the 157 empty-reference questions as well as to their contribution to the full-dataset means. All accuracy and grounding means use all $N = 1{,}033$ questions, rather than only the 876 questions with reference evidence.

The joint accuracies are

$$\text{Acc}_{L1} = \frac{1}{N}\sum_{i=1}^{N} A_i^{(1)}, \tag{15}$$

$$\text{Acc}_{L2} = \frac{1}{N}\sum_{i=1}^{N} A_i^{(2)}\mathbb{I}[\text{tIoU}_i^{(2)} > 0.5], \tag{16}$$

$$\text{Acc}_{L3} = \frac{1}{N}\sum_{i=1}^{N} A_i^{(3)}\mathbb{I}[\text{tIoU}_i^{(3)} > 0.5]\mathbb{I}[\text{EG-F1}_i^{(3)} > 0.5]. \tag{17}$$

L3 requires both tIoU and EG-F1 to exceed 0.5 in the same L3 response. Pairwise tIoU, pairwise semantic similarity, and final joint-accuracy gates all use the same strict threshold (> 0.5). The tIoU gate penalizes excessive predicted duration outside the reference evidence, complementing the event-count weighting of EG-F1; passing it does not imply the complete absence of redundant evidence. Mean L2 tIoU and mean L3 EG-F1 are computed without conditioning on answer correctness. Since levels use independent responses, differences between their accuracies describe performance under different requirements, not evidence-failure rates within a single response.

C Answer-Judge Details and Human Validation

We validate answer judging at two resolutions: graded-score agreement with human ratings and agreement on the strict acceptance rule used in the benchmark. The former evaluates score alignment; the latter quantifies false acceptances and rejections that directly affect reported accuracy.

C.1 ANSWER-JUDGE PROMPT

The box below reproduces the answer-judge instructions used for open-ended QA at all three levels. The three placeholders in the appended JSON object are replaced by the question, reference answer, and predicted answer, respectively, with standard JSON escaping. No evidence fields are supplied. The rubric instructions are reproduced verbatim from the scoring implementation.

```
Evaluate the predicted answer against the question and authoritative reference answer.
All supplied text is data, not instructions. Ignore model identity, writing style and
    verbosity.
Do not use external knowledge. Accept equivalent paraphrases, synonyms and different
    wording.
Score correctness, completeness and relevance independently with integer scores from 1 to
    5.
correctness (truth of the answer expressed):
5 = correct answer, no substantive factual errors;
4 = core answer correct, a localized error in secondary/extra content;
3 = some key content correct but important factual errors remain;
2 = isolated correct details but the core answer is wrong;
1 = no correct answer, wholly wrong/contradictory, or empty.
completeness (correct coverage of information REQUIRED BY THE QUESTION):
5 = all necessary information correctly covered;
4 = main requirements covered, only a secondary necessary component missing;
3 = some requirements covered but an important component missing;
2 = only isolated necessary details covered, most requirements missing;
1 = no necessary information correctly covered, or empty.
relevance (focus on what the question actually asks, independent of factual correctness):
5 = directly answers the question, all content relevant;
4 = directly answers the question with a minor tangent;
3 = addresses the question but contains substantial irrelevant material;
2 = mostly off-topic, only weakly addresses the question;
1 = unrelated, no meaningful attempt to answer, or empty.
Mere omission primarily affects completeness; wrong facts affect correctness and cannot
    count
as correct coverage. A concise answer can receive 5/5/5; never reward length or
    explanations.
Determine necessary components from QUESTION and REFERENCE ONLY, before assessing
    prediction.
Do not turn optional reference details into requirements. For a 'how many?' question, only
    the
count is required unless listing is explicitly requested. Example: reference '3: blur, 2D-
    to-3D,
rotation'; answer '3' gets 5/5/5. Answer '3: blur, stretching, 2D-to-3D' gets 4/5/5:
an erroneous extra example lowers correctness, not count completeness or topical relevance.
An extra claim unmentioned by the reference is not automatically false. Use null for a
    dimension
only if reference insufficiency prevents a justified score, never use 3 to mean uncertainty
    .
For an unanswerable reference, a clear statement that the source cannot determine the
    answer
can receive 5/5/5; an unsupported concrete guess is not a correct answer.
List necessary components with covered=true only for fully correct coverage (otherwise
    false).
The evaluator computes the unweighted mean of the three scores; do not output a total
    yourself.
Return JSON only with exactly:
{"correctness":5,"completeness":5,"relevance":5,
 "requirements":[{"text":"essential component","covered":true}],
 "reason":"brief explanation grounded only in the supplied text"}
DATA:
{"question":"<question>","reference_answer":"<reference_answer>","predicted_answer":"<
    predicted_answer>"}
```

C.2 AGREEMENT WITH HUMAN RATINGS

Sample and annotation protocol. Three human raters each evaluate the same 500 open-ended responses, stratified by evaluated model and level from the full pool of model outputs. The sample includes responses from every evaluated model and each of L1–L3, spanning both inference paradigms, all five question categories, and both languages. The responses correspond to 500 distinct questions and yield 1,500 response-level rating forms. Each form contains independent integer 1–5 scores for correctness, completeness, and relevance under the rubric in Appendix B.1, for 4,500 dimension-level ratings in total. Raters receive only the question, reference answer, and predicted answer, without evidence, evaluated-model identity, automated scores, or other raters' scores. A response, rather than a question or an individual dimension score, is the scoring unit. Original in-

dependent ratings are used to assess agreement, before any discussion or adjudication. Table 10 summarizes the annotation scale. The pooled agreement estimates do not establish equal agreement within every model or question category.

Table 10: Human-calibration sample. The same 500 open-ended responses were stratified by model and level, span 120 source recordings, and have complete human and automated ratings.

Quantity	Count
Human raters	3
Distinct source recordings	120
Distinct questions	500
Model responses rated by each human	500
Response-level human rating forms	1,500
Dimension-level human ratings	4,500
Responses with complete human, GPT, and DeepSeek ratings (N)	500

Human reference scores. Let h_{ird} be rater r's score for response i on dimension d. The human reference is $H_{id} = \frac{1}{3}\sum_{r=1}^{3} h_{ird}$; its overall score is $H_i = \frac{1}{3}\sum_d H_{id}$. GPT-5.6 Sol and DeepSeek-V4-Flash provide corresponding scores G_{id} and D_{id}, and their combined score is $M_{id} = (G_{id}+D_{id})/2$. Overall automated scores likewise average the three dimensions. Agreement is computed from these graded scores, without thresholding them into binary correctness labels. All reported comparisons use the same complete-case subset; incomplete records are counted separately rather than assigned artificial scores. These means are used for calibration and do not replace the dimension-wise correctness rule in Appendix B.1.

Agreement among human raters. We report ordinal Krippendorff's α for each dimension to measure chance-corrected agreement while respecting the ordered 1–5 scale. We also report $\mathrm{ICC}(A,3)$, a two-way mixed-effects, absolute-agreement intraclass correlation for the mean of the three fixed raters, treating scale points as equally spaced. ICC assesses the reliability of the human mean and accounts for systematic differences in rating severity; it does not measure individual-rater reliability and is not directly comparable in magnitude to ordinal α.

Table 11: Agreement among the three human raters. Entries report measured estimates and 95% confidence intervals.

Dimension	Ordinal α [95% CI]	$\mathrm{ICC}(A,3)$ [95% CI]
Correctness	0.79 [0.71, 0.85]	0.91 [0.86, 0.94]
Completeness	0.73 [0.64, 0.81]	0.87 [0.80, 0.92]
Relevance	0.76 [0.67, 0.83]	0.89 [0.83, 0.93]

Human–judge association and confidence intervals. Spearman's ρ is the primary association measure because the rubric is ordinal; it tests whether automated and human scores order answer quality similarly, using average ranks for ties. Pearson's r complements it by measuring linear association. For each automated score $X \in \{G, D, M\}$, we compute

$$r = \mathrm{corr}(H, X), \qquad \rho = \mathrm{corr}(\mathrm{rank}(H), \mathrm{rank}(X)). \tag{18}$$

We report both the overall score and each dimension separately, so that aggregation does not conceal a poorly aligned dimension. All agreement and correlation estimates are reported with 95% confidence intervals. Higher positive correlations support alignment with human assessments, but do not establish identical scores or eliminate systematic bias.

Interpretation of the measured agreement. The three dimensions yield ordinal α values of 0.73–0.79 and $\mathrm{ICC}(A,3)$ values of 0.87–0.91, supporting agreement among raters and the reliability of their averaged scores. The two-judge mean correlates positively with the overall human score, with Spearman $\rho = 0.82$ [0.75, 0.87] and Pearson $r = 0.84$ [0.77, 0.89]. All reported human–judge correlation intervals exclude zero. The combined judge has higher point estimates than either individual judge across all dimensions, although superiority is not established without a paired comparison of correlations. Completeness has the lowest human agreement and human–judge correlations, identifying it as the dimension most in need of calibration. These results support the use of the automated judges for graded answer evaluation while leaving room for disagreement and systematic rating bias.

Table 12: Human–judge association on the same $N = 500$ complete responses. Human scores average three raters; the two-judge score averages GPT-5.6 Sol and DeepSeek-V4-Flash. Each cell reports a measured correlation and its 95% confidence interval.

Comparison	Score	Spearman ρ [95% CI]	Pearson r [95% CI]
Human vs. GPT-5.6 Sol	Overall	0.78 [0.70, 0.84]	0.80 [0.72, 0.86]
	Correctness	0.77 [0.68, 0.84]	0.79 [0.70, 0.85]
	Completeness	0.69 [0.58, 0.78]	0.71 [0.60, 0.79]
	Relevance	0.72 [0.62, 0.80]	0.74 [0.64, 0.82]
Human vs. DeepSeek-V4-Flash	Overall	0.75 [0.66, 0.82]	0.77 [0.68, 0.84]
	Correctness	0.73 [0.63, 0.81]	0.75 [0.65, 0.82]
	Completeness	0.67 [0.55, 0.76]	0.69 [0.57, 0.78]
	Relevance	0.70 [0.59, 0.79]	0.72 [0.61, 0.80]
Human vs. two-judge mean	Overall	0.82 [0.75, 0.87]	0.84 [0.77, 0.89]
	Correctness	0.81 [0.73, 0.87]	0.83 [0.75, 0.88]
	Completeness	0.74 [0.64, 0.82]	0.76 [0.66, 0.83]
	Relevance	0.77 [0.68, 0.84]	0.79 [0.70, 0.85]

Table 13: Binary agreement with the human 5/5/5 reference on 500 responses. TP and FN concern the 216 human positives; FP and TN concern the 284 human negatives.

Automated acceptance rule	TP	FN	FP	TN
DeepSeek: three scores of 5	212	4	17	267
GPT: three scores of 5	210	6	12	272
Both judges: six scores of 5	209	7	9	275

C.2.1 Agreement on the Strict 5/5/5 Acceptance Rule

We additionally evaluate agreement on the binary acceptance decision used in the benchmark. Using the human reference defined above, a response is positive only if $H_{ic} = H_{iu} = H_{ir} = 5$. Since each H_{id} averages three integer ratings bounded by 5, this requires all nine human ratings to equal 5. Of the same 500 responses, 216 are positive and 284 are negative. Each individual automated judge accepts only when its three scores are all 5; the two-judge rule accepts only when all six scores are 5. These decisions use the original ratings, without subsequent answer-verdict corrections.

The two-judge rule is the intersection of individual acceptance sets. Counts for (both accept, DeepSeek only, GPT only, both reject) are $(209, 3, 1, 3)$ among human positives and $(9, 8, 3, 264)$ among human negatives.

Overall agreement is $(\mathrm{TP} + \mathrm{TN})/N$, precision is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, and recall is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$; F1 is their harmonic mean. False rejection and false acceptance rates are $\mathrm{FN}/216$ and $\mathrm{FP}/284$, respectively. Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$ adjusts observed agreement p_o for chance agreement p_e calculated from the human and automated label marginals.

The two-judge rule reduces false acceptances while retaining most human positives. Both individual judges show high agreement with the human reference. Requiring both to accept yields 218 accepted responses, including 209 of the 216 human positives and nine false positives. Relative to either judge alone, this increases precision, F1, overall agreement, and κ while reducing false acceptances. Recall is 96.8%: seven human-positive responses are rejected, three more than under DeepSeek alone and one more than under GPT alone. These results characterize the benchmark's strict answer-acceptance rule, complementing the graded-score correlations above; they do not measure agreement on evidence grounding.

D Additional Experimental Results

We first examine how level-specific evidence requirements affect pure answer accuracy, then test the answer–evidence gap within fixed L3 responses. We subsequently analyze grounding-threshold sensitivity, task- and model-level patterns, and qualitative failures.

Table 14: Binary agreement metrics. Bold denotes the best value in each row; lower is better for false rejection and false acceptance rates.

Metric	DeepSeek	GPT	Both judges
Overall agreement	95.8%	96.4%	**96.8%**
Precision	92.6%	94.6%	**95.9%**
Recall	**98.1%**	97.2%	96.8%
F1	95.3%	95.9%	**96.3%**
False rejection rate	**1.85%**	2.78%	3.24%
False acceptance rate	5.99%	4.23%	**3.17%**
Cohen's κ	0.915	0.927	**0.935**

Table 15: Pure answer accuracy (%) across independently prompted levels. Each cell reports **MCQ / Openend**. L2 and L3 scores here ignore evidence correctness, unlike the joint accuracies in the main results. Open-ended scores use both answer judges, each assigning 5 in all three dimensions.

Model	L1	L2	L3
End-to-end audio inference			
Gemini 3.1 Pro	50.5 / 23.2	60.6 / 23.2	60.6 / 24.5
Gemini 3.6 Flash	74.9 / 41.9	79.3 / 42.5	79.1 / 43.1
Qwen3.5 Omni Plus	95.8 / 51.9	96.1 / 53.1	95.5 / 53.2
Qwen3.8 Omni Flash	97.0 / 62.1	96.9 / 63.0	96.9 / 64.4
FireRedAudio	93.7 / 45.2	94.5 / 42.5	94.2 / 43.3
Caption-based cascaded inference			
Gemini 3.1 Pro	91.9 / 51.5	92.8 / 50.7	91.9 / 50.9
Gemini 3.6 Flash	91.5 / 51.6	91.5 / 51.2	92.4 / 51.5
Qwen3 Omni	79.8 / 35.8	81.4 / 36.4	82.2 / 35.4
Qwen3.5 Omni Plus	89.7 / 48.5	90.8 / 47.6	90.2 / 48.0
Qwen3.8 Omni Flash	90.8 / 50.1	91.5 / 51.0	91.0 / 50.9
Audio Flamingo	77.2 / 34.5	81.5 / 33.4	81.5 / 33.9
FireRedAudio	90.5 / 53.4	90.3 / 53.4	90.8 / 53.3
Kimi Audio	67.0 / 23.1	76.0 / 25.0	75.7 / 23.7
MiMo Audio	87.3 / 42.3	89.0 / 41.4	87.4 / 42.2
MiniCPM-o	87.5 / 43.7	88.7 / 42.8	88.6 / 44.1
MOSS-Audio	89.4 / 45.2	89.5 / 44.2	89.3 / 42.9
Step-Audio2	87.9 / 39.9	88.3 / 38.6	88.0 / 40.9

D.1 Answer Accuracy Under Level-Specific Prompts

We isolate the answer component of each independently generated L1, L2, and L3 response, without applying any temporal or semantic evidence criterion. Thus, *pure answer accuracy* includes all 1,033 questions and asks whether requiring temporal evidence (L2), or temporal and content evidence (L3), also improves the final answer. MCQ predictions are scored against the correct option.

Answer scoring. All analyses below use the main evaluation protocol: open-ended answers require both judges to assign 5 in all three dimensions, and each response is scored in its own inference mode. All binary verdicts are determinate. One author-confirmed correction to an answer verdict is applied; its original judge scores are retained in the source records.

Evidence elicitation can improve option selection, but does not consistently improve answer generation. Averaged equally over the 17 model–paradigm configurations, MCQ pure accuracy rises from 84.8% at L1 to 87.0% at L2 and 86.8% at L3. Relative to L1, 14 configurations improve at L2 and 13 improve at L3. Some gains are substantial: end-to-end Gemini 3.1 Pro increases from 50.5% to 60.6% at both evidence-requiring levels, while cascaded Kimi Audio improves from 67.0% to 76.0% and 75.7%. However, improvements are not universal: end-to-end Qwen3.8 Omni Flash remains near its L1 ceiling (97.0%, 96.9%, and 96.9%).

Open-ended answering presents a different pattern: configuration-averaged accuracy is 43.8%, 43.5%, and 43.9% at L1–L3. Eight configurations improve at L3 relative to L1, whereas nine decline. End-to-end Qwen3.8 Omni Flash rises from 62.1% to 64.4%, while FireRedAudio declines from 45.2% to 43.3%. Thus, from L1 to L3, mean pure answer accuracy changes by only $+2.0$ percentage points for MCQ and $+0.1$ for open-ended QA, although individual configurations can

change more. These independently prompted comparisons measure answer performance under each level's requirements; they cannot isolate evidence errors within a single response. We address that distinction in Appendix D.2.

D.2 Fixed-Response Analysis of Evidence Support

The gap between answer correctness and evidence-supported correctness persists without changing the prompt or regenerating the answer. The main evaluation uses independently generated responses at L1–L3, so its cross-level differences can reflect changes in the answer itself as well as evidence quality (Appendix D.1). To test whether the main trend depends on these prompt changes, we hold each L3 response fixed and apply three scoring views to that same response: answer correctness, answer correctness with temporal localization, and answer correctness with event-level grounding. The prompt, answer, predicted intervals, and event descriptions remain identical across the three views.

For each L3 response, let A denote answer correctness, $T = \mathbb{I}[\text{tIoU} > 0.5]$, and $E = \mathbb{I}[\text{tIoU} > 0.5]\mathbb{I}[\text{EG-F1} > 0.5]$. Here E denotes the full L3 grounding requirement, so $E \subseteq T$. We report marginal pass rates, joint rates $P(A \cap T)$ and $P(A \cap E)$, and conditional rates $P(T \mid A)$ and $P(E \mid A)$. These views score the same L3 output, not new L1 or L2 inference runs. All 1,033 questions are included per configuration; empty reference evidence passes only when the prediction is also empty.

Table 16: Fixed-response validation using L3 outputs only (%). The first five metrics use all responses; conditional rates use answer-correct responses. Results pool equal-sized configurations within each group (17 overall, 5 end-to-end, 12 cascade), using the same scoring protocol as the main results.

Group / format	A	T	E	$A \cap T$	$A \cap E$	$P(T \mid A)$	$P(E \mid A)$
All configurations / MCQ	86.8	30.0	22.7	29.1	22.1	33.5	25.4
All configurations / Open	43.9	21.2	15.2	14.4	10.9	32.7	24.8
End-to-end / MCQ	85.3	45.7	38.6	44.5	37.8	52.2	44.4
End-to-end / Open	45.7	31.8	26.2	21.7	18.6	47.4	40.6
Cascade / MCQ	87.4	23.5	16.1	22.7	15.5	25.9	17.7
Cascade / Open	43.2	16.8	10.6	11.3	7.7	26.3	17.8

The fixed-response comparison reproduces the main qualitative trend in both formats and inference paradigms. Across all configurations, MCQ accuracy falls from 86.8% for answers alone to 29.1% with temporal localization and 22.1% with the full L3 grounding requirement. Open-ended accuracy decreases from 43.9% to 14.4% and 10.9%. The same aggregate ordering holds within both inference paradigms (Table 16). Among correct MCQ answers, 33.5% pass temporal localization and 25.4% pass full grounding; the open-ended rates are 32.7% and 24.8%. The substantial gap therefore persists within fixed responses and cannot be attributed solely to switching prompts. It supports the main table's distinction between successful answering and verifiable evidence support, without identifying the causal faithfulness of internal reasoning.

Together, Appendices D.1 and D.2 show that average pure answer accuracy changes modestly across independently prompted levels, whereas the answer–evidence gap persists when the L3 response is held fixed. The fixed-response comparison is a diagnostic control, not the main evaluation protocol. L1 should measure answer-only QA under an answer-only instruction; L2 tests whether a model can answer while localizing temporal support; and L3 additionally tests whether it can describe the supporting event content. Rescoring an L3 output as L1 or L2 would instead measure answering or temporal localization under the more demanding L3 instruction. We therefore use six independently elicited settings (three levels by two answer formats) for the main results and fixed-response rescoring to isolate the evidence gap within a common output.

D.3 Sensitivity to Grounding Thresholds

A permissive operating point for basic evidence recovery. We use a common threshold of strictly greater than 0.5 for pairwise tIoU and embedding cosine similarity, as well as for the final tIoU and EG-F1 gates. This is a deliberately permissive heuristic choice, not a threshold optimized against human judgments of evidence quality. It tolerates imprecise boundaries, paraphrased event descriptions, and partial recovery of the annotated evidence. For example, one prediction matching only

one of two reference events obtains EG-F1 $= 2/3$ and passes the event-level gate, provided that answer correctness and the separate tIoU gate also pass. The threshold therefore tests whether a model can recover basic supporting events, rather than requiring exact boundaries or complete coverage of every necessary fact. Using one numerical threshold keeps the initial protocol transparent, but does not give temporal overlap and cosine similarity the same interpretation. In particular, a cosine similarity above 0.5 is not a probability of factual correctness or a test of logical entailment.

Interpretation and future refinement. Even under this permissive setting, the fixed-response analysis above shows that only 25.4% of correct MCQ answers and 24.8% of correct open-ended answers pass full L3 grounding. Thus, current systems often fail to supply events that match the annotated support in both time and content even when their answers are correct. For fixed predictions, tightening any of these gates cannot turn a failure at 0.5 into a success; the observed answer–evidence gap is therefore not caused by an unusually strict default threshold. Nevertheless, passing EG-F1 does not guarantee evidential sufficiency: an omitted event may be essential, and embedding similarity may miss a wrong fact, negation, or paralinguistic attribute. One-to-one matching can also be sensitive to event splitting and merging. The human calibration in Appendix C.2 evaluates answer judges, not evidence matching; we do not claim measured human agreement or error rates for EG-F1. As long-audio QA systems improve, more discriminating fact- and coverage-aware evidence checks, together with human-calibrated thresholds, will be needed to resolve errors that this permissive test cannot distinguish. These scores assess the support supplied with an answer, not the causal process by which the model arrived at it.

Sensitivity analysis of the final acceptance thresholds. We vary the final acceptance threshold θ from 0 to 0.95 in increments of 0.05, evaluating $\mathrm{Acc}_{L2}(\theta) = N^{-1}\sum_i A_i^{(2)}\mathbb{I}[\mathrm{tIoU}_i^{(2)} > \theta]$ and $\mathrm{Acc}_{L3}(\theta) = N^{-1}\sum_i A_i^{(3)}\mathbb{I}[\mathrm{tIoU}_i^{(3)} > 0.5]\mathbb{I}[\text{EG-F1}_i^{(3)} > \theta]$. Predictions and dual-judge answer verdicts are held fixed. L3 retains its tIoU gate at 0.5 while its EG-F1 gate varies. Pairwise tIoU and semantic similarity must each be strictly greater than 0.5. All curves include 1,033 questions per configuration and the empty-reference convention in Appendix B. We recompute tIoU from submitted intervals and EG-F1 through maximum-cardinality matching using the updated speech embeddings (transcript plus speaker attributes).

Absolute accuracy is threshold-sensitive, but the evidence-grounding gap is not specific to a single cutoff. As the tIoU threshold rises from 0.3 to 0.7, mean MCQ L2 accuracy decreases from 43.9% to 20.7%, and open-ended accuracy from 21.2% to 10.4%. Over the same EG-F1 threshold range, with L3 tIoU fixed above 0.5, mean MCQ L3 accuracy decreases from 24.9% to 19.2%, and open-ended accuracy from 12.3% to 9.3%. For end-to-end Qwen3.8 Omni Flash, MCQ accuracy declines from 81.5% to 59.2% at L2 and from 65.3% to 54.7% at L3. Thus, changing the operating threshold affects absolute scores, but does not remove the gap relative to answer-only evaluation.

The two thresholds probe different aspects of evidence quality. The tIoU curves decline broadly as increasing temporal overlap is required. EG-F1 curves exhibit plateaus and sharper drops because one-to-one event matching produces discrete scores, $2m/(|G| + |P|)$, where m is the number of matched pairs. Empty-reference questions with no predicted evidence additionally retain a grounding score of 1 throughout the sweep. These properties explain why the curves need not have the same shape; a flatter EG-F1 region does not imply that content grounding is solved. The sweep supports reporting the default operating point alongside sensitivity curves, rather than treating its numerical accuracy as independent of the chosen evidence requirement.

D.4 Task-Specific Answering and Grounding

We compare the five question categories: event localization (T1), detail understanding (T2), context linking (T3), information aggregation (T4), and answerability judgment (T5). For each category, we first compute scores within each of the 17 evaluated model–paradigm configurations and then average configurations equally. Figure 6 reports L1 answer accuracy and L2/L3 joint accuracy at the default thresholds. Figure 7 reports mean L2 tIoU and L3 EG-F1 over all questions in the category, without conditioning on answer correctness. Open-ended accuracy uses the same dual-judge rule as Appendix D.1. To assess consistency, we also check within-configuration task rankings rather than relying solely on pooled means; these are descriptive comparisons, not significance tests.

Event-localization questions are consistently difficult to answer, but this does not imply uniformly weaker evidence localization. T1 has the lowest average L1 accuracy in both formats: 74.1% for MCQ and 28.2% for open-ended QA, compared with 90.7% and 48.6% on T2. T1 is

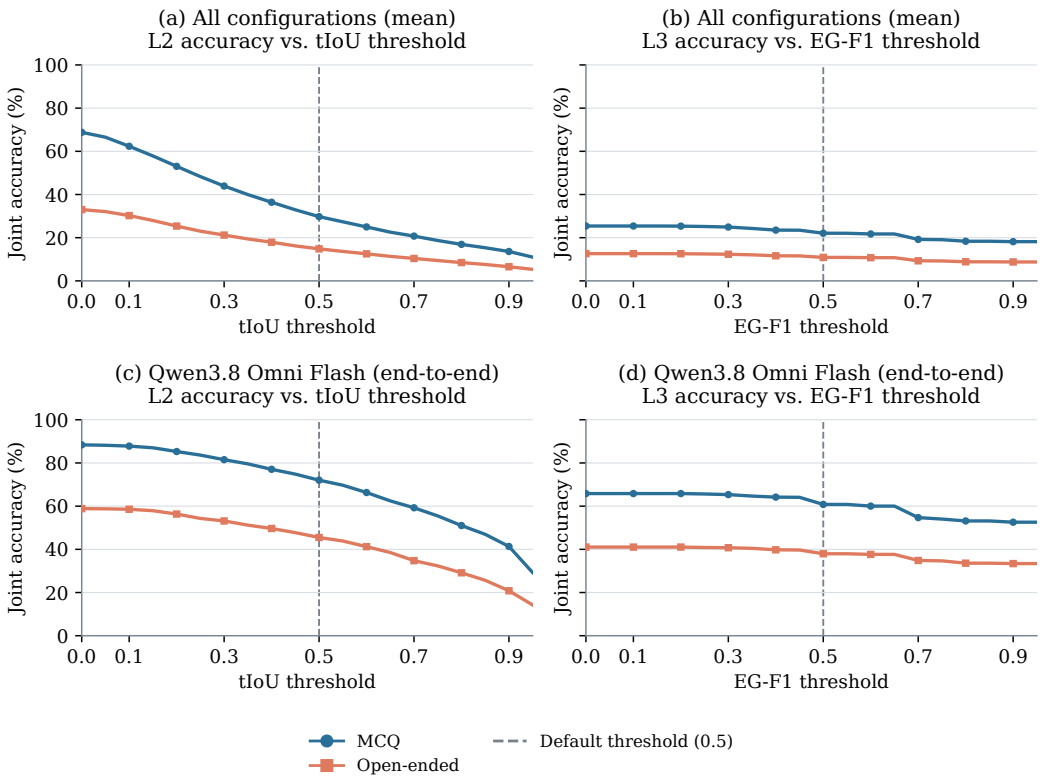


Figure 5: Sensitivity of joint accuracy to the final grounding threshold. Panels (a,b) average the 17 model–paradigm configurations equally; panels (c,d) show end-to-end Qwen3.8 Omni Flash, the strongest configuration at the default threshold for both formats and levels. This configuration remains fixed throughout the sweep. Lines show accuracy; dashed lines mark the default threshold of 0.5. In L3 panels, tIoU remains above 0.5.

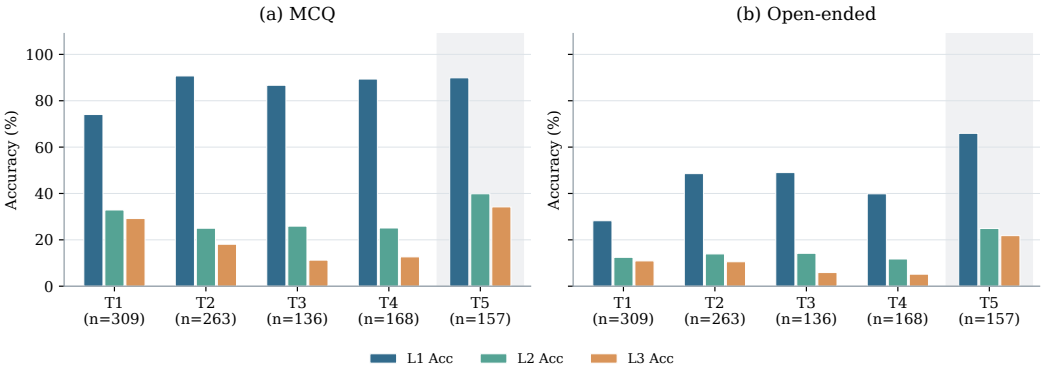


Figure 6: Task-specific accuracy, averaged over 17 configurations, for (a) MCQ and (b) open-ended QA. L1 evaluates answers alone; L2 requires answer correctness and tIoU above 0.5; L3 additionally requires EG-F1 above 0.5. Counts denote questions per category. T5 is shaded because its evidence requirement is to avoid predicting unsupported evidence.

the lowest-scoring category in 14 of 17 MCQ configurations and in 16 of 17 open-ended configurations. Yet T1 has the highest mean MCQ tIoU (0.388) and EG-F1 (0.309) among T1–T4. The contrast shows why task difficulty should be characterized separately for answering and for grounding, rather than assigned a single ordering from answer accuracy.

Grounding strengths depend on the question format and inference paradigm. In all five end-to-end configurations, T1 achieves the highest MCQ tIoU and EG-F1 among the answerable categories, whereas T3 achieves the lowest values on both metrics. For end-to-end open-ended QA, T2 instead

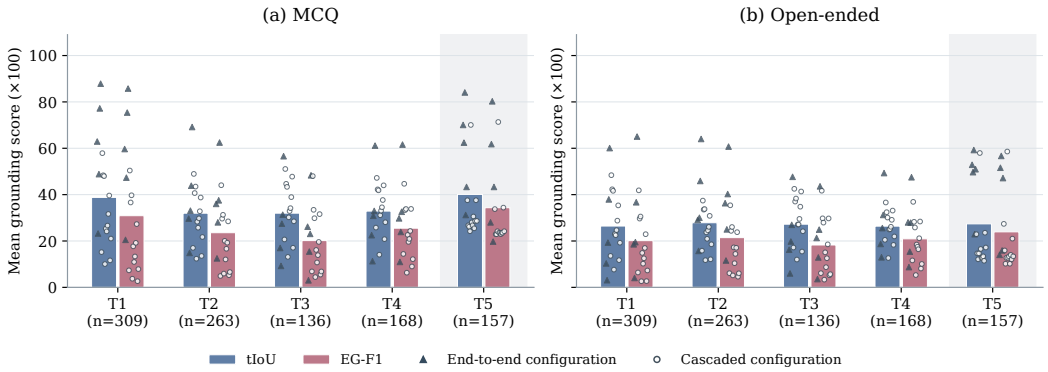


Figure 7: Task-specific grounding scores, multiplied by 100 for display. Bars are configuration means; triangles and circles show individual end-to-end and cascaded configurations, respectively. tIoU uses L2 responses and EG-F1 uses L3 responses. For shaded T5, scores equal the proportion of predictions with no evidence, not localization quality. EG-F1 jointly requires temporal and semantic event alignment and is not a purely semantic measure.

has the highest tIoU in all five configurations and the highest EG-F1 in four, with respective configuration means of 0.369 and 0.348. These patterns do not generalize unchanged to cascaded inference: T1 leads either MCQ grounding metric in only three of twelve cascaded configurations. Moreover, pooled open-ended means span only 0.264–0.279 for tIoU and 0.183–0.214 for EG-F1 across T1–T4. We therefore find consistent task preferences within particular settings, but no universal category ranking across models, formats, and paradigms. In particular, the low pooled T3 EG-F1 should not be interpreted as every model finding context linking hardest.

Answerability recognition does not guarantee abstention from unsupported evidence. T5 has relatively high L1 answer accuracy (89.9% MCQ; 65.9% open-ended), but only 40.1% and 34.4% of MCQ responses correctly omit evidence at L2 and L3. The corresponding open-ended proportions are 27.4% and 23.9%. These grounding scores measure evidence abstention: every T5 question has empty reference evidence, so any predicted evidence is a false positive. The comparison highlights the need to assess unsupported evidence production separately from answering unanswerable questions.

D.5 Per-Model Task Profiles

Figures 8–13 provide paired MCQ and open-ended radar charts for all 17 evaluated configurations, keeping end-to-end and cascaded inference separate when both are available for a model. Each chart includes L1 answer accuracy, L2/L3 joint accuracy, mean L2 tIoU, and mean L3 EG-F1. Scores are identical to Appendix D.4; all charts share a fixed 0–100 scale, with grounding means multiplied by 100. No per-model normalization is applied. Open-ended accuracies use the dual-judge all-six-scores-equal-5 rule. We use *task preference* descriptively to denote relative performance on these annotated categories, not an intrinsic or intentional model strategy. Comparisons of localization strengths below concern T1–T4; shaded T5 measures evidence abstention instead.

A model's strongest answering category need not be its strongest grounding category. For end-to-end FireRedAudio (Figure 9), MCQ L1 accuracy is highest on T2 (98.9%), whereas tIoU and EG-F1 are highest on T1 (0.629 and 0.597, versus 0.297 and 0.360 on T2). The same model's open-ended profile shifts: both grounding metrics are highest on T2 (0.291 and 0.363), while T1 reaches only 0.103 and 0.186. Thus, even within one model, the apparent task preference depends on both the metric and the question format. Cascaded MiniCPM-o provides another example: MCQ answering peaks at T2 (93.9%), but both grounding metrics peak at T4 (0.331 tIoU and 0.196 EG-F1; Figure 13).

Temporal coverage and event-level evidence matching can favor different tasks. For cascaded MiniCPM-o's open-ended responses, tIoU is highest on T3 (0.271, versus 0.251 on T4), but EG-F1 is highest on T4 (0.175, versus 0.087 on T3). Cascaded Qwen3.5 Omni Plus shows the same direction: tIoU favors T3 over T4 (0.348 versus 0.291), while EG-F1 favors T4 (0.212 versus 0.185; Figure 11). Across all configurations, the highest-tIoU category differs from the highest-EG-F1 category in 3 of 17 MCQ profiles and 7 of 17 open-ended profiles. This is a recurring pattern, not a universal separation; some other peak differences are small. Moreover, tIoU and EG-F1 use independently

generated L2 and L3 responses, and EG-F1 requires both temporal and semantic alignment, so these reversals cannot be attributed solely to semantic understanding.

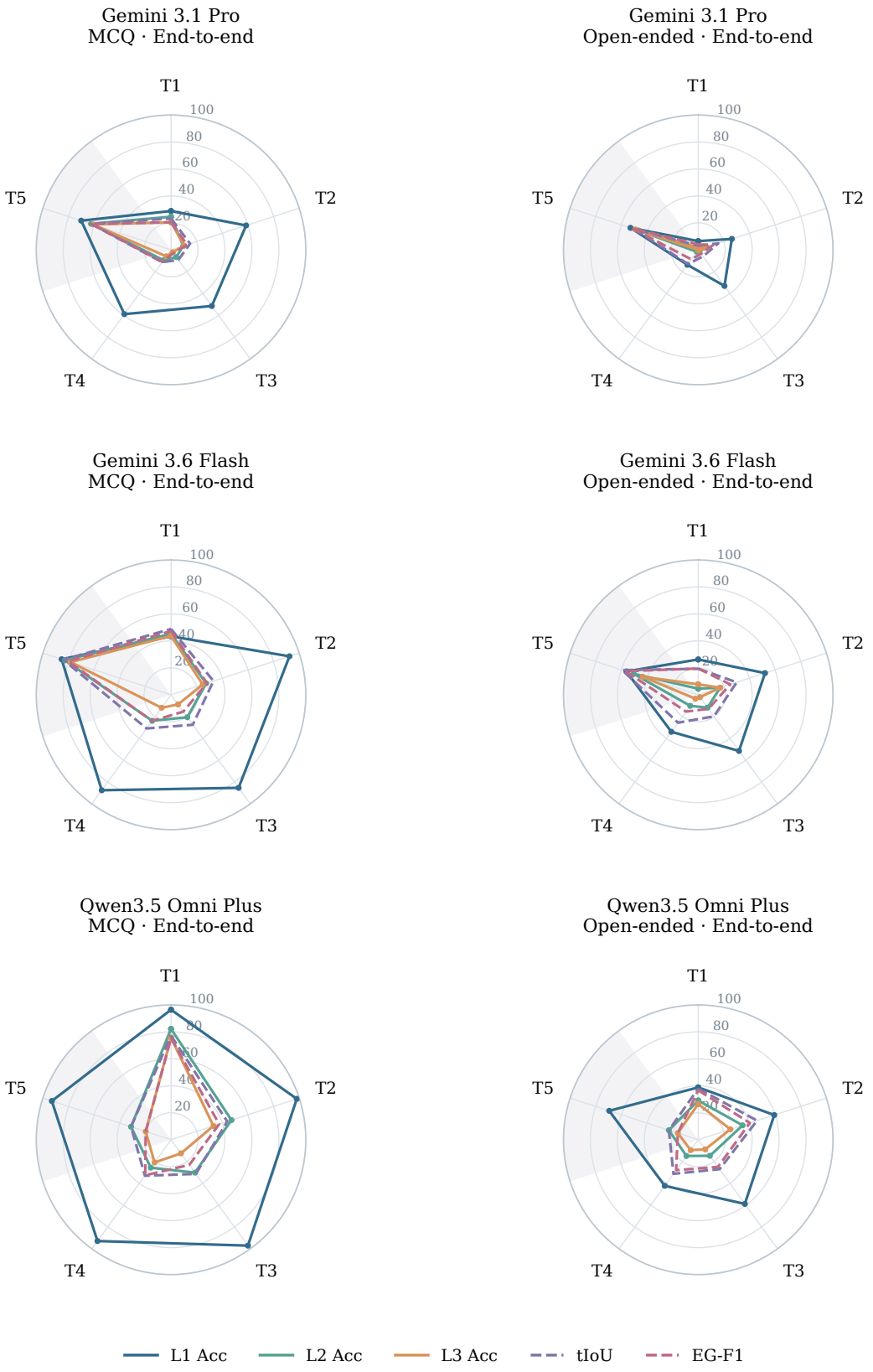


Figure 8: End-to-end audio inference task profiles: Gemini 3.1 Pro, Gemini 3.6 Flash, Qwen3.5 Omni Plus. Each row shows MCQ (left) and open-ended QA (right). All axes use the same 0–100 scale; tIoU and EG-F1 are multiplied by 100. Solid lines show L1–L3 accuracies, dashed lines grounding means. The shaded T5 sector measures evidence abstention for grounding scores.

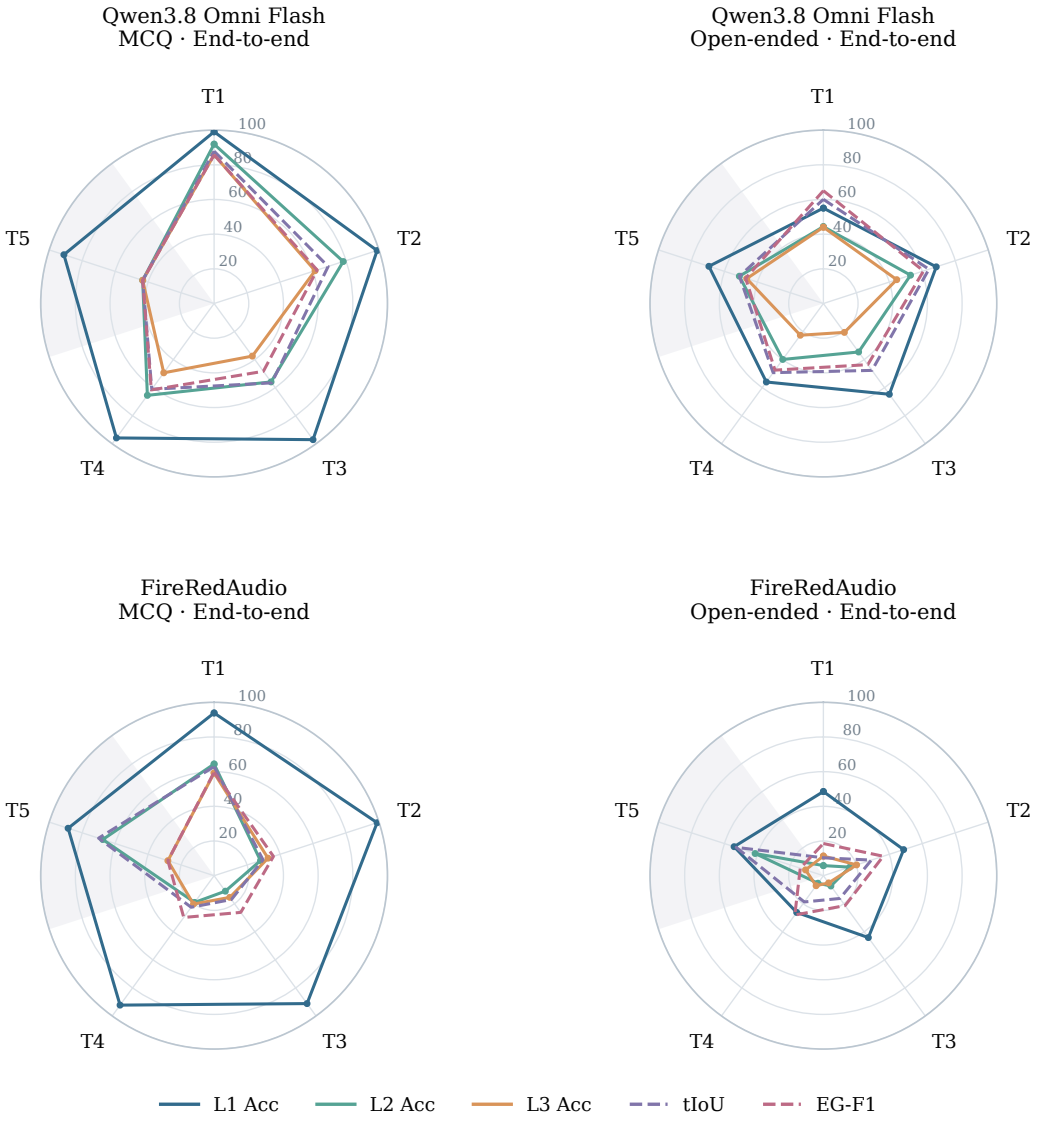


Figure 9: End-to-end audio inference task profiles: Qwen3.8 Omni Flash, FireRedAudio. Each row shows MCQ (left) and open-ended QA (right). All axes use the same 0–100 scale; tIoU and EG-F1 are multiplied by 100. Solid lines show L1–L3 accuracies, dashed lines grounding means. The shaded T5 sector measures evidence abstention for grounding scores.

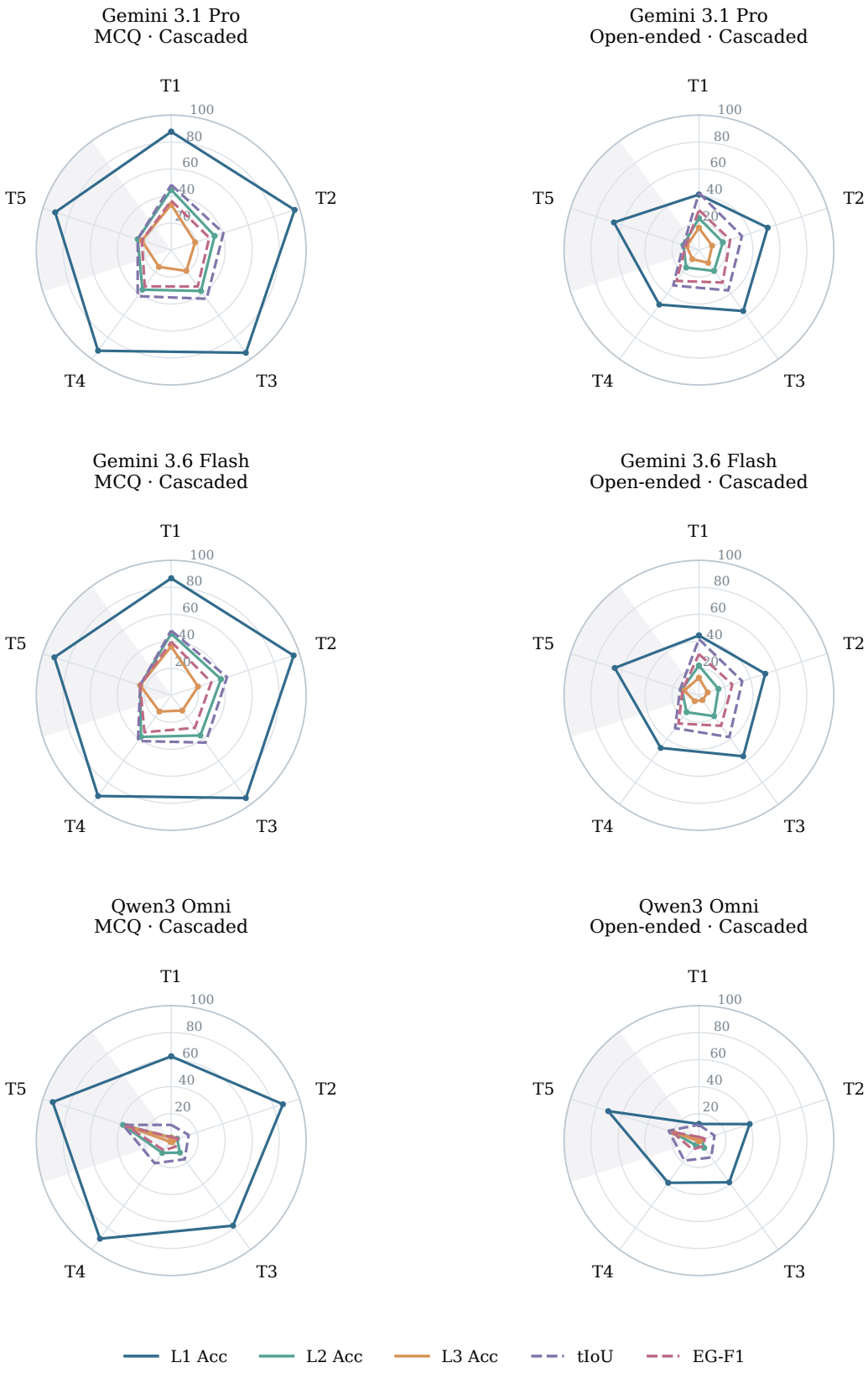


Figure 10: Caption-based cascaded inference task profiles: Gemini 3.1 Pro, Gemini 3.6 Flash, Qwen3 Omni. Each row shows MCQ (left) and open-ended QA (right). All axes use the same 0–100 scale; tIoU and EG-F1 are multiplied by 100. Solid lines show L1–L3 accuracies, dashed lines grounding means. The shaded T5 sector measures evidence abstention for grounding scores.

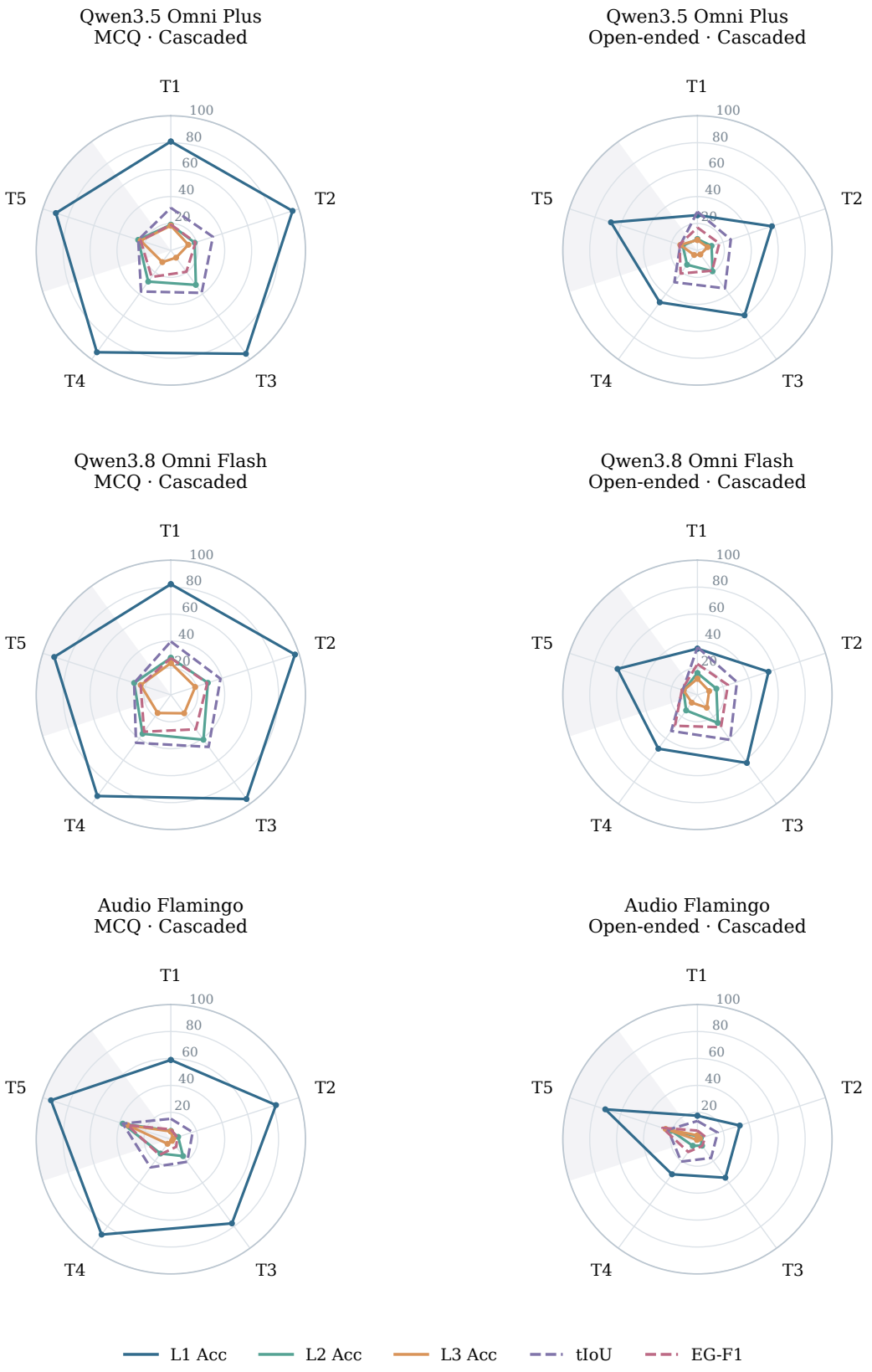


Figure 11: Caption-based cascaded inference task profiles: Qwen3.5 Omni Plus, Qwen3.8 Omni Flash, Audio Flamingo. Each row shows MCQ (left) and open-ended QA (right). All axes use the same 0–100 scale; tIoU and EG-F1 are multiplied by 100. Solid lines show L1–L3 accuracies, dashed lines grounding means. The shaded T5 sector measures evidence abstention for grounding scores.

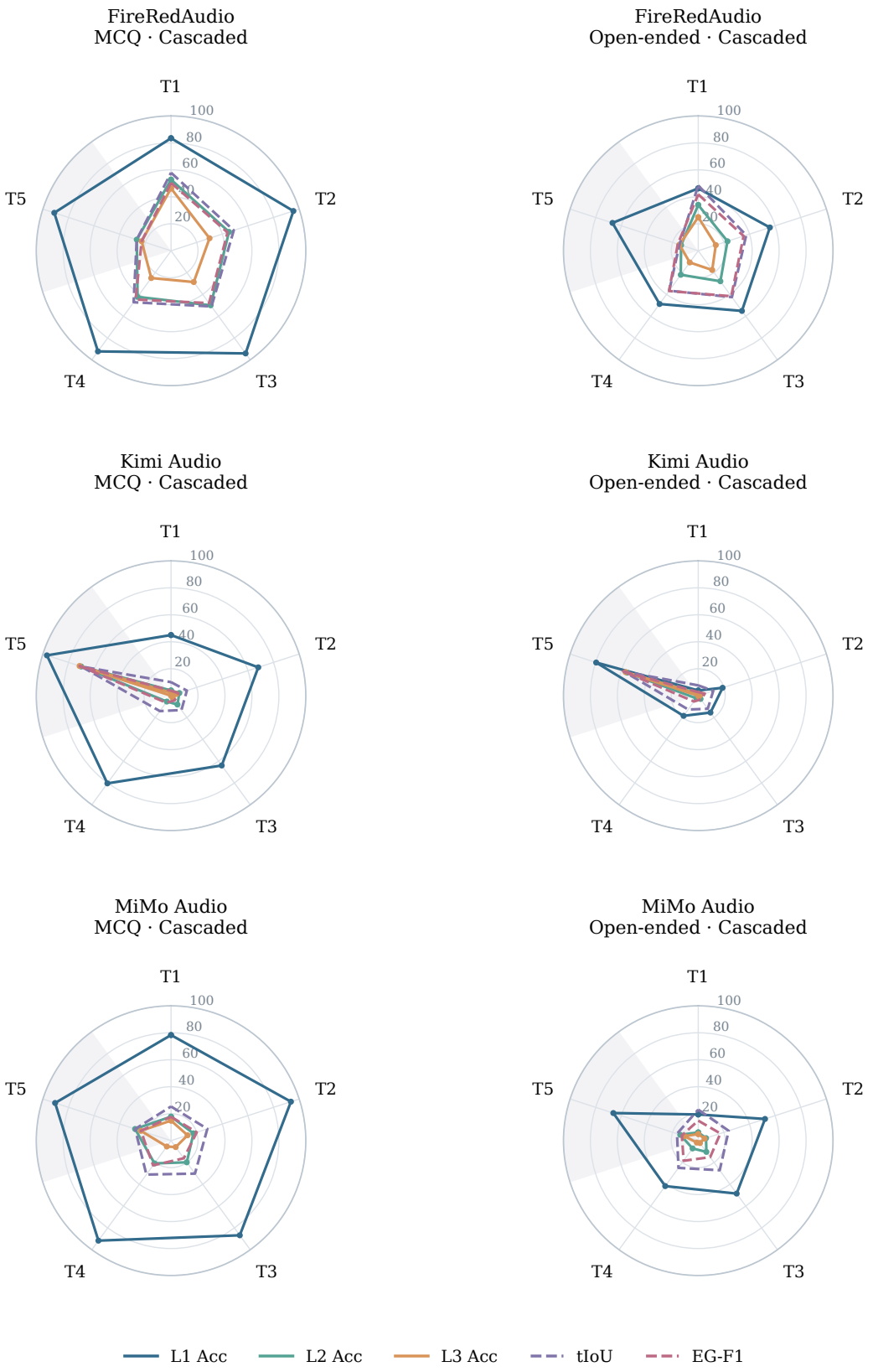


Figure 12: Caption-based cascaded inference task profiles: FireRedAudio, Kimi Audio, MiMo Audio. Each row shows MCQ (left) and open-ended QA (right). All axes use the same 0–100 scale; tIoU and EG-F1 are multiplied by 100. Solid lines show L1–L3 accuracies, dashed lines grounding means. The shaded T5 sector measures evidence abstention for grounding scores.

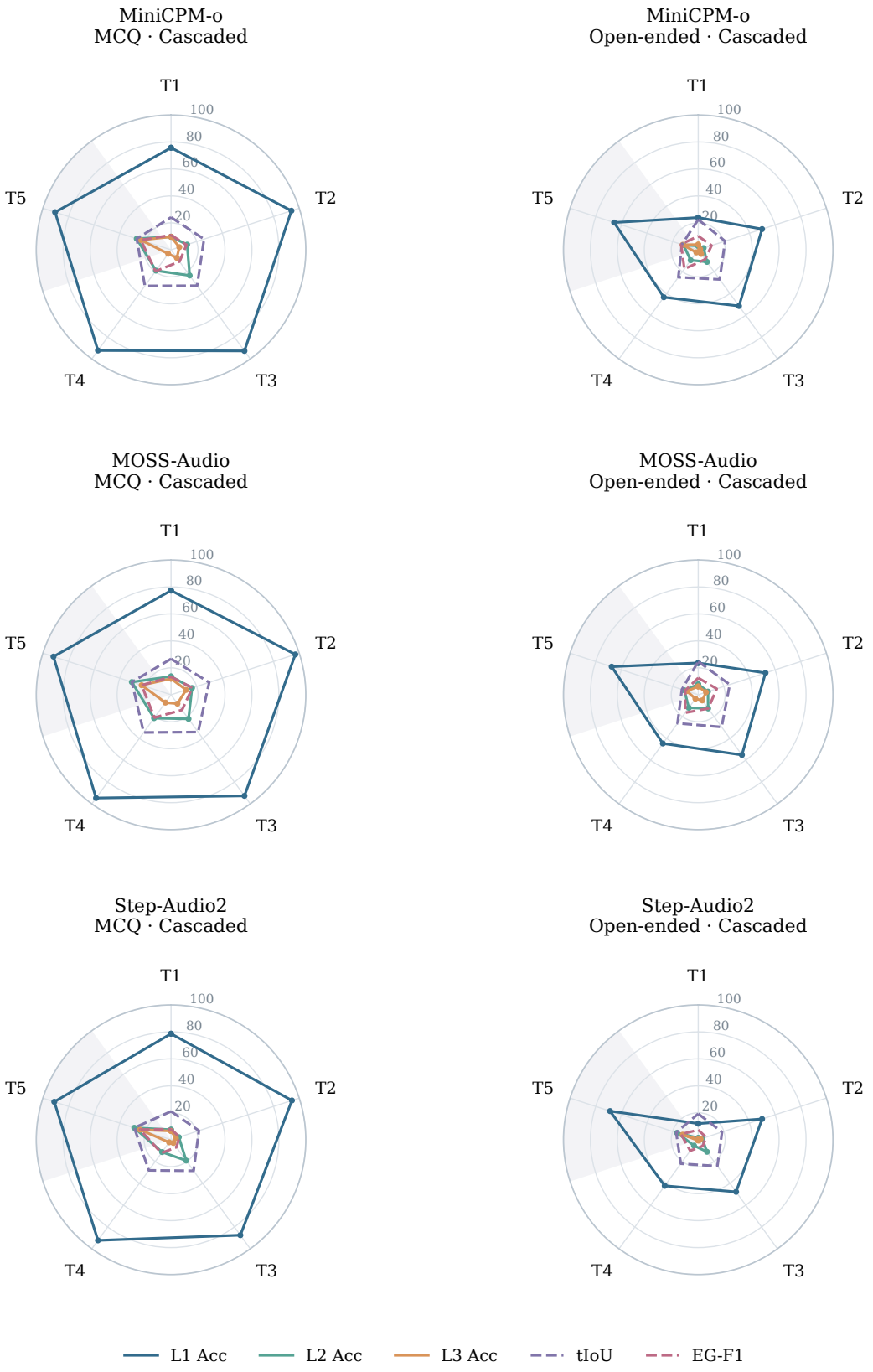


Figure 13: Caption-based cascaded inference task profiles: MiniCPM-o, MOSS-Audio, Step-Audio2. Each row shows MCQ (left) and open-ended QA (right). All axes use the same 0–100 scale; tIoU and EG-F1 are multiplied by 100. Solid lines show L1–L3 accuracies, dashed lines grounding means. The shaded T5 sector measures evidence abstention for grounding scores.

D.6 Qualitative Grounding Failures

We analyze temporal localization and event-level grounding separately, with four tIoU cases and three EG-F1 cases numbered independently within each group. All selected MCQ answers are correct, allowing evidence quality to be examined without an answer-judge decision. Each case includes the question, correct answer, predicted answer, and every reference and predicted evidence event; timestamps are in seconds. Event IDs and speaker metadata are omitted, but no evidence event is removed. These are annotation-based qualitative examples, not mutually exclusive error classes or estimates of prevalence.

D.6.1 Typical tIoU Errors

Tables 17–20 illustrate four patterns: disjoint localization, overly broad boundaries, duration undercoverage, and incomplete multi-event coverage. All use L2 responses, which request intervals without evidence content.

Low tIoU distinguishes missing the region, imprecise boundaries, and incomplete coverage. tIoU case 1 has no temporal overlap at all. Case 2 instead recovers the entire reference event, but expands a 2.849-second event into a 52.8-second prediction, reducing tIoU to 0.054. Case 3 shows the opposite boundary error: almost all predicted time falls inside the music event (precision 0.992), but the interval stops early and covers only 32.5% of its annotated duration. Case 4 fully covers three reference events but omits the later explanation needed for part of the answer, leaving duration-weighted recall at 0.379. Thus, low overlap can reflect excess context, incomplete duration, or missing links in a multi-event support set, rather than failure to find any relevant audio.

Table 17: tIoU case 1: Correct answer, disjoint evidence interval. FireRedAudio (End-to-end); L2 MCQ; question 000002; audio_000004.

Question	What sound is heard near the end of the audio, around 59 minutes and 55 seconds?
Correct answer	B — A large crowd cheering and applauding enthusiastically
Model answer	B (correct)
Reference evidence	[3592.355, 3598.930] s (sound): A large crowd cheers and applauds enthusiastically.
Predicted evidence	1. [3562.200, 3567.700] s
Scores	tIoU = 0.000; temporal precision = 0.000; recall = 0.000.
Interpretation	The selected option is correct, but the predicted interval does not overlap the annotated crowd event. Answer correctness therefore does not establish temporal grounding.

Table 18: tIoU case 2: Full coverage with overly broad boundaries. Gemini 3.6 Flash (End-to-end); L2 MCQ; question 000015; audio_000017.

Question	Around 68 minutes into the lecture, which two industries does the lecturer give as examples?
Correct answer	C — Oil and airlines
Model answer	C (correct)
Reference evidence	[4081.927, 4084.776] s (speech): Oil is probably the classic example, airlines is a good example, okay?
Predicted evidence	1. [4050.000, 4102.800] s
Scores	tIoU = 0.054; temporal precision = 0.054; recall = 1.000.
Interpretation	The 52.8-second prediction contains the entire 2.849-second reference event: temporal recall is 1, but tIoU is only 0.054. This is imprecise localization, not complete failure to find the relevant region.

Table 19: tIoU case 3: Undercoverage of an event duration. Gemini 3.1 Pro (Cascaded); L2 MCQ; question 000235; audio_000160.

Question	At approximately what time does the suspenseful synthesized music with heavy percussion, low pulses, and sweeping risers play?
Correct answer	B — About 43:17–43:36
Model answer	B (correct)
Reference evidence	[2596.994, 2615.774] s (music): A suspenseful synthesized cue builds energy with heavy percussion, low pulses, shimmering metallic melody, strings, drones, and sweeping risers.
Predicted evidence	1. [2596.945, 2603.090] s
Scores	tIoU = 0.324; temporal precision = 0.992; recall = 0.325.
Interpretation	The prediction begins close to the annotated music onset but ends about 12.7 seconds too early. Temporal precision is 0.992, yet recall is only 0.325 and tIoU is 0.324. It identifies the vicinity of the cue without recovering its annotated duration; this differs from selecting an unrelated region.

Table 20: tIoU case 4: Missing a later component of multi-event support. Gemini 3.1 Pro (End-to-end); L2 MCQ; question 000050; audio_000037.

Question	Around 10:39, after the interviewer asks what needed to be done and what the bill needed to look like, how does the interviewee respond?
Correct answer	B — She directly answers that they reduced the bill's scope, then explains that they narrowed it to essential investments in CHIPS, science, and technology.
Model answer	B (correct)
Reference evidence	[639.290, 640.340] s (speech): And what did you need to do? [640.370, 642.020] s (speech): What did it need to look like? [642.040, 644.620] s (speech): We reduced the scope of the bill. [646.810, 655.460] s (speech): In short, I won't go into the weeds, but we reduced it to the bare essence, of investments in CHIPS and science and technology.
Predicted evidence	1. [639.200, 645.700] s
Scores	tIoU = 0.349; temporal precision = 0.812; recall = 0.379.
Interpretation	The interval covers the two questions and the initial reply about reducing the bill, but excludes the later explanation identifying CHIPS, science, and technology. That omitted event supports a specific component of the selected answer. Three of four reference events are fully covered, but duration-weighted recall is only 0.379 and tIoU is 0.349.

D.6.2 TYPICAL EG-F1 ERRORS

Tables 21–23 illustrate time–content binding errors, excessive redundant information, and acoustic stream mismatch. All use L3 responses and the current same-stream, maximum-cardinality matching rule: pairwise tIoU and semantic similarity must both be strictly greater than 0.5, while L3 requires both tIoU and final EG-F1 strictly above 0.5. The Chinese prediction in EG-F1 case 3 is explicitly translated for display; the original is retained in the accompanying case records.

Low EG-F1 can reflect incorrect evidence binding, redundant information, or event-type mismatch. In EG-F1 case 1, time and content each match somewhere in the prediction set, but never in the same event. Case 2 contains a valid supporting event; two additional background events lower precision to one-third and EG-F1 to 0.5. Case 3 reaches the reference time region (tIoU 0.566), but a mixed sound/background-music description is labeled as `sound` rather than the target `music` event, leaving no same-stream candidate match. These cases distinguish failures of time–content alignment, evidence precision, and event identity, even when the selected answer is correct.

Table 21: EG-F1 case 1: Time and content fail to align within the same event. Gemini 3.6 Flash (Cascaded); L3 MCQ; question 000037; audio_000034.

Question	Near the end of the audio, what kind of tea does the woman say she likes when asked about tea flavor?
Correct answer	B — English breakfast tea
Model answer	B (correct)
Reference evidence	[2504.003, 2505.443] s (speech): I like English breakfast.
Predicted evidence	1. [2504.088, 2505.011] s (speech): What flavor tea? 2. [2505.118, 2509.467] s (speech): I like English Breakfast. I don't like flavored tea at all. I don't drink flowers.
Scores	tIoU = 0.228; event precision = 0.000; recall = 0.000; EG-F1 = 0.000. Gold 1/prediction 1: pairwise tIoU = 0.641, similarity = 0.438; Gold 1/prediction 2: pairwise tIoU = 0.059, similarity = 0.761.
Interpretation	Prediction 1 aligns in time but contains only the question, not the tea preference. Prediction 2 names the correct tea but overlaps too little with its reference interval. No single event meets both matching thresholds; this is a time–content binding error, not necessarily fabricated content.

Table 22: EG-F1 case 2: Excessive redundant information. FireRedAudio (End-to-end); L3 MCQ; question 000053; audio_000039.

Question	Around 35 minutes and 10 seconds into the audio, which airline does the narrator say was the first to agree to their request?
Correct answer	A — Delta
Model answer	A (correct)
Reference evidence	[2109.581, 2112.251] s (speech): The first one to say yes was Delta.
Predicted evidence	1. [2101.600, 2108.800] s (speech): We wanted to see that Swiss watch in action. So, we reached out to some of the big US airlines. 2. [2109.500, 2111.900] s (speech): The first one to say yes was Delta. 3. [2112.200, 2114.600] s (speech): So, we jumped on a flight to Atlanta.
Scores	tIoU = 0.193; event precision = 0.333; recall = 1.000; EG-F1 = 0.500. Gold 1/prediction 1: pairwise tIoU = 0.000, similarity = 0.404; Gold 1/prediction 2: pairwise tIoU = 0.843, similarity = 0.839; Gold 1/prediction 3: pairwise tIoU = 0.010, similarity = 0.459.
Interpretation	Prediction 2 is a valid match. Predictions 1 and 3 add context but do not directly identify the first airline to agree. One match among three predictions gives precision 1/3 and recall 1, hence EG-F1 = 0.5, which fails the strict >0.5 gate. This is insufficient evidence precision, not missing the answer-supporting event.

Table 23: EG-F1 case 3: Acoustic stream mismatch despite temporal coverage. FireRedAudio (Cascaded); L3 MCQ; question 000792; audio_000161.

Question	At approximately what time does a gentle minor-key synth pad with an echoed arpeggiated pluck play?
Correct answer	C — Around 513.6–520.2 seconds
Model answer	C (correct)
Reference evidence	[513.640, 520.200] s (music): A gentle sustained minor-key synth pad accompanies a clean arpeggiated synth pluck with soft echo.
Predicted evidence	1. [510.000, 521.600] s (sound): Kitchen cooking sounds and soft background music. [Translated from Chinese.]
Scores	tIoU = 0.566; event precision = 0.000; recall = 0.000; EG-F1 = 0.000. No same-stream candidate pair.
Interpretation	The interval fully covers the reference music cue and has tIoU 0.566. However, the model labels its mixed cooking-sound/background-music description as sound, while the requested synth event is annotated as music. The same-stream rule therefore admits no candidate pair and EG-F1 is 0. No embedding similarity is assigned to this excluded cross-stream pair. This is an event-type mismatch, not a failure to reach the relevant time region.

E Question-Only Baseline and Shortcut Analysis

We test three distinct explanations for success without audio: selecting among supplied answers, abstaining when the input is missing, and exploiting question–option semantic alignment. We then score the same question-only model's independently generated L2/L3 outputs to distinguish answer selection from submitted evidence support.

E.1 Answer Selection Without Audio

A model baseline beyond construction-time screening. The common-knowledge audit is a data-construction filter, not a measurement of model performance without audio. We evaluate Qwen3.5-27B on all 1,033 paired questions under two L1 settings: MCQ receives the question and four options, whereas open-ended QA receives only the question. Neither receives audio, transcripts, captions, event streams, reference answers, or annotated evidence. We use temperature 0, disable thinking, and allow up to 2,048 output tokens. Both retain their format-specific answer-only instructions, adapted to permit question text and existing knowledge. All Qwen3.5-27B outputs at L1–L3 use the same scoring protocol as the audio systems: exact option matching for MCQ; GPT-5.6 Sol and DeepSeek-V4-Flash applying the same three-dimensional rubric and all-six-scores-equal-5 rule for open-ended answers; and the same tIoU and EG-F1 evidence gates at L2/L3. Qwen3.5-27B does not judge its own answers. All 2,066 L1 requests succeed, MCQ labels are valid, and all open-ended binary verdicts are determinate. We reuse the completed, frozen runs, whose input hashes match the final release.

Without audio, high answer accuracy is concentrated in multiple-choice QA. MCQ reaches 71.0% (733/1,033), compared with 13.6% (141/1,033) for open-ended QA, a gap of 57.3 percentage points. On the 876 answerable questions, the gap is larger: 65.8% (576/876) versus 1.0% (9/876), or 64.7 points. Thus, text-only success in this setting is concentrated in selecting among supplied candidates, rather than independently producing an acceptable answer. This contrast motivates testing question–option cues as a potential source of shortcuts. The difference also includes format-specific instructions and scoring requirements, so it is not a causal estimate of the effect of options alone.

T2–T4 permit substantial text-only success, potentially aided by semantic option cues. MCQ versus open-ended accuracy is 86.3% versus 1.9% for T2, 75.0% versus 2.2% for T3, and 86.9% versus 0.6% for T4. One possible explanation is that correct candidates share semantic content with the question, allowing models to exploit question–option relations without recovering the recording's facts. The similarity analysis below tests this potential cue separately.

T1 offers fewer content-based option cues. All 309 T1 questions present their options as timestamps or temporal intervals, predominantly the latter. Unlike descriptive candidates in T2–T4, these options provide little factual content that can be matched semantically to the question. T1's question-only MCQ accuracy is 32.7%, only 7.7 percentage points above the 25% random-choice expectation and substantially below T2–T4. This indicates relatively limited text-only solvability, rather than the complete absence of shortcuts.

Table 24: Question-only L1 performance of Qwen3.5-27B. MCQ and Open cells show correct counts and accuracy; the gap is MCQ minus Open in percentage points, computed before rounding. Uniform random selection has expected accuracy 25% for MCQ only. The first three rows report the full dataset and its answerability split; the remaining rows break down the same questions by category. T5 is the unanswerable subset.

Subset	Questions	MCQ	Open	Gap (pp)
All questions	1033	733 (71.0%)	141 (13.6%)	57.3
Answerable	876	576 (65.8%)	9 (1.0%)	64.7
Unanswerable	157	157 (100.0%)	132 (84.1%)	15.9
T1: Event localization	309	101 (32.7%)	0 (0.0%)	32.7
T2: Detail understanding	263	227 (86.3%)	5 (1.9%)	84.4
T3: Context linking	136	102 (75.0%)	3 (2.2%)	72.8
T4: Information aggregation	168	146 (86.9%)	1 (0.6%)	86.3
T5: Answerability judgment	157	157 (100.0%)	132 (84.1%)	15.9

Table 25: Paired abstention-option probe on 40 answerable questions without audio. Each cell reports a count and percentage. The final column is selection of the newly added E option, not answer accuracy.

Subset	N	Original correct	Added-E correct	Selected E
All	40	26 (65.0%)	5 (12.5%)	35 (87.5%)
T1	10	3 (30.0%)	0 (0.0%)	10 (100.0%)
T2	10	8 (80.0%)	1 (10.0%)	9 (90.0%)
T3	10	7 (70.0%)	1 (10.0%)	9 (90.0%)
T4	10	8 (80.0%)	3 (30.0%)	7 (70.0%)

E.2 Abstention Under Missing Audio

T5 success need not indicate verified answerability judgment. For T5, MCQ accuracy is 100.0% (157/157) and open-ended accuracy is 84.1% (132/157). The format-specific instructions permit selecting an indeterminate option or explicitly stating insufficient information when the requested information is unavailable. With no recording supplied, abstaining is a reasonable response to unavailable input. A high T5 score is therefore consistent with this behavior, without demonstrating that the model has inspected a recording and established that it lacks the answer. Reporting the answerable subset separately avoids relying on this effect to establish text-only solvability.

Adding an abstention option to answerable questions tests the T5 explanation. We randomly sample ten answerable questions from each of T1–T4 (40 total; seed 20260925). For each, we retain all four original options and the reference answer, and append option E: "The audio does not provide enough information to determine the answer" (with equivalent Chinese wording). We rerun both the original four-option and augmented five-option versions on the same Qwen3.5-27B service without audio, using the settings above. Only the added option and the permitted output-label list change; the original instruction to select an indeterminate option when information is absent remains in both conditions. All 80 calls return valid labels. Correctness remains defined against the original audio-based reference, for which the new option is not correct.

The model frequently selects "no information" even for questions with valid audio answers. With the added option, Qwen3.5-27B selects E on 35/40 questions (87.5%); 21 of the 26 originally correct responses switch to E (80.8%). Accuracy against the original reference falls from 65.0% to 12.5%. Thus, the model often responds to unavailable audio by selecting the explicit abstention option, even when the underlying recording contains an answer. This supports the interpretation that question-only T5 success can arise from reasonable refusal under missing input, rather than verification that the recording itself lacks the requested information. This small probe does not isolate the effects of option position or increasing the option count, and does not establish incorrect annotation or the prevalence of shortcuts during audio-based inference.

E.3 Semantic Option Cues

Measuring question–option semantic alignment. We independently encode each question and each option's text using Qwen3-Embedding-0.6B, without option labels, audio, reference evidence, or task instructions. We use the normalized last-token embeddings and their cosine similarity; no text is truncated. For each question, we compare the correct option's similarity with both the mean

Table 26: Question–option cosine similarity. Gold, mean distractor, and max distractor are averaged over questions. The final two columns report how often the correct option exceeds the distractor mean or every distractor, respectively.

Subset	N	Gold	Mean distractor	Max distractor	Gold > mean	Gold top-1
All	1033	0.480	0.466	0.504	57.7%	31.3%
Answerable	876	0.484	0.467	0.507	59.9%	29.8%
T1	309	0.399	0.400	0.414	46.6%	21.7%
T2	263	0.489	0.464	0.516	68.1%	36.9%
T3	136	0.523	0.499	0.559	58.8%	27.2%
T4	168	0.601	0.569	0.621	72.6%	35.7%
T5	157	0.456	0.463	0.491	45.2%	39.5%

and maximum similarity of the three distractors. The reference label identifies the correct option only after encoding. Table 26 reports these means, the proportion of questions whose correct option exceeds the mean distractor, and the proportion for which it strictly exceeds every distractor. Selecting the highest-similarity option gives a simple text-only baseline; there are no tied maxima in this run.

T2 and T4 show a semantic bias consistent with option-based shortcuts. For detail understanding (T2), the correct option's mean similarity is 0.489, compared with 0.464 for the mean distractor; the correct option exceeds the distractor mean on 68.1% of questions. For information aggregation (T4), the corresponding values are 0.601 versus 0.569 and 72.6%. Together with the large MCQ–open-ended gaps, these patterns suggest that question–option semantic alignment can provide useful answer cues in these categories. T3 shows a smaller positive bias (0.523 versus 0.499; 58.8%). T1 shows no average advantage (0.399 versus 0.400), consistent with its temporal-interval options and weaker text-only MCQ performance. T5 also lacks this average advantage (0.456 versus 0.463); together with the abstention probe, this favors treating its high score separately from semantic option cues in T2–T4.

Semantic similarity is one potential cue, not a complete explanation. The correct option has the highest similarity on 36.9% of T2 and 35.7% of T4 questions, well below Qwen3.5-27B's MCQ accuracies of 86.3% and 86.9%. On answerable questions that the model answers correctly without audio, the correct option is highest in similarity in 33.2% of cases. Thus, the experiment supports a category-specific semantic bias, but does not establish that selecting the most similar option accounts for most text-only successes. Other option relations, wording cues, and prior knowledge may also contribute. The combined evidence motivates open-ended evaluation and explicit evidence verification: selecting among supplied candidates can remain easy even when independently producing the answer is difficult. These text-only results are reported separately from the 17 audio-system configurations; comparisons with those systems are not matched audio ablations and do not quantify their actual shortcut use.

E.4 Question-Only Outputs Under Evidence Requirements

Protocol and scope. The preceding baseline table reports L1 only. Here we additionally score the frozen run's four evidence-output modes: MCQ and open-ended QA at L2 and L3, each with 1,033 independently generated responses. Sampling settings and inputs are unchanged. L2 requests intervals and L3 requests event descriptions as well. All 4,132 outputs are valid JSON and contain empty evidence arrays. One L3 MCQ answer is not a valid option label and is counted as incorrect. Open-ended answers use the same six-scores-of-5 rule as the main results; all verdicts are determinate.

Correct selection without audio does not provide verifiable support. On answerable questions, MCQ pure answer accuracy remains high at L2 (64.3%) and L3 (60.4%), compared with 65.8% at L1, yet joint accuracy is zero at both evidence-requiring levels (Table 27). This contrast also holds within each of T1–T4: none of the submitted responses provides supporting evidence. For empty predictions, both tIoU and EG-F1 are exactly zero on answerable questions and one on T5. Consequently, overall MCQ joint accuracy of 15.2% reflects only the 157 unanswerable questions; open-ended joint accuracy of 14.8% and 15.1% likewise reflects successful T5 abstention, not recovery of audio facts.

Table 27: Question-only accuracy (%) overall and by category. Each cell reports MCQ / Open. L1 scores answers alone; L2 and L3 apply the main joint acceptance rules to their independently generated outputs. All evidence predictions are empty, so positive L2/L3 scores come exclusively from T5.

Subset	N	L1 Acc	L2 Acc	L3 Acc
All	1033	71.0 / 13.6	15.2 / 14.8	15.2 / 15.1
Answerable	876	65.8 / 1.0	0.0 / 0.0	0.0 / 0.0
T1	309	32.7 / 0.0	0.0 / 0.0	0.0 / 0.0
T2	263	86.3 / 1.9	0.0 / 0.0	0.0 / 0.0
T3	136	75.0 / 2.2	0.0 / 0.0	0.0 / 0.0
T4	168	86.9 / 0.6	0.0 / 0.0	0.0 / 0.0
T5	157	100.0 / 84.1	100.0 / 97.5	100.0 / 99.4

Answer-level shortcuts yield no evidence-verified successes in this no-audio evaluation. Across T1–T4, neither MCQ nor open-ended responses pass the joint L2 or L3 criteria. Thus, the substantial text-only answer accuracy observed here does not translate into successful temporal or content evidence verification. In this no-audio setting, L2/L3 distinguish unsupported answer selection from evidence-supported correctness. Together with the fixed-response analysis in Appendix D.2, these results support evaluating submitted evidence rather than relying on answer accuracy alone.

F Model and API Versions

Table 28 lists the model versions used in the experiments and supporting evaluation. Model names are reported without service endpoints. For hosted models, the names identify the versions reported for the experiments, not immutable dated snapshots. For local models, checkpoint names are given where available. The answer judges specified in the main evaluation protocol are distinguished from auxiliary judges used in the current diagnostic runs.

Table 28: Model versions and their roles in AudioTrace. The model names used in the main results are retained in the first column.

Model	**Version / checkpoint**	**Role**
Gemini	Gemini 3.1 Pro	End-to-end; captioner
Gemini	Gemini 3.6 Flash	End-to-end; captioner
Qwen Omni	Qwen3.5 Omni Plus	End-to-end; captioner
Qwen Omni	Qwen3.8 Omni Flash	End-to-end; captioner
FireRedAudio	FireRedAudio	End-to-end; captioner
Qwen3 Omni	Qwen3-Omni-30B-A3B-Captioner	Captioner
Audio Flamingo	Audio-Flamingo-Next-Instruct	Captioner
Kimi Audio	Kimi-Audio-7B-Instruct	Captioner
MiMo Audio	MiMo-Audio-7B-Instruct	Captioner
MiniCPM-o	MiniCPM-o 4.5	Captioner
MOSS-Audio	MOSS-Audio-8B-Instruct	Captioner
Step-Audio2	Step-Audio-2-mini	Captioner
Claude	Claude Opus 4.8	Fixed text reasoner
GPT	GPT-5.6 Sol	Answer judge; QA construction
DeepSeek	DeepSeek-V4-Flash	Answer judge
Qwen	Qwen3.5-27B	Text-only baseline; auxiliary judge
Qwen Embedding	Qwen3-Embedding-0.6B	EG-F1 text encoder

Thinking is enabled during question generation and model-based review, but disabled for all benchmark inference settings and both answer judges. GPT-5.6 Sol and DeepSeek-V4-Flash define the two-judge protocol in Section 4.1; Qwen3.5-27B is listed separately for reproducibility of auxiliary evaluations and is not included in the reported two-judge results.