LAMAR-BENCH: EVALUATING LONG-FORM AND MULTI-AUDIO UNDERSTANDING AND REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-language models are increasingly able to process long recordings and multiple related audios. Existing benchmarks capture important parts of these settings, but the evaluation landscape remains fragmented, with fine-grained paralinguistic and speech-artifact analysis, long-range semantic dependencies, detailed audio-version comparison, and goal-conditioned decisions still underexplored. We introduce LAMAR-Bench, a capability-based benchmark that organizes long-form and multi-audio evaluation into three complementary families: fine-grained understanding, cross-audio relation reasoning, and decisions under goals and constraints. It contains 1,597 questions across 10 tasks, constructed from shared audio facts with traceable supervision, with up to 19 audio clips and 59.9 minutes of input per question. Evaluating 14 audio-language and omni-modal systems reveals substantial room for improvement: the best paralinguistic-analysis and acoustic-localization accuracies reach only 51.09% and 54.38%, respectively, while detailed cross-audio difference analysis remains challenging. We further conduct paired analyses over answer format, presentation order, and requested goals, revealing limitations in direct localization, order robustness, and goal adaptation. These findings highlight persistent gaps in evidence-grounded understanding and reliable goal-conditioned reasoning over long and multiple audio inputs.

1 INTRODUCTION

Large audio-language and omni-modal models are increasingly able to process long recordings and multiple audio inputs (Xu et al., 2025; Zhou et al., 2026; Wang et al., 2026). This capability enables audio workflows that go beyond understanding an isolated short clip. For example, meeting analysis may require locating a particular discussion, identifying who said what, and connecting information across distant parts of a recording. Podcast and content editing may require comparing an original recording with alternative versions, identifying what has changed, and selecting or verifying an outcome according to requirements such as information preservation, acoustic continuity, or duration. Such workflows require models not only to recover evidence from audio, but also to relate evidence across time or recordings and use it to make decisions under explicit goals or constraints.

Existing benchmarks cover important aspects of these demands, including diverse auditory skills (Sakshi et al., 2025; Ma et al., 2026), long-context understanding (Ahia et al., 2025; Huang et al., 2026; Dai et al., 2026), and reasoning, comparison, or retrieval over multiple audio inputs (Kumar et al., 2026; Yang et al., 2026b; Li & Lee, 2026). However, the evaluation landscape remains fragmented across capabilities and evidence requirements. Long-form benchmarks mainly focus on understanding and locating information within a recording, while multi-audio benchmarks typically target specific forms of matching, comparison, reasoning, or retrieval. Less attention has been given to fine-grained paralinguistic and speech-artifact analysis, long-range dependencies, detailed changes across related audio versions, and decisions based on explicit goals or multiple requirements. This fragmentation makes it difficult to assess, within a unified framework, what models can accomplish over long and multiple audio inputs.

We therefore organize the problem around three capability questions. First, can a model recover fine-grained semantic and acoustic evidence from a long recording, including speaker and paralinguistic properties, speech artifacts, repeated occurrences, and information that must be connected across distant parts of the input? Second, when multiple related audio materials are presented, can

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

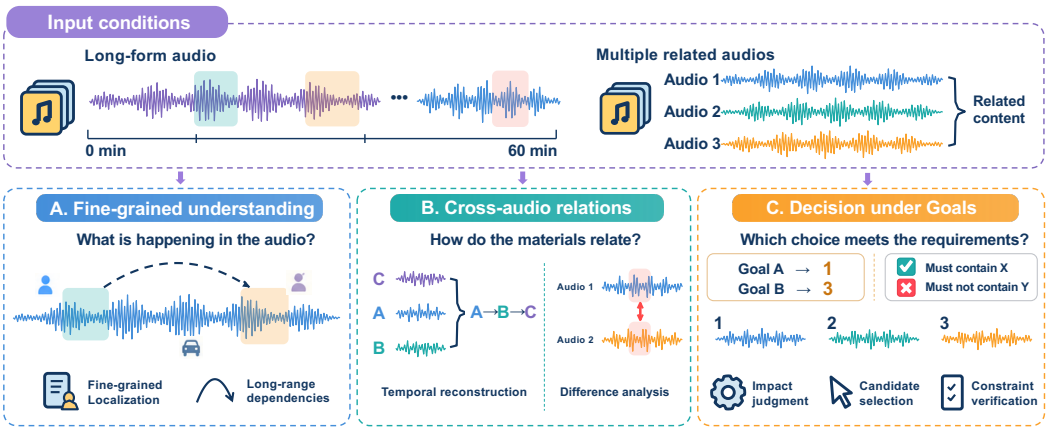

Figure 1: LAMAR-Bench organizes long-form and multi-audio evaluation into three complementary capability families: fine-grained understanding (A), cross-audio relation reasoning (B), and decisions under goals and constraints (C).

the model determine how they relate, for example by reconstructing their temporal organization or identifying where and how different versions change? Third, can the model assess audio candidates or outcomes according to an explicit goal or a set of constraints?

To address these questions, we introduce LAMAR-Bench (**L**ong-form **A**udio and **M**ulti-**A**udio **R**easoning benchmark), comprising 1,597 questions across 10 tasks and three capability families (Figure 1): *fine-grained understanding*, covering speaker and paralinguistic analysis, acoustic and speech-artifact localization, content retrieval, and long-range semantic dependencies; *cross-audio relation reasoning*, covering temporal reconstruction and detailed differences among related audios; and *decisions under goals and constraints*, covering impact judgment, candidate selection, and constraint verification. Questions contain 1–19 audio clips and up to 59.9 minutes of input, and are constructed from shared audio facts with traceable supervision. Evaluating 14 audio-language and omni-modal systems with task-specific metrics reveals substantial room for improvement: the best paralinguistic-analysis and acoustic-localization accuracies reach only 51.09% and 54.38%, respectively, while audio difference analysis remains challenging. We further conduct paired analyses of answer format, presentation order, and requested goals, revealing recognition-to-localization gaps, order-related selection changes, and failures to adapt across goals.

Our contributions are:

- We introduce LAMAR-Bench, a capability-based benchmark that jointly evaluates fine-grained understanding, cross-audio relation reasoning, and decisions under goals and constraints across long-form and multi-audio inputs.
- We construct a 1,597-question suite spanning 10 tasks, including fine-grained paralinguistic analysis, speech-artifact localization, long-range semantic dependencies, audio-version difference analysis, and goal- and constraint-conditioned decisions, using shared audio facts and task-specific derivations with traceable supervision.
- We evaluate 14 audio-language and omni-modal systems with task-specific metrics and textual reference baselines, and complement the main evaluation with paired analyses that expose limitations in direct localization, order robustness, and adaptation to changed goals.

2 Related Work

Table 1 compares representative benchmarks by input conditions and the capabilities they explicitly evaluate. Rather than grouping benchmarks only by input format, we organize prior work around the three capability families studied in LAMAR-Bench.

Table 1: Comparison of input conditions and capabilities across audio benchmarks. Duration is the mean summed audio input per instance (minutes). ✓: explicitly evaluated; –: not explicitly evaluated; N/R: not reported; ∼: approximate; pool: retrieval corpus.

Benchmark	Input		Audio Understanding		Cross-audio Reasoning		Goal-based Decisions	
	Duration	Clips	Fine-grained Localization	Long-range Dependencies	Difference Comparison	Temporal Reconstruction	Goal-based Selection	Constraint Verification
MMAU (Sakshi et al., 2025)	0.17	1	–	–	–	–	–	–
BLAB (Ahia et al., 2025)	51.0	1	✓	✓	–	–	–	–
MMAR (Ma et al., 2026)	0.33	1	–	–	–	–	–	–
MMAU-Pro (Kumar et al., 2026)	N/R	1–3	–	✓	✓	–	–	–
AudioMarathon (He et al., 2025)	3.55	1	–	✓	–	–	–	–
MUGEN (Yang et al., 2026b)	0.76	5–6	–	–	✓	–	✓	–
INSPIRE (Li & Lee, 2026)	N/R	1+pool	–	–	–	–	✓	–
LongAudioSpan (Huang et al., 2026)	∼50	1	✓	✓	–	–	–	–
TAG-Bench (Dai et al., 2026)	5.13	1	✓	–	–	–	–	–
LAMAR-Bench	**37.41**	**1–19**	✓	✓	✓	✓	✓	✓

Fine-grained audio understanding. MMAU (Sakshi et al., 2025) and MMAR (Ma et al., 2026) evaluate broad speech, sound, and music understanding, while LongAudioBench (Ghosh et al., 2025), BLAB (Ahia et al., 2025), and AudioMarathon (He et al., 2025) extend evaluation to long recordings. LongAudioSpan (Huang et al., 2026) and TAG-Bench (Dai et al., 2026) further emphasize temporal grounding. Together, these benchmarks establish the difficulty of long-context comprehension and localization, but fine-grained evidence requirements such as paralinguistic ordering, speech-artifact localization, complete occurrence retrieval, and dependencies across distant utterances remain less systematically studied. LAMAR-Bench evaluates these requirements jointly within long recordings.

Cross-audio relation reasoning. MAE (Chen et al., 2024), MMAU-Pro (Kumar et al., 2026), and MUGEN (Yang et al., 2026b) evaluate reasoning across multiple audio inputs, including matching, attribute comparison, and question answering; MUGEN further studies performance as the number of candidates increases. These tasks establish the importance of multi-audio reasoning, but they provide more limited coverage of relations that require reconstructing a global temporal organization or characterizing localized changes between related audio versions. LAMAR-Bench explicitly evaluates both temporal order reconstruction and detailed difference analysis.

Goal-conditioned audio decisions. MUGEN (Yang et al., 2026b) and INSPIRE (Li & Lee, 2026) evaluate criterion-dependent audio selection and retrieval, with INSPIRE varying the retrieval goal for the same query and MUGEN studying presentation-order sensitivity. MMAU-Pro additionally tests output constraints such as response length and format. These works advance goal-sensitive evaluation, but offer limited assessment of whether modifications preserve desired properties, whether candidate audios satisfy multiple content or acoustic requirements, and whether models adapt their choices when the requested goal changes. LAMAR-Bench broadens this setting to impact judgment, candidate selection, and constraint verification.

3 LAMAR-BENCH

3.1 OVERVIEW

LAMAR-Bench evaluates audio understanding and reasoning under long-form and multi-audio input conditions through 1,597 questions across 10 tasks, spanning 935 distinct source materials and 606.7 hours of audio. Each question presents 1–19 clips. Table 2 groups the tasks into three capability families according to their evidence requirements and prediction targets. Audio understanding tasks (1,135 questions) assess fine-grained localization, complete occurrence retrieval, and dependencies between distant utterances. Cross-audio reasoning tasks (178 questions) require reconstructing segment order and identifying version differences, including acoustic changes that preserve the words. Goal-based decision tasks (284 questions) assess impact judgment, candidate selection, and constraint satisfaction. Appendix A.2 provides examples of the tasks.

Table 2: Task suite and prediction targets across three capability families: audio understanding, cross-audio reasoning, and goal-based decisions.

Task	Prediction target
A: Audio understanding	
A1 Speaker analysis	Speaker counts, identities, locations, attributions, and roles
A2 Paralinguistic analysis	Rate- and pause-based spans, and speaking-rate order
A3 Acoustic localization	Acoustic event and artifact locations
A4 Content retrieval	Requested content locations
A5 Long-range semantic reasoning	Locations of semantically linked utterances
B: Cross-audio reasoning	
B1 Temporal order reconstruction	Original order of audio segments
B2 Audio difference analysis	Audio differences with locations, types, classes, and durations
C: Goal-based decisions	
C1 Impact judgment	C1a: Information loss and semantic disruption C1b: Preferred splice based on acoustic continuity
C2 Candidate selection	Best-matching candidate for the stated goal and requirements
C3 Constraint verification	Compliance verdict for each requirement

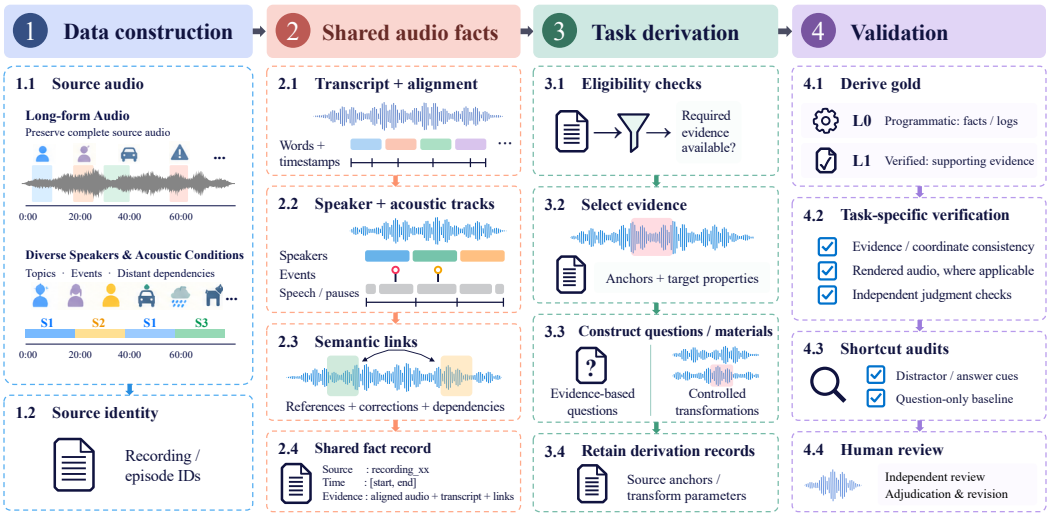

Figure 2: Construction pipeline of LAMAR-Bench. Shared facts extracted from source recordings support task derivation, followed by verification and human review.

3.2 Dataset construction

We construct the task suite from shared audio facts, then verify the resulting questions and reference answers through task-specific checks and human review.

Construction pipeline. The pipeline in Figure 2 combines basic audio parsing with shared derived annotations to organize each recording into four complementary components: transcripts and alignment, speaker information, acoustic information, and semantic links. Task constructors reuse these annotations and select recordings with suitable evidence, avoiding repeated full-recording parsing for each task. For example, reliable speaker turns support speaker localization, while verified links between distant utterances support long-range semantic reasoning. Each constructor selects evidence anchors and derives questions or controlled audio transformations. Gold is computed from source facts or transformation logs (L0), or obtained through independently verified judgments (L1); these labels distinguish derivation routes, not quality levels. Appendix A.3 describes the four components and illustrates task routing with two selected examples.

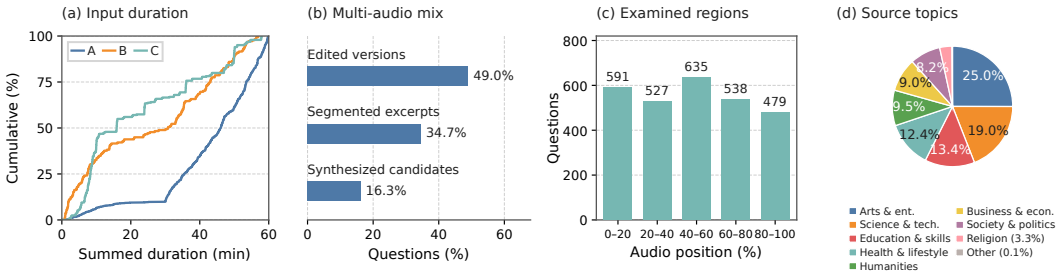

Figure 3: Dataset profile: (a) summed input duration; (b) multi-audio composition; (c) examined regions and search scopes; (d) automatically labeled source topics.

Quality control. Quality control combines independent human review, adjudication and manual revision, and a final independent spot check. Human review assesses media match, question clarity, evidence sufficiency, answer correctness and uniqueness, temporal accuracy, and target completeness. Our protocol calls for two independent reviewers per question and listening judgments before consulting reference answers. Flagged issues and disagreements underwent third-party adjudication, followed by the required revisions and a separate recheck of revised items. All required revisions and final checks have been completed. Appendix A.4 describes this process and reports pre-adjudication reviewer agreement rates, with 93.04% pooled agreement across the seven review dimensions.

3.3 Dataset Analysis

We examine the input demands represented in the benchmark and the breadth of its evaluation coverage.

Complementary input demands. The task suite combines extensive audio inputs with comparisons among closely related recordings. Figure 3(a) shows that long inputs are routine: 75.7% of questions exceed 30 minutes of summed audio, with extended inputs in all three capability families. Within the multi-audio subset shown in Figure 3(b), edited versions account for nearly half of the questions, alongside same-source excerpts and synthesized candidates. This composition gives substantial weight to distinguishing changes within shared content and assessing their consequences, supporting both comparison of existing materials and evaluation of alternative outcomes under explicit requirements.

Broad evaluation coverage. These demands are examined across varied temporal positions and subject matter. Figure 3(c) shows that examined regions and search scopes reach the beginning, middle, and end of recordings, with comparable question counts across the five position bins. This supports fine-grained evaluation across the input timeline, including its later portions. In Figure 3(d), the largest source topic accounts for one quarter of the collection; the remainder spans science, education, health, humanities, and other subjects. This adds topical breadth within the speech-centered collection. Appendix A.1 details the position-counting convention and automated topic annotation.

3.4 Evaluation Metrics

We use task-specific metrics to reflect distinct prediction targets and aggregate scores within each capability family to summarize performance.

Task metrics. Understanding tasks (A1–A5) use multiple-choice accuracy. Temporal order reconstruction (B1) measures ordering agreement with Kendall's τ; audio difference analysis (B2) uses F1 to evaluate the detection and characterization of differences between audio versions. Table 3 lists these metrics and the correctness requirements for decision tasks (C1–C3). Detailed scoring rules and supplementary diagnostic metrics appear in Appendices B.1 and B.2, respectively.

Table 3: Task metrics used in the main evaluation.

Task	Metric	Scoring rule
A1–A5	Accuracy	Correct multiple-choice option.
B1	Mean Kendall's τ	Pairwise ordering agreement.
B2	Mean micro F1	Pool annotated differences across questions; average F1 over matching onset, then additionally object class, difference type, and duration.
C1a	Macro F1	Classify semantic integrity as safe or unsafe.
C1b	Both-order accuracy	Select the correct audio in both the original and reversed input orders.
C2	Selection accuracy	Correct candidate for the stated goal.
C3	All-constraint accuracy	Every constraint verdict in a question is correct.

Score aggregation. Each family **Avg** equally averages its task scores: A1–A5, B1–B2, or C1–C3. C1 first averages semantic integrity (C1a) and acoustic continuity (C1b). Scores are multiplied by 100. We compare models within each family and report no overall score because metrics and chance levels differ across families.

4 Experiments

4.1 Experimental setup

To characterize model capabilities in long-form and multi-audio settings, we evaluate 14 audio-language and omni-modal systems spanning hosted and open-source model families, alongside textual reference baselines.

Evaluated models. The eight hosted systems are Gemini-3.1-Pro and Gemini-3.5/3.6/3.7/3.8-Flash[1], Qwen3.5-Omni-plus (Team, 2026), Doubao-Seed 2.0 Lite[2], and Muse Spark 1.2[3]. The six open-source systems are Voxtral-Mini-3B and Voxtral-Small-24B (Liu et al., 2025), Baichuan-Omni-1.5 (Li et al., 2025), MOSS-Audio-8B-Instruct (Yang et al., 2026a), Phi-4-Multimodal-Instruct (Abouelenin et al., 2025), and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025). We construct two textual baselines using GPT-5.6-Sol[4]: *question-only*, which measures performance supported by question cues, and *transcript-only*, which additionally provides transcript words to assess performance with access to linguistic content.

Evaluation protocol. All systems are evaluated on the complete set of 1,597 questions, with 49 additional acoustic continuity requests using reversed input order. Section 4.5 adds fresh repeated decision diagnostics for four hosted systems. Audio inputs retain their full duration; Doubao and Muse use matched MP3 conversion (†) when required by service file-size limits. Appendix B.3 details checkpoint identities, inference settings, audio adaptation, and transcript construction.

4.2 Performance across capability families

Table 4 summarizes performance across the three capability families. Gemini-3.7/3.8-Flash obtain the two highest family means throughout, although the leading model varies by task. Hosted systems consistently outperform the evaluated open-source models on family means and exceed the question-only baseline in all three families. Open-source understanding scores are close to the 25.00 uniform-choice reference. These results indicate a substantial performance gap between the evaluated model groups under the tested inference settings, while the task-level results reveal distinct challenges within each capability family.

[1] Gemini model documentation: `https://ai.google.dev/gemini-api/docs/models`.

[2] `https://ark.volcengine.com/region:cn-beijing/model/detail?Id=doubao-seed-2-0-lite`

[3] `https://research.meta.ai/blog/introducing-muse-code-and-muse-spark-1-2`

[4] `https://developers.openai.com/api/docs/models/gpt-5.6-sol`

Table 4: Main results on LAMAR-Bench. Bold: maximum observed audio-model scores.

	A: Understanding						B: Relations			C: Decisions				
System	A1	A2	A3	A4	A5	**Avg**	B1	B2	**Avg**	C1a	C1b	C2	C3	**Avg**
Baselines														
GPT-5.6-Sol (transcript)	70.91	27.51	32.50	52.68	53.76	47.47	59.63	1.55	30.59	88.37	12.24	64.89	45.92	53.71
GPT-5.6-Sol (question-only)	38.55	21.83	33.75	29.19	24.28	29.52	-2.13	0.00	-1.07	30.65	20.41	22.34	36.73	28.20
Random	25.00	25.00	25.00	25.00	25.00	25.00	0.00	—	—	49.70	25.00	23.19	16.76	25.77
Open-source audio models														
Voxtral-Mini-3B	30.91	22.71	24.38	25.50	26.01	25.90	1.07	0.50	0.78	44.16	0.00	35.11	11.22	22.80
Voxtral-Small-24B	24.73	24.89	23.13	23.83	28.32	24.98	4.08	0.00	2.04	52.22	2.04	38.30	15.31	26.91
Baichuan-Omni-1.5	28.73	24.02	26.25	26.51	29.48	27.00	-1.16	0.00	-0.58	43.91	4.08	23.40	28.57	25.32
MOSS-Audio-8B-Instruct	34.91	26.64	29.38	24.50	31.79	29.44	-2.35	0.00	-1.18	46.40	0.00	26.60	26.53	25.44
Phi-4-Multimodal	22.18	21.40	23.75	26.51	29.48	24.66	-2.97	0.00	-1.48	45.25	0.00	25.53	7.14	18.43
Qwen3-Omni-Instruct	45.09	19.65	18.75	27.85	28.32	27.93	3.93	0.18	2.06	39.50	8.16	35.11	26.53	28.49
Hosted audio models														
Gemini-3.1-Pro	56.36	31.00	31.88	51.34	71.10	48.34	41.85	2.44	22.15	56.08	30.61	76.60	31.63	50.52
Gemini-3.5-Flash	58.18	24.45	29.38	41.61	68.21	44.37	42.33	4.57	23.45	42.07	34.69	67.02	32.65	46.02
Gemini-3.6-Flash	63.27	27.51	31.88	57.72	70.52	50.18	51.52	15.16	33.34	65.04	30.61	68.09	46.94	54.28
Gemini-3.7-Flash	85.09	49.78	**54.38**	**83.56**	84.39	**71.44**	62.54	**36.34**	**49.44**	81.39	**61.22**	77.66	**66.33**	**71.76**
Gemini-3.8-Flash	**86.18**	**51.09**	52.50	80.54	**84.97**	71.06	**64.88**	26.18	45.53	**83.58**	53.06	**80.85**	64.29	71.15
Qwen3.5-Omni-plus	70.55	48.47	50.63	74.16	76.88	64.14	26.76	16.52	21.64	46.43	40.82	72.34	46.94	54.30
Doubao-Seed 2.0 Lite†	55.27	27.51	38.75	46.98	48.55	43.41	11.77	6.78	9.27	52.07	51.02	65.96	57.14	58.21
Muse Spark 1.2†	46.55	21.40	31.88	36.58	34.10	34.10	23.81	0.46	12.14	53.49	16.33	46.81	18.37	33.36

Within fine-grained understanding, performance is strongest on speaker analysis, content retrieval, and long-range semantic reasoning, where the highest multiple-choice accuracies exceed 80%. Paralinguistic analysis and acoustic localization remain more challenging, with maximum accuracies of 51.09% and 54.38%, respectively. This contrast suggests that current models are less reliable at resolving fine-grained acoustic and temporal details, even when they perform well on other forms of understanding.

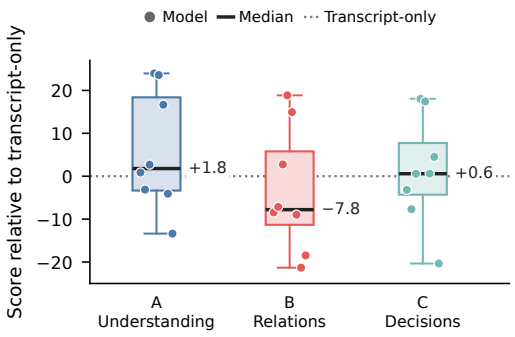

Figure 4: Scores of hosted audio models relative to the transcript-only baseline. Boxes show quartiles and median; whiskers show value range; dots denote individual models.

Extending understanding to cross-audio reasoning and goal-based decisions introduces further challenges. Audio difference analysis reaches a maximum task score of 36.34, and stronger understanding does not consistently translate into better decisions: Qwen3.5-Omni-plus outperforms Doubao on understanding, but their ranking reverses on decisions. Figure 4 further shows that hosted models have a median score difference of -7.8 points from the transcript-only reference on relations and only $+0.6$ on decisions; three and five of the eight models, respectively, exceed this reference. This limited advantage over the textual reference underscores the difficulty of integrating evidence across recordings and using it to make goal-based judgments.

4.3 Fine-grained understanding

We further evaluate fine-grained audio understanding through open-ended perception and content retrieval at varying granularities.

Open-ended perception. We extend the evaluation to direct responses, where models produce counts, orders, or temporal spans from audio without candidate answers. We compare these responses with multiple-choice predictions on 206 matched questions covering one setting from each

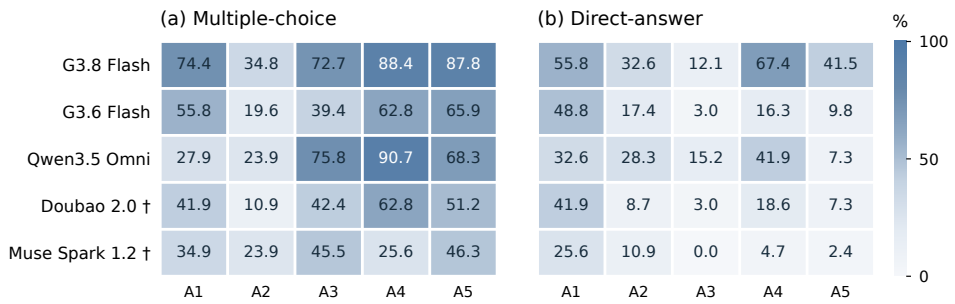

Figure 5: Multiple-choice accuracy and direct-answer success rates on 206 matched questions across five audio understanding sub-tasks. G denotes Gemini.

understanding task, using identical audio across formats within each model. The five evaluated models are Gemini-3.8-Flash, Gemini-3.6-Flash, Qwen3.5-Omni-plus, Doubao-Seed 2.0 Lite, and Muse Spark 1.2; Appendix C.1 details the protocol. Figure 5 shows relatively small differences for counting and rate ordering, but consistently lower success rates for direct localization and retrieval. For example, Gemini-3.8-Flash scores 72.7% on multiple-choice event localization and 12.1% on direct responses, which require locating every target event. These results identify precise localization and complete occurrence recovery as particularly demanding aspects of open-ended audio perception.

Retrieval granularity. We further analyze content retrieval at the levels of unique entities, specified occurrences, and phrases, using results from the six Gemini/Qwen hosted models. Across these question sets, median accuracy is 74.49% for unique entities, 53.33% for specified entity occurrences, and 37.50% for specified phrase occurrences (Table 5). Every model scores lower on phrase-level than entity-level nth-occurrence retrieval. The difference associated with query paraphrasing is smaller, with median accuracy changing from 82.43% to 75.58%. This pattern indicates that distinguishing repeated occurrences and retrieving phrase-level targets remain more challenging than matching query wording in these evaluated settings. Additional exhaustive-retrieval results are reported in Appendix C.2; Appendix C.3 analyzes overlap tolerance, partial recovery, and output validity.

Table 5: Content retrieval comparisons for the six reference hosted models.

Comparison	**Median (%)**
Verbatim → paraphrase	82.43 → 75.58
Unique → nth occurrence	74.49 → 53.33
Entity → phrase (nth)	53.33 → 37.50

4.4 Cross-audio ordering and difference analysis

We examine temporal ordering and audio difference analysis using Gemini-3.8/3.7-Flash, Qwen, Doubao, and Muse at Table 6.

Table 6: Cross-audio reasoning across four model families ($\times 100$): (a) temporal ordering by segment count; (b) audio difference analysis under progressively stricter matching criteria.

Model	**4–5**	**6–7**	**8–10**	**11–13**	**14–16**	**17–19**
Gemini-3.8-Flash	94.29	60.41	65.71	47.82	64.70	55.55
Gemini-3.7-Flash	88.10	50.75	70.85	46.05	63.55	54.15
Qwen3.5-Omni-plus	52.86	25.03	31.22	7.66	27.05	15.29
Doubao 2.0 Lite†	23.33	2.72	23.03	2.35	8.32	9.06
Muse Spark 1.2†	27.14	26.80	29.54	16.79	16.58	25.29

Model	**Onset**	**+Cls**	**+Type**	**+Dur**
Gemini-3.8-Flash	33.90	33.52	24.86	12.43
Gemini-3.7-Flash	46.34	46.34	34.15	18.54
Qwen3.5-Omni-plus	24.78	23.91	12.61	4.78
Doubao 2.0 Lite†	8.90	8.47	6.78	2.97
Muse Spark 1.2†	1.11	0.37	0.37	0.00

Temporal order reconstruction. Table 6(a) compares segment-count groups across 85 questions with varying content and duration. Gemini-3.8-Flash scores 94.29 on 4–5 segments and 55.55 on 17–19 segments in $\tau \times 100$, with fluctuations across intermediate groups. This pattern may reflect the difficulty of integrating local discourse cues into a global sequence: larger sets introduce more

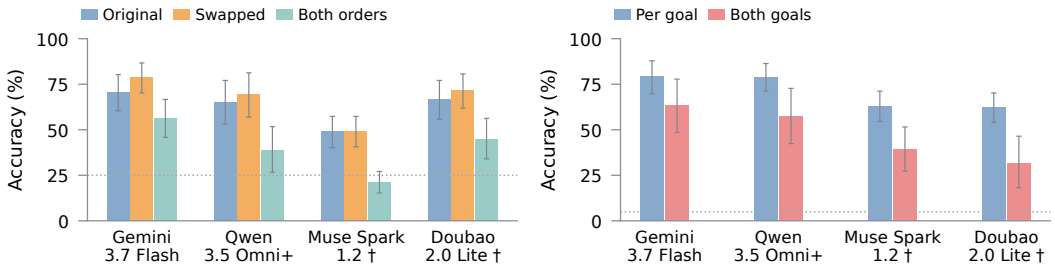

Figure 6: Decision accuracy across three repetitions: input orders (left) and goals (right). Error bars show 95% intervals; dotted lines indicate paired guessing.

candidate transitions, requiring models to reconcile references and discourse continuity across more segments while maintaining a consistent ordering.

Audio difference analysis. Under the nested matching criteria in Table 6(b), Gemini-3.7-Flash matches 95 of 246 gold onsets, all with correct object classes; 70 also match difference type and 38 additionally match duration. Gemini-3.8-Flash and Qwen similarly show larger losses when type and duration are required. This pattern may arise from the need to establish correspondence between versions: insertions and deletions shift timing, while acoustic changes can preserve spoken words. Identifying the difference type and its temporal extent therefore requires aligning corresponding regions and resolving local acoustic differences beyond identifying the affected object category.

4.5 Decision reliability under order and goal changes

To separate condition effects from response variability, we obtain three fresh responses per condition for four hosted models on 49 order pairs and 33 goal pairs (Figure 6). Appendix C.4 provides protocols, paired contrasts, and five-response checks for Doubao and Muse.

Order changes beyond repeat variability. All models have positive estimated excess cross-order disagreement (Table 11). Qwen has 89.80% same-order repeat agreement, yet its excess recording disagreement is 46.94 percentage points (95% interval: $[33.33, 60.42]$); 22/49 pairs select the same input position across all six responses. These order-related selection changes persist beyond repeated-response variability. All four paired accuracy-difference intervals include zero. For Gemini, original/reversed accuracies are 70.75%/78.91%, with a difference of $+8.16$ points ($[-4.08, 20.84]$). Choice instability and mean accuracy differences therefore describe distinct aspects of order dependence.

Persistent failures across goals. Qwen and Doubao have 94.95% and 86.87% same-goal repeat agreement, respectively, yet their both-goal accuracies are 57.58% and 31.31%. Each retains one candidate across all six responses in 11/33 pairs, although the correct candidate changes with the goal; Gemini does so in 5/33 pairs. These unchanged selections persist under replication and demonstrate failures to adapt to the requested criterion on those pairs. Muse's lower same-goal repeat agreement (54.55%) highlights the additional contribution of response variability.

5 Conclusion

We introduced LAMAR-Bench, a benchmark of 1,597 questions evaluating fine-grained understanding, cross-audio relations, and goal-based decisions over long-form and multi-audio inputs. Evaluation of 14 systems reveals persistent challenges in acoustic perception, precise evidence recovery, and audio difference analysis. Paired diagnostics further show that correct answer recognition does not consistently extend to direct localization, while repeated decision diagnostics identify order-related selection changes beyond within-condition variability and persistent unchanged selections across opposing goals. These findings highlight the need for evaluation beyond answer accuracy and motivate further work on evidence-grounded audio understanding and reasoning.

AI USE STATEMENT

Generative AI tools support audio transcription and annotation, question construction, proposed reference judgments, and speech synthesis for constructed audio variants. We also use language models for the textual baselines and source-topic annotation described in Section 4.1 and Appendix A.1, and AI assistance for code development and manuscript editing. Model-assisted annotations and proposed answers are checked against source evidence through the task-specific verification and human-review protocol in Section 3.2. Reference answers derived from timestamps or transformation logs are computed from those records. Reported task scores are computed by the scoring rules in Appendix B.1, rather than by an unconstrained language-model judge. Automatically assigned topic labels describe dataset coverage and do not determine task scores. The authors take responsibility for the methods, results, claims, and final manuscript.

REPRODUCIBILITY STATEMENT

Section 3.2 and Appendix A.3 detail the task construction and verification procedures. Appendices B.1–B.3 document the scoring rules, handling of invalid outputs, paired diagnostic protocols, inference settings, audio adaptation procedures, and textual-baseline inputs. To facilitate reproducibility, we plan to release question manifests, source and transformation metadata, prompts, recorded inference configurations, prediction records, and scripts for scoring and analysis.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

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

Yiming Chen, Xianghu Yue, Xiaoxue Gao, Chen Zhang, Luis Fernando D'Haro, Robby T Tan, and Haizhou Li. Beyond single-audio: Advancing multi-audio processing in audio large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10917–10930, 2024.

Yuhang Dai, Xin Shu, Zengxi Li, Lei Xie, Xiangang Li, and Jianwei Yu. Tag-bench: Benchmarking temporal audio grounding in large audio language models. *arXiv preprint arXiv:2609.01542*, 2026.

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

Peize He, Zichen Wen, Yubo Wang, Yuxuan Wang, Xiaoqian Liu, Jiajie Huang, Zehui Lei, Zhuangcheng Gu, Xiangqi Jin, Jiabing Yang, et al. Audiomarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio llms. *arXiv preprint arXiv:2510.07293*, 2025.

Wen Huang, Yunfei Chu, Meng Gao, Haolin He, and Jin Xu. Audiospan: Spanning the duration and depth of audio comprehension. *arXiv preprint arXiv:2608.26431*, 2026.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22688–22697, 2026.

Chen-An Li and Hung-yi Lee. Inspire: A benchmark for instruction-aware speech retrieval. *arXiv preprint arXiv:2608.16203*, 2026.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, et al. Voxtral. *arXiv preprint arXiv:2507.13264*, 2025.

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. *Advances in Neural Information Processing Systems*, 38, 2026.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

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

Hui Wang, Jinghua Zhao, Yifan Yang, Shujie Liu, Junyang Chen, Yanzhe Zhang, Shiwan Zhao, Jinyu Li, Jiaming Zhou, Haoqin Sun, et al. Speechllm-as-judges: Towards general and interpretable speech quality evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7675–7700, 2026.

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. *arXiv preprint arXiv:2509.17765*, 2025.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, et al. Moss-audio technical report. *arXiv preprint arXiv:2606.01802*, 2026a.

Chih-Kai Yang, Yun-Shao Tsai, Yu-Kai Guo, Ping-Le Tsai, Yen-Ting Piao, Hung-Wei Chen, Ting-Lin Hsiao, Yun-Man Hsu, Ke-Han Lu, and Hung-yi Lee. Mugen: Evaluating and improving multi-audio understanding of large audio-language models. *arXiv preprint arXiv:2603.09714*, 2026b.

Jiaming Zhou, Hongjie Chen, Shiwan Zhao, Jian Kang, Jie Li, Enzhi Wang, Yujie Guo, Haoqin Sun, Hui Wang, Aobo Kong, et al. Diffa: Large language diffusion models can listen and understand. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35112–35120, 2026.

A BENCHMARK STATISTICS, EXAMPLES, AND CONSTRUCTION

A.1 DETAILED BENCHMARK STATISTICS

Task composition. The 1,597 questions comprise 1,135 audio understanding questions across five tasks, 178 cross-audio reasoning questions across two tasks, and 284 goal-based decision questions across three tasks. Table 7 expands the audio understanding tasks in Table 2 into 26 settings, with 32–51 questions per setting. These settings distinguish operations or evidence requirements within speaker analysis (A1), paralinguistic analysis (A2), acoustic localization (A3), content retrieval (A4), and long-range semantic reasoning (A5). For example, speaker analysis separates counting, diarization, matching, localization, attribution, and role classification, while content retrieval varies query wording and occurrence requirements, including unique, nth, and all-occurrence retrieval.

Table 7: Question counts for the 26 settings across the five audio understanding tasks.

Task / setting	#Qs	Task / setting	#Qs
A1 Speaker analysis		**A3 Acoustic localization**	
Speaker counting	43	Event localization	37
Speaker diarization	51	Defect localization	51
Speaker matching	50	Filler localization	40
Speaker localization	47	Repetition localization	32
Speaker attribution	44	**A4 Content retrieval**	
Role classification	40	T0 Verbatim retrieval	37
A2 Paralinguistic analysis		T1 Paraphrase retrieval	43
Fastest passage	50	T2 Unique-entity retrieval	49
Slowest passage	49	T3 Entity: nth occurrence	45
Speaking-rate ordering	46	T4 Entity: all occurrences	41
Longest pause	42	T5 Phrase: nth occurrence	40
Pause-dense passage	42	T6 Phrase: all occurrences	43
A5 Long-range semantic reasoning			
Coreference resolution	42	Question–answer dependency	38
Definition reference	49	Correction chain	44
Total (A)			**1,135**

Per-question input duration. We sum the durations of all clips presented in each question, including each presented version or candidate. Across the 1,597 questions, inputs contain 1–19 clips and total 0.83–59.9 minutes per question, with a mean of 37.41 minutes and a median of 41.82 minutes. Of these questions, 1,209 (75.7%) exceed 30 minutes of summed input.

Multi-audio counting convention. We count the audio clips actually presented in each question, including context recordings and candidate takes. Of the 1,597 questions, 1,020 have one audio input and 577 have multiple inputs. Each multi-audio question is assigned once according to the relationship among its presented inputs: same-source excerpts (115 speaker-analysis and 85 temporal-order questions), edited versions (93 difference-analysis, 92 impact-judgment, and 98 constraint-verification questions), or context with synthesized candidates (94 candidate-selection questions). These categories describe input relationships, not mutually exclusive audio-production operations: an edited version may itself contain synthesized speech. The 94 candidate-selection questions contain three to five candidate takes plus one context recording; their 33 goal pairs count as 66 questions because each goal defines a separate benchmark item.

Examined-region conventions. We map question anchors, correct-answer regions, explicit search scopes, and relevant construction annotations to the timelines of the presented clips. Speaker analysis anchors additionally use source construction records and exact transcript matches. Paralinguistic analysis includes all candidates explicitly compared in a question; localization distractors are

not automatically counted as evidence. Version edits, deletions, and constraint targets are mapped to the corresponding original and output timelines. Each clip is normalized independently before question-level aggregation; clips are not concatenated or normalized by an unseen parent recording. A question enters a position bin if any eligible region intersects it, with duplicate hits counted once and point annotations assigned to their containing bin. This measures bin coverage, not coverage of every instant within a bin. For questions combining local targets and whole-clip judgments, only their local regions enter Figure 3(c). The 112 questions with only whole-clip task scope are excluded from that panel; non-positional constraints such as duration limits do not fill all bins.

Source-topic annotation. We count each distinct source ID once, including each source part separately, yielding 582 sources for audio understanding, 178 for cross-audio reasoning, and 175 for goal-based decisions. GPT-5.6-Sol assigns one primary topic using titles, part titles, descriptions, tags, and six distributed ASR excerpts of up to 550 characters each. Source material with uncertain labels and a stratified sample of other material undergo reannotation with up to ten 1,400-character excerpts and longer descriptions, without the initial label. This verification covers 273 sources. The taxonomy distinguishes subject matter from presentation format; for example, physics lectures are classified as science and technology. We retain an other/unclear category and 94 review flags for topic or format ambiguity or quotation-matching issues.

A.2 Question examples

The examples illustrate evidence recovery, relations across recordings, and judgments under goals and constraints. The audio understanding examples emphasize how distant evidence and fine-grained localization interact: a voice heard near the beginning identifies later speaking turns (A1), brief pauses are compared across distant passages (A2), a subsecond sound must be localized (A3), separated mentions must be retrieved completely (A4), and a later reference must be linked to a short earlier introduction (A5). Questions are translated and abridged from the benchmark; the audio understanding examples retain their multiple-choice format.

A1 Speaker analysis — **distant voice cue, local turns**

Input	one 48.0-minute recording
Question	someone says “So she said that she’s trying to step…” in the recording. Within 1411–1618, which set contains all of this person’s speaking turns?
Options	**(A) 1431.60–1433.52, 1447.12–1506.42, 1566.64–1609.86**; (B) 1431.60–1433.52, 1521.60–1537.60, 1566.64–1609.86; (C) 1410.72–1427.92, 1431.60–1433.52, 1566.64–1609.86; (D) 1431.60–1433.52, 1538.08–1555.44, 1566.64–1609.86
Evidence	the quoted voice cue occurs at 119.80–124.04. The same speaker returns in the three gold intervals; the substituted interval in each distractor belongs to another speaker.
Gold	**A**: all three turns, including a 1.92-second interjection

The voice cue precedes the first target turn by over 21 minutes. The model must connect that earlier voice to precise turn boundaries in a later discussion. All options share two genuine turns, so the remaining turn determines the answer.

A2 Paralinguistic analysis — **pause density**

Input	one 59.6-minute recording
Question	a pause is a gap between utterances of at least 0.3 seconds with no intelligible speech; music or environmental sound may remain. Which of these four 15-second passages contains the most pauses? Judge the pauses, irrespective of what is said.
Options	(A) 1145.00–1160.00; (B) 735.00–750.00; (C) 2360.00–2375.00; **(D) 765.00–780.00**
Evidence	the annotated target at 765.00–780.00 contains seven pauses
Gold	**D**: 765.00–780.00

Candidate start times span over 27 minutes, while each 15-second window must be examined for pauses meeting the 0.3-second threshold. The model must count speech interruptions and compare their frequency across passages.

A3 Acoustic localization — **a subsecond event**

Input one 43.3-minute recording
Question within 1226.3–1526.3, which set of intervals contains the clearly audible mouse clicks?
Options (A) none in this range; **(B) 1525.80–1526.00**; (C) 1319.34–1319.90; (D) 1347.76–1347.92
Gold **B**: a single 0.20-second click at 1525.80–1526.00

Within the specified five-minute search region, the target occupies only 0.20 seconds, near the end. Selecting its interval requires distinguishing a brief acoustic event at the candidate locations; a summary of the spoken content does not identify it.

A4 Content retrieval — **all occurrences**

Input one 38.5-minute recording
Question which set contains exactly all sentences mentioning "Howard," with none missing or added?
Options **(A) 831.80–835.88, 1810.30–1818.78, 1820.84–1826.92**; (B) 1820.84–1826.92, 2291.94–2298.02; (C) 545.44–551.52, 1820.84–1826.92; (D) 438.50–444.58, 680.40–686.48, 1820.84–1826.92
Gold **A**: all three sentence intervals

The three sentence intervals last 4.08–8.48 seconds and extend from about minute 14 to minute 30. Every option contains the genuine occurrence at 1820.84–1826.92. Correct selection therefore requires recovering the other mentions while excluding unrelated sentences.

A5 Long-range semantic reasoning — **a distant antecedent**

Input one 44.8-minute recording
Question consider the statement: "For those of you who don't know, I do the evening news in Fort Wayne, Indiana, and you know the anchorman there is a sinner, no doubt about that." Which interval contains the antecedent of "the anchorman"?
Options (A) 2307.24–2309.00; **(B) 12.50–14.66**; (C) 2354.24–2356.24; (D) 1216.96–1219.52
Evidence at 12.50–14.66: "I'm Mark Mellinger with the Gospel Coalition." The same speaker makes the quoted remark at 726.96–734.56.
Gold **B**: the 2.16-second self-introduction identifies Mark Mellinger as the referent

The reference occurs about 12 minutes after the introduction. Resolving it requires recognizing the speaker's self-reference and linking it to his earlier name, then selecting the short interval containing that introduction. The target sentence itself never says "anchorman".

B1 Temporal order reconstruction — **discourse across segments**

Input five segments from one episode, presented in shuffled order
Question reconstruct the original order, using every segment label exactly once.
Evidence abbreviated transcript excerpts in presentation order: S1: "We focus quite concretely on the websites, apps"; S2: "if you go at it with one approach"; S3: "So basically we do interviews, workshops, surveys"; S4: "Whereas if you use a different approach or a different methodology"; S5: "we provide a link to the end customer"
Gold **`S5,S1,S3,S2,S4`**

The sequence moves from the team's role to its focus and methods, then contrasts reliance on one approach with using another. Discourse links across segments support recovery of the original sequence.

B2 Audio difference analysis — **voice change and pause deletion**

Input version 1 is the original; version 2 is the edited recording
Question report every difference with its onset and duration on the original timeline, difference type, object class, and a description.
Gold the complete answer contains two differences:
(1) onset 263.42, duration 3.68, type **modification**, class **speech**; "All right, now we each solve it on our own blackboard" (same words, different speaker voice)
(2) onset 925.27, duration 0.90, type **deletion**, class **pause**; a pause is removed

Neither difference involves a change in the spoken words. Their detection requires comparing voice identity and timing across recordings, then locating and characterizing both changes.

C1 Impact judgment — **semantic and acoustic consequences**

Semantic integrity	version 2 deletes the sentence at 1358.1–1366.2 in version 1, with everything else unchanged. Does the deletion cause information loss or break continuity? (A) Yes; **(B) No**.
Evidence	the retained statement is "The mind is not in darkness." The deleted sentence repeats it: "The mind is not in darkness, that's what I mean, really."
Gold	**B**: the deleted restatement adds no new information
Acoustic continuity	two versions remove the same content with a seam near 678.5. Which seam sounds more natural, without clipped words, abrupt noise or loudness changes, or clicks? (A) Version 1; **(B) Version 2**.
Evidence	version 1 has mismatched room sound after the seam, with raised background noise and slight reverberation
Gold	**B**: version 2 has the more natural splice

The affected positions are given. The judgments concern whether information survives an edit and whether the resulting sound remains continuous, beyond merely identifying what changed.

C2 Candidate selection — **same audio, different goals**

Input	the current cut has a line missing at 3:56, followed by four candidate takes. Both questions present identical audio in the same order.
Requirements	no excessive fillers and no microphone pops; unlisted aspects need not be considered
Evidence	take 1 has excessive fillers; take 2 uses polished, written-style phrasing; take 3 uses conversational phrasing; take 4 has microphone pops
Goal 1 / gold	conversational and natural, as in ordinary conversation → **take 3**
Goal 2 / gold	more polished and written, with fewer colloquialisms → **take 2**

The requirements exclude takes 1 and 4. The requested style then changes which of the remaining takes is preferred, while the audio stays fixed.

C3 Constraint verification — **per-requirement outcomes**

Input	a delivered cut and its original recording; the cut uses deletions only
Question	check each requirement; for any violation, report its position, content, and type (not removed, wrongly removed, or over limit)
Requirement 1	remove door knocks, car horns, and door closes → **violated**: a door knock remains; its onset is 169.47 on the original timeline and its duration is 3.02 seconds
Requirement 2	keep the live laughter → **satisfied**
Requirement 3	keep the cut at most 1215 seconds long → **violated**: the cut exceeds the duration limit
Gold	per-requirement verdicts: **false, true, false**; overall `pass`: **false**

A retained sound violates one requirement while retained laughter satisfies another. Acceptance also depends on a whole-recording duration limit, so the result requires separate judgments for all three requirements.

A.3 Construction details

Basic parsing and shared derived annotations organize each recording into four complementary components, linked to source timestamps.

1. **Transcripts and alignment.** Sentence- and word-level transcripts record what was said and when, providing text anchors and temporal boundaries for retrieval and audio transformations.
2. **Speaker information.** Speaker assignments link utterances to consistent identities within a recording. Voiceprint clustering and audio-model assignments are cross-checked to identify reliable turns; speaker profiles add supported name or role information.
3. **Acoustic information.** Acoustic-event annotations record sound types and intervals, while voice activity detection supplies speech and pause tracks. Together with word timings, these support event localization and speaking-rate or pause measurements.
4. **Semantic links.** Entity and discourse annotations connect related utterances, including coreference, question–answer dependencies, corrections, and definition references. A separate verification pass checks proposed relations and retains their timestamped evidence.

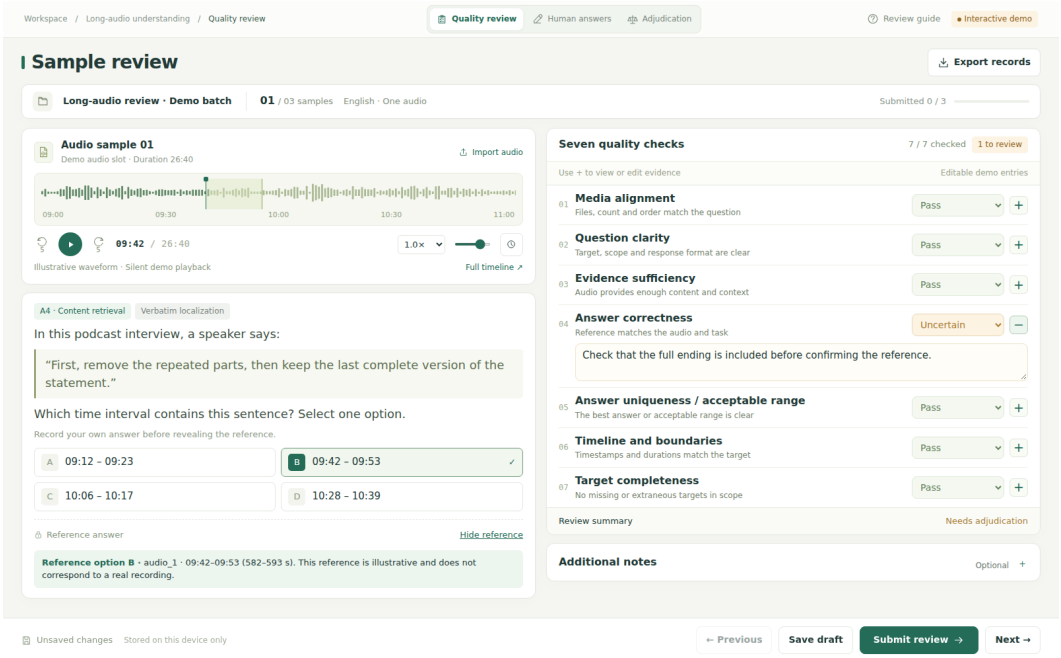

Figure 7: Quality-review interface with audio playback, a revealable reference answer, seven quality checks, and evidence notes. The displayed sample and waveform are illustrative.

Task routing and illustrative examples. Task routing matches the shared annotations to the evidence requirements of the task suite in Table 2. A recording can support multiple task settings when the required evidence is available. The following two examples illustrate this process for selected settings within speaker analysis (A1) and long-range semantic reasoning (A5). First, reliable turns from multiple speakers support speaker localization: a content anchor identifies the target speaker, and that speaker's turns within the requested range provide the reference intervals. Second, a verified link between a later reference and an earlier introduction supports coreference localization: the later reference forms the query, and the earlier introduction provides the answer interval. In both cases, construction proceeds from checking evidence availability to selecting query anchors and deriving the target answer, while reusing the shared annotations.

A.4 Quality-control protocol and review criteria

Quality control proceeds through independent human review, adjudication and manual revision, and a final independent spot check.

Independent human review. Thirty reviewers assessed all 1,597 questions using the seven criteria in Table 8, with two independent reviews per question, producing 3,194 review records. Reviewers worked independently, listened before consulting the reference answer, and recorded evidence and reasons for flagged issues. Figure 7 shows the interface. Agreement between initial ratings, pooled across the seven dimensions, was 93.04%; Table 8 reports agreement and pass rates by dimension. These statistics precede adjudication and manual revision.

Adjudication and manual revision. Nine adjudicators resolved flagged issues and disagreements. Before manual revision, 1,428 of 1,597 questions passed adjudication (89.42%); the remaining 169 (10.58%) required follow-up, and the identified issues were addressed through manual corrections to questions and reference answers. The proportions requiring follow-up were A1: 13.45%, A2: 6.55%, A3: 12.50%, A4: 7.38%, A5: 6.36%, B1: 10.59%, B2: 10.75%, C1: 9.78%, C2: 19.15%, and C3: 18.37%, with C1 combining its semantic and acoustic subtasks. Original review judgments and adjudication decisions were retained.

Table 8: Initial quality-review criteria and results, before adjudication and manual revision. Agreement measures matching nonblank rating pairs, retaining distinct rating categories. Pass rate uses all 1,597 questions as the denominator and counts only questions where both reviewers rated the dimension as pass.

Dimension	Review criterion	Agreement (%)	Pass rate (%)
Media match	All audio is playable, complete, and consistent with the question's input labels.	98.56	98.25
Question clarity	The target, scope, and required output are unambiguous.	94.67	94.36
Evidence sufficiency	The supplied inputs contain the context needed to answer.	96.23	95.87
Answer correctness	Audible evidence supports the proposed reference answer.	80.21	78.40
Answer uniqueness	No alternative is equally valid, or the acceptable range is specified.	93.38	92.74
Time axis and boundaries	Positions refer to the stated audio timeline and correctly delimit the target.	91.64	91.05
Target completeness	All required targets are included; absence claims cover the full search scope.	96.56	94.99

Independent final verification. Revised items were rechecked against the audio and supporting evidence. Independent reviewers then conducted a final spot check with questions sampled uniformly across tasks. All inspected question–answer pairs were judged correct, giving 100% accuracy on the inspected sample after manual revision.

B Evaluation methodology

B.1 Scoring details

Coverage and invalid outputs. Family means require all constituent scores. Missing requests and transport failures withhold the affected task score and family mean; completed but invalid answers receive the task-specific failure treatment below. Missing requests remain in diagnostic denominators and count as failures. Audio understanding multiple-choice questions use correct-option accuracy; invalid answers count as incorrect.

Structured relations (B). For a valid temporal order reconstruction permutation of n segments with D discordant pairs, $\tau = 1-2D/\binom{n}{2}$; invalid permutations score zero and remain in the question average. Audio difference analysis greedily pairs annotated differences with unused predictions within 3 seconds of the gold onset, preferring more matching attributes and then smaller onset distance. These pairs are reused across all four criteria. Object class and difference type require exact agreement; duration requires $|\hat{d}-d| \le \max(0.2d, 0.30\,\mathrm{s})$. For M_ℓ matches satisfying criterion ℓ, micro F1 is $F_1^{(\ell)} = 2M_\ell/(N_{\mathrm{pred}}+N_{\mathrm{gold}})$, with all counts pooled across audio difference analysis questions. The task score is the mean over the four criteria, $\frac{1}{4}\sum_{\ell=1}^{4} F_1^{(\ell)}$. Unmatched differences remain in the micro F1 denominators. Unparseable responses yield no predicted differences, while their annotated differences remain counted.

Decision scores and chance references. Semantic integrity uses macro F1 after mapping options to safe/unsafe labels for each question. An unparseable answer contributes a false negative to its true class; a class with zero F1 denominator receives zero. For acoustic continuity and candidate selection pairs, a missing or unparseable answer sets the corresponding $c_g^{(k)}$ to zero (Appendix B.2). Constraint verification compares verdicts by constraint identity; missing or non-Boolean verdicts are scored as incorrect. The overall pass flag and violation locations are not scored. Analytic chance assumes independent uniform guesses. Its expected semantic integrity macro F1 is 49.70 on the fi-

nite, nearly balanced 43-question set; candidate selection accuracy averages $1/n_{\text{cand}}$ over questions. For constraint verification, Random supplies a valid answer with an independent fair Boolean guess for each required constraint. With k_i constraints in question i, its expected all-constraint accuracy is $\frac{1}{N}\sum_{i=1}^{N} 2^{-k_i} = 16.76\%$ across the 98 questions. Audio difference analysis has no analytic chance entry because a distribution over random difference counts, positions, and attributes is unspecified. Dashes in result tables denote unavailable scores, including undefined chance entries.

B.2 Diagnostic metrics

Answer format and evidence recovery. We compare multiple-choice accuracy with direct-answer success on matched questions to assess evidence recovery without candidate answers. Speaker counts and speaking-rate orders require exact agreement. A single-location answer succeeds if a returned interval has intersection over union (IoU) ≥ 0.5 with the gold interval. For interval sets, predictions are processed in output order, each matched to the unused gold interval with highest IoU meeting the same threshold. With m matches, precision and recall are $P = m/|\widehat{\mathcal{I}}|$ and $R = m/|\mathcal{I}|$ for predicted and gold sets $\widehat{\mathcal{I}}$ and $\mathcal{I}$. We report mean per-question set F1 for partial recovery; complete success requires $P = R = 1$. Two empty sets receive full credit; exactly one empty set receives zero F1. Missing or unparseable outputs fail. These direct-answer scores are supplementary and excluded from family means.

Presentation order and requested goals. For two conditions on the same audio materials, joint accuracy is $\text{Acc}_{\text{pair}} = \frac{1}{G}\sum_{g=1}^{G} c_g^{(1)} c_g^{(2)}$, where G counts pairs and $c_g^{(k)} \in \{0, 1\}$ indicates correctness under condition k. We compare this with individual-condition accuracy, $\frac{1}{2G}\sum_g (c_g^{(1)} + c_g^{(2)})$. Acoustic continuity exchanges input order for all 49 questions: both-order accuracy on the original paired predictions is its main task score. Candidate selection changes the goal for 33 pairs (66 of 94 questions): both-goal accuracy is supplementary, while accuracy over all 94 questions enters the family mean. These conditions reuse source material and do not increase the benchmark question count. Fresh repeated-response diagnostics, including agreement and repeat-averaged paired contrasts, are detailed in Appendix C.4. These additional diagnostics do not replace the main benchmark scores.

B.3 Inference settings and baseline construction

We evaluate each system on all 1,597 benchmark questions and 49 additional reversed-order presentations for acoustic continuity. The evaluated models are introduced in Section 4.1. The repeated decision experiments follow the protocol in Appendix C.4.

We use the released checkpoints of Voxtral-Mini-3B[5], Voxtral-Small-24B[6], Baichuan-Omni-1.5[7], MOSS-Audio-8B-Instruct[8], Phi-4-Multimodal-Instruct[9], and Qwen3-Omni-30B-A3B-Instruct[10].

Inference protocol. Each question is evaluated independently using a single user message that preserves the order of the text and audio segments. Prompts specify the required answer format: an option label for multiple-choice tasks, a segment permutation for ordering, or structured JSON for edit detection and constraint verification. Hosted models receive an additional JSON-format instruction for the A tasks. MOSS and Baichuan use the system instruction "You are a helpful assistant."; the remaining systems use their native chat templates without an additional system instruction. Reference answers and construction annotations are excluded from model inputs. Gemini, Qwen, and the GPT baselines use temperature zero, while Doubao and Muse use provider-default generation settings. Local models use greedy decoding. Top-p and generation seeds are not explicitly set for the main runs, and hosted reasoning settings follow provider defaults. Transient request failures are retried; answer correctness does not determine whether a request is repeated. Responses are parsed

[5] https://huggingface.co/mistralai/Voxtral-Mini-3B-2507
[6] https://huggingface.co/mistralai/Voxtral-Small-24B-2507
[7] https://huggingface.co/baichuan-inc/Baichuan-Omni-1d5
[8] https://huggingface.co/openmoss/MOSS-Audio-8B-Instruct
[9] https://huggingface.co/microsoft/Phi-4-multimodal-instruct
[10] https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct

with deterministic, task-specific rules, and empty or malformed answers receive the failure treatment in Appendix B.1.

Audio preprocessing. Models receive the full duration of each presented clip. The source files comprise 16-kHz mono PCM16 WAV or MP3 and 48-kHz stereo MP3; local model frontends operate at 16 kHz. Model-native feature processing preserves the complete input, including Voxtral's padded 30-second blocks and Qwen's full-length audio frontend. Hosted requests transmit complete files, with response streaming where supported. To accommodate service limits, Doubao and Muse share a matched conversion policy: original files are retained when each file is at most 25 MiB and the serialized request is at most 35 MiB; otherwise, full-duration mono MP3 encoding uses 64 kb/s, or 48 kb/s when further compression is required. The two systems receive identical files in identical order. This adaptation preserves temporal coverage, although lossy compression may affect acoustic cues.

Textual baselines. Both textual conditions use GPT-5.6-Sol with the original question and output specification. The question-only condition replaces audio inputs with numbered placeholders. The transcript-only condition supplies reference text derived from cached `qwen3-asr-flash-filetrans`[11] transcriptions. The ASR service uses channel 0 with word-level output enabled and inverse text normalization disabled, without a natural-language prompt.

C Supplementary results and diagnostics

C.1 Paired answer-format diagnostic protocol

The answer-format comparison uses 206 matched questions from the benchmark, with multiple-choice and direct-answer predictions for all five evaluated models.

Matched question selection. Figure 5 covers five settings: speaker counting (A1), rate ordering (A2), event localization (A3), paraphrase retrieval (A4), and coreference localization (A5). Eligible questions have matching audio, question content apart from answer-format instructions, and gold targets across formats. The shared subset comprises 43 speaker-counting, 46 rate-ordering, 33 event-localization, 43 paraphrase-retrieval, and 41 coreference-localization questions. Selection depends on task coverage and cross-format compatibility, independently of prediction outcomes, and uses the same subset for every model.

Comparison and scoring. We compare multiple-choice accuracy with direct-answer success on the same questions for all five models. Direct answers must provide exact speaker counts or speaking-rate orders, or temporal intervals matching the reference at IoU ≥ 0.5; event localization requires complete recovery of the target set. Malformed answers count as failures. The cross-format results provide descriptive comparisons of task performance, and direct-answer scores are excluded from audio understanding averages. Doubao and Muse receive identical audio across formats and between systems: original files for 169 questions and full-duration 48 kb/s MP3 adaptations for 37 questions.

C.2 Exhaustive content retrieval

We extend the answer-format diagnostic to retrieving every occurrence of an entity or phrase. Table 9 compares multiple-choice accuracy, complete-set accuracy, and mean per-question set F1 on 41 entity-all (T4) and 41 phrase-all (T6) questions. These common subsets are shared by Gemini-3.1-Pro, Gemini-3.5/3.6/3.7/3.8-Flash, and Qwen3.5-Omni-plus. Both answer formats use identical audio and matched question content and reference targets. Two phrase-all questions with differences in question content or reference targets across formats are excluded for every model, independently of prediction outcomes. These supplementary scores do not enter the main task averages.

For Gemini-3.8-Flash, entity-all accuracy is 87.8% with choices but 14.6% under complete direct recovery, with set F1 of 33.6. Phrase-all results are 95.1%, 0.0%, and 10.3, respectively. Partial-credit scores therefore remain low at the default IoU threshold of 0.5: requiring every occurrence

[11] `https://help.aliyun.com/zh/model-studio/qwen-asr-api-reference`

Table 9: Exhaustive content retrieval diagnostic for six hosted models ($\times 100$) on 41 entity-all questions (T4) and 41 phrase-all questions (T6) matched across answer formats for all models. MC is multiple-choice accuracy; Complete and F1 evaluate direct answers. Complete is complete-set accuracy; F1 is mean per-question set F1. Interval matching uses IoU ≥ 0.5.

	T4: Entity, all			T6: Phrase, all		
Model	**MC**	**Complete**	**F1**	**MC**	**Complete**	**F1**
Gemini-3.1-Pro	78.0	0.0	5.3	65.9	0.0	2.7
Gemini-3.5-Flash	41.5	0.0	7.0	41.5	0.0	0.2
Gemini-3.6-Flash	70.7	0.0	6.5	61.0	0.0	4.6
Gemini-3.7-Flash	75.6	7.3	24.7	92.7	2.4	20.3
Gemini-3.8-Flash	87.8	14.6	33.6	95.1	0.0	10.3
Qwen3.5-Omni-plus	68.3	2.4	12.2	75.6	0.0	12.1

to be recovered does not alone account for the observed difficulty. The threshold analysis below examines how these scores depend on temporal overlap tolerance.

C.3 Sensitivity to localization scoring

The answer-format comparisons in Appendices C.1 and C.2 evaluate direct interval predictions at IoU ≥ 0.5, requiring every target to be recovered for set-valued tasks. To assess how these scoring choices affect the observed difficulty, we first vary the overlap threshold while holding model responses fixed. We then examine partial recovery and output validity at the default threshold, distinguishing complete-set failures from partial successes and malformed responses. Together, these analyses clarify how much the reported performance depends on temporal tolerance and complete recovery, without attributing individual errors to a specific cause.

C.3.1 Diagnostic scope and protocol

The analysis covers two existing cohorts. The localization cohort from Appendix C.1 contains 117 questions: event localization (33), paraphrase retrieval (43), and coreference localization (41). Speaker counting and rate ordering do not use IoU and are excluded. The exhaustive-retrieval cohort from Appendix C.2 contains 41 entity-all and 41 phrase-all questions. Both retain their original cross-format compatibility checks, with no further outcome-based filtering.

The full threshold sweep includes all five models in the localization cohort and all six in the exhaustive-retrieval cohort. Figure 8 summarizes the localization cohort using one system from each model family: Gemini-3.8-Flash, Qwen3.5-Omni-plus, Doubao-Seed 2.0 Lite, and Muse Spark 1.2. Gemini-3.8-Flash is the latest evaluated Gemini generation in this cohort; Gemini-3.6-Flash remains in the full analysis.

Fixed-response protocol. We rescore the same predictions at IoU thresholds of 0.3, 0.5, and 0.7, holding audio, questions, reference intervals, and parsing fixed. Single-target A4/A5 success requires a returned interval matching the reference; A3 and exhaustive retrieval retain prediction-order greedy one-to-one matching and require complete set recovery for success. Set F1 grants partial credit. Invalid outputs remain failures.

C.3.2 Sensitivity to overlap tolerance

Event, paraphrase, and coreference localization. Relaxing IoU from 0.5 to 0.3 raises Gemini-3.8-Flash's event-localization success from 12.1% to 21.2%, while its matched multiple-choice accuracy is 72.7%. Its paraphrase- and coreference-localization success rates rise from 67.4% to 76.7% and from 41.5% to 56.1%, respectively. Across all 15 model–setting combinations in the full five-model localization analysis, success at IoU 0.3 remains below the corresponding multiple-choice accuracy, with gaps of 11.6–57.6 percentage points. Relaxing IoU can admit coarse or oversized intervals as well as small boundary offsets, so the sweep alone does not identify the cause of a localization error.

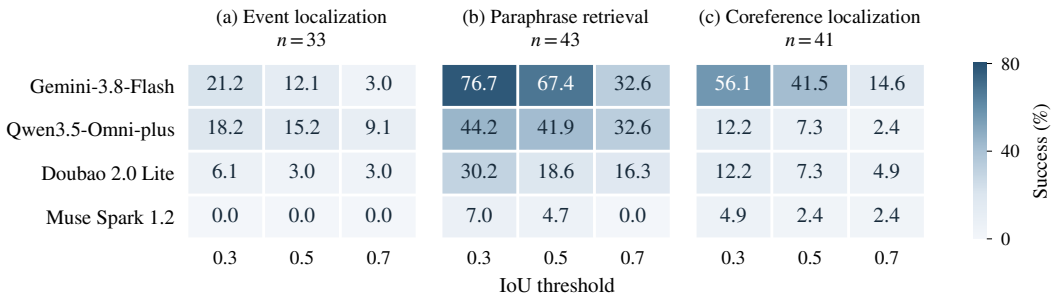

Figure 8: Localization success (%) across IoU thresholds: complete set recovery for events, and a matching interval for paraphrase and coreference retrieval (n: questions).

Table 10: Partial and complete set recovery at IoU 0.5 ($\times 100$, except Invalid counts). P/R pool interval counts; F1 averages per-question set F1. These have different weighting. Invalid answers receive zero success/F1; their gold intervals remain in the recall denominator.

Model	Micro P	Micro R	Macro F1	Complete	Invalid
A3: Event localization ($n = 33$)					
Gemini-3.8-Flash	17.0	17.0	18.2	12.1	0
Gemini-3.6-Flash	3.2	4.3	3.8	3.0	4
Qwen3.5-Omni-plus	14.3	23.4	23.2	15.2	0
Doubao-Seed 2.0 Lite†	5.2	6.4	6.3	3.0	0
Muse Spark 1.2†	4.2	2.1	0.7	0.0	0
T4: Entity, all ($n = 41$)					
Gemini-3.1-Pro	5.1	2.8	5.3	0.0	0
Gemini-3.5-Flash	19.1	3.3	7.0	0.0	19
Gemini-3.6-Flash	18.7	3.5	6.5	0.0	17
Gemini-3.7-Flash	19.8	16.9	24.7	7.3	2
Gemini-3.8-Flash	29.6	20.4	33.6	14.6	5
Qwen3.5-Omni-plus	7.9	16.9	12.2	2.4	0
T6: Phrase, all ($n = 41$)					
Gemini-3.1-Pro	2.4	1.3	2.7	0.0	0
Gemini-3.5-Flash	0.8	0.3	0.2	0.0	2
Gemini-3.6-Flash	6.5	3.0	4.6	0.0	0
Gemini-3.7-Flash	17.2	15.4	20.3	2.4	0
Gemini-3.8-Flash	13.0	9.0	10.3	0.0	5
Qwen3.5-Omni-plus	9.9	12.7	12.1	0.0	0

Exhaustive retrieval. In the separate exhaustive-retrieval cohort, relaxing IoU from 0.5 to 0.3 raises Gemini-3.8-Flash's entity-all and phrase-all set F1 from 33.6 and 10.3 to 43.4 and 23.0. Complete recovery at IoU 0.3 remains 17.1% and 2.4%. Score magnitudes therefore depend on boundary tolerance, while complete recovery remains difficult under the relaxed criterion. Table 9 reports all six models at the default threshold, and the full threshold sweep accompanies the analysis data.

C.3.3 Partial recovery and output validity

The threshold sweep tests sensitivity to temporal overlap, but complete-set accuracy still treats a partially recovered set as a failure. We therefore return to IoU 0.5 and examine the three set-valued settings: event localization, entity-all retrieval, and phrase-all retrieval. Table 10 reports pooled interval precision and recall alongside mean per-question set F1, complete recovery, and invalid-response counts. Precision and recall pool matched, predicted, and gold intervals; set F1 averages questions equally. Invalid responses receive zero success and F1, and their reference intervals remain in the recall denominator. All questions in these subsets have nonempty reference sets.

Partial recovery. For Gemini-3.8-Flash, event-localization set F1 is 18.2, compared with 12.1% complete recovery, despite no invalid responses. Its entity-all and phrase-all recall values are 20.4% and 9.0%, respectively. These results show that granting partial credit still leaves substantial recovery errors at the default threshold. Low recall can reflect missed targets, misplaced boundaries, or selection of incorrect content, which these aggregate metrics do not distinguish.

Output validity. Gemini-3.5-Flash and Gemini-3.6-Flash have 19 and 17 invalid entity-all responses out of 41, respectively, so their low scores also reflect output-validity failures. Gemini-3.8-Flash has five invalid responses in each exhaustive task. On its 36 valid responses per task, entity-all and phrase-all set F1 are 38.3 and 11.7; these descriptive values do not replace the full-denominator scores.

C.4 Paired decision diagnostics

The decision experiments in Section 4.5 test two complementary requirements: a model should preserve its preferred recording when only presentation order changes, and select a different candidate when the requested goal requires it. A single response per condition cannot separate these behaviors from ordinary response variability. We therefore use repeated paired trials to distinguish repeatability, order invariance, and correct goal adaptation. Below, we describe the shared design, analyze each intervention, and check whether the conclusions persist with additional repetitions.

C.4.1 Paired design and interpretation

Controlled conditions and repetitions. We retain all 49 acoustic-continuity order pairs (48 source clusters) and 33 candidate-selection goal pairs (33 source clusters). Order reversal preserves the question and preferred recording but reverses its answer label. Goal changes preserve the audio and input order but change the correct candidate. Each condition is run three times for Gemini and Qwen and five times for Doubao and Muse. The primary comparison uses preassigned repetitions 1–3 for every model; historical responses are excluded. Audio and prompts are fixed within conditions, and calls are interleaved across conditions and repetitions. Gemini and Qwen use temperature 0, a 16,000-token output budget, and requested seeds 2026092501–2026092503 matched across conditions; other generation settings use provider defaults. Seed acceptance does not establish deterministic backend behavior. Doubao and Muse retain their provider-default settings and full-audio adaptations, without a fixed seed. Only transport failures are retried; empty or malformed responses count as incorrect.

What the metrics measure. For pair $g \in \{1, \dots, G\}$, condition $k \in \{1, 2\}$, and repetition $r \in \{1, \dots, R\}$, let c_{gkr} indicate correctness and $\bar{c}_{gk} = R^{-1} \sum_r c_{gkr}$. Condition accuracy, both-condition accuracy, and the paired accuracy contrast are

$$\widehat{A}_k = \frac{1}{G} \sum_g \bar{c}_{gk}, \qquad \widehat{A}_{\text{pair}} = \frac{1}{GR} \sum_{g,r} c_{g1r} c_{g2r}, \qquad \widehat{\Delta} = \frac{1}{G} \sum_g (\bar{c}_{g2} - \bar{c}_{g1}).$$

The product in $\widehat{A}_{\text{pair}}$ uses matched repetition indices. Within each condition, repeat agreement S_k averages the fraction of unordered repetition pairs with identical valid choices; the tables report $S = (S_1 + S_2)/2$. Invalid outputs receive no agreement credit. High S means that a choice is repeatable, regardless of whether it is correct. The *Both* column instead requires correct decisions in both conditions; it is necessarily no higher than either condition accuracy, so a lower value alone does not establish sensitivity. Independent uniform paired guessing yields 25% accuracy for order pairs and 4.95% for goal pairs.

Uncertainty. Every statistic, including $\widehat{\Delta}$ and the excess disagreement defined below, is recomputed in 5,000 source-cluster bootstrap draws (seed 20260909), retaining all conditions and repetitions within each source. Tables report pointwise 95% percentile intervals without multiplicity adjustment. Repetitions are not treated as independent source materials, and contrast intervals are computed directly rather than by subtracting marginal interval endpoints.

C.4.2 ORDER SENSITIVITY BEYOND RESPONSE VARIABILITY

Order invariance concerns the identity of the selected recording, not its displayed label. We therefore map reversed-order choices back to recording identities. Let D_{cross} be one minus valid cross-order agreement over repetition pairs with distinct indices. We compare it with the same-condition disagreement baseline:

$$E = D_{\text{cross}} - (1 - S).$$

Excluding equal repetition indices makes both terms compare distinct seeds where seeds are supplied. Invalid comparisons contribute disagreement in both terms. A positive E indicates more selection disagreement across orders than under an unchanged order; $\widehat{\Delta}$ instead measures the net change in correctness.

Table 11: Presentation order ($R = 3$). S: same-order repeat agreement; Both: correct under both orders; Δ: reversed minus original accuracy; E: cross-order disagreement beyond the repeat baseline. Accuracy and S are percentages; Δ and E are percentage points. Brackets: 95% source-cluster intervals.

Model	S	Original	Reversed	Both	Δ	E
Gemini-3.7-Flash	75.51 [68.75, 82.00]	70.75 [60.42, 80.27]	78.91 [70.14, 86.67]	56.46 [45.83, 66.67]	+8.16 [-4.08, 20.84]	+12.24 [4.67, 20.41]
Qwen3.5-Omni-plus	89.80 [84.03, 94.67]	65.31 [53.06, 77.08]	69.39 [56.94, 81.25]	38.78 [26.67, 51.70]	+4.08 [-15.97, 23.81]	+46.94 [33.33, 60.42]
Doubao-Seed 2.0 Lite†	75.51 [68.03, 82.99]	66.67 [55.78, 77.08]	71.43 [61.81, 80.67]	44.90 [34.00, 56.25]	+4.76 [-10.42, 20.41]	+21.09 [9.72, 32.67]
Muse Spark 1.2†	46.26 [40.00, 53.12]	48.98 [40.14, 57.33]	48.98 [40.67, 57.33]	21.09 [15.28, 27.08]	+0.00 [-13.89, 14.29]	+10.20 [0.69, 20.14]

Stable choices can depend on presentation position. Table 11 shows positive excess disagreement for all four systems, with all four pointwise intervals above zero. Qwen illustrates why the repeat baseline matters: its same-order repeat agreement is 89.80%, yet its excess cross-order disagreement is 46.94 points. Thus, much of its cross-order inconsistency exceeds the variability observed when the inputs are unchanged. At the pair level, Qwen selects the same displayed position in all six responses for 22/49 pairs: the first position in 19 and the second in three. Since reversal changes the recording occupying that position, these are persistent failures of recording invariance. The predominance of the first position in this subset suggests a positional tendency, but does not establish a universal preference across questions.

Average accuracy can conceal changes in individual decisions. All four order-contrast intervals include zero, despite the positive excess-disagreement intervals. These findings address different questions: improvements on some pairs can offset regressions on others, producing little net accuracy change even when many selected recordings change. Qwen's small mean contrast therefore does not imply order invariance. Muse requires a different emphasis: its same-order agreement is only 46.26%, so substantial inconsistency is already present before order changes. Its positive E indicates an additional order-related component, while its lower E than Qwen's should not be read as greater overall reliability.

C.4.3 GOAL ADAPTATION VERSUS REPEATABILITY

For goal pairs, changing the selected candidate is necessary because the correct candidate changes. Selection agreement across goals is therefore not the desired behavior; correctness under both goals is. Table 12 reports both-goal accuracy alongside same-goal repeat agreement to distinguish reliable repetition from successful adaptation.

Repeatable selection can persist under the wrong goal. Qwen and Doubao have high same-goal repeat agreement (94.95% and 86.87%), but both-goal accuracies of 57.58% and 31.31%. The gap between these metrics motivates inspecting individual pairs rather than treating repeatability as evidence of goal following. Each model retains one candidate across all six responses in 11/33

Table 12: Requested goal ($R = 3$). S: same-goal repeat agreement; Both: correct under both goals; Δ: Goal 2 minus Goal 1 accuracy. Goal indices are local to each pair. Accuracy and S are percentages; Δ is in percentage points. Brackets: 95% source-cluster intervals.

Model	S	Goal 1	Goal 2	Both	Δ
Gemini-3.7-Flash	89.90 [83.84, 94.95]	84.85 [73.74, 94.95]	73.74 [59.60, 86.87]	63.64 [48.48, 77.78]	-11.11 [-28.28, 5.05]
Qwen3.5-Omni-plus	94.95 [90.91, 98.99]	77.78 [63.64, 90.91]	79.80 [66.67, 91.92]	57.58 [42.40, 72.73]	+2.02 [-18.18, 23.23]
Doubao-Seed 2.0 Lite†	86.87 [78.79, 93.94]	53.54 [38.38, 67.70]	70.71 [56.57, 83.84]	31.31 [18.18, 46.46]	+17.17 [-6.06, 41.41]
Muse Spark 1.2†	54.55 [45.45, 63.64]	68.69 [58.59, 77.78]	57.58 [44.44, 69.70]	39.39 [27.27, 51.52]	-11.11 [-27.27, 5.05]

pairs, compared with 5/33 for Gemini. Because the reference candidate changes, a constant selection must fail at least one goal on every repetition. These cases directly demonstrate persistent failures to adapt, beyond the mechanical reduction caused by requiring two correct answers. They do not by themselves distinguish misunderstanding the requested criterion from misjudging the candidate audio.

Response variability and goal differences. Muse's same-goal agreement is 54.55%, indicating that its goal-pair failures also occur amid substantial variation under identical inputs. This differs from the persistent unchanged selections above and limits explanations based only on candidate preference. Moreover, Goal 1/2 follows the stored brief index, not a common semantic contrast across pairs. The corresponding $\widehat{\Delta}$ is descriptive: neither its sign nor the two column averages identifies a generally easier goal or a shared direction of goal sensitivity.

C.4.4 Stability with additional repetitions

Using all five Doubao and Muse repetitions retains positive excess order disagreement: 20.82 points ($[11.84, 31.00]$) and 9.18 points ($[2.92, 16.25]$), respectively. Their same-goal repeat agreements remain close to the primary estimates, at 85.15% and 56.36%. The additional responses therefore preserve the distinction between order-related changes and within-condition variability, including the much lower repeatability of Muse's goal decisions. This is a sensitivity check using overlapping responses, not an independent replication. The small numbers of sources and repetitions, together with differences in provider generation settings, limit model comparisons. The diagnostics establish observed decision patterns under these conditions; they do not isolate an internal mechanism or guarantee behavior on unseen goals and recordings.