AFTER THE SECOND BELL: CAN AUDIO-LANGUAGE MODELS REASON WITH TEMPORALLY GROUNDED EVENTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-language models (ALMs) increasingly answer questions about speech, environmental sounds, and music, but temporal operations such as ordering, counting, and localization are often evaluated as separate skills. We study a more specific dependency: whether a model can identify a particular event occurrence, use its temporal position as a reference, and execute a downstream computation conditioned on that reference. We introduce TEROS (Temporal Event-Reference and Relational Reasoning), a diagnostic benchmark spanning speech, environmental sound, and music. The benchmark contains 67,466 questions in Task 1 across four source datasets and 2,955 matched questions in Task 2 over short MusicNet clips. Task 2 fixes two temporal intervals and asks models to count within each interval, determine interval overlap, and compute their intersection and union without double counting. Ground truth is generated from structured annotations and retained using cross-model agreement, followed by expert verification. Across the evaluation suite, models show a consistent gap between coarse temporal judgments and exact temporal counting: binary questions are often answered more reliably than numerical counting, while Task 2 intersection and union counting remain difficult even when interval-overlap recognition is substantially higher. We further find that Task 1 category scores are strongly affected by answer type and therefore should not be interpreted as a difficulty hierarchy. TEROS exposes a narrower failure mode: temporally grounded information can be recognized locally without being reliably carried into a subsequent relational computation.

1 INTRODUCTION

Audio-language models (ALMs) have moved from speech recognition and captioning toward question answering and reasoning over heterogeneous audio. Recent benchmarks evaluate temporal ordering, event counting, compositionality, and long-context understanding (Ghosh et al., 2024; Bhattacharya et al., 2025; Kumar et al., 2025; Luo et al., 2026; Ahia et al., 2025). A more specific question is whether a model can identify one particular event occurrence and then use it to constrain a later computation.

Consider the query: *How many events occur after the second bell?* Solving it requires identifying the event type, selecting its second occurrence, locating it in time, and using that temporal boundary to filter another event stream before counting. We call this capability *temporal event-reference reasoning*. A broad family of such queries can be expressed as

$$\mathrm{AGG}\big(E_B \mid R(G(E_A))\big), \tag{1}$$

where E_A is a reference event type, G identifies a particular occurrence or temporal boundary, R is a relation such as BEFORE, AFTER, or BETWEEN, and AGG is a downstream operation such as counting or comparison. The critical dependency is that the output of G becomes an operand for the subsequent computation.

Prior work establishes important components of this problem. CompA studies compositional event ordering and attribute binding; TREA evaluates temporal ordering, counting, duration, and consistency; and MMAU-Pro broadens evaluation across speech, sound, music, long-form, and multi-audio

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.

settings (Ghosh et al., 2024; Bhattacharya et al., 2025; Kumar et al., 2025). Recent work further studies long-form temporal understanding and temporal grounding, including ChronosAudio, BLAB, Listening with Time, MusTBENCH, SpotSound-Bench, TEMPO, and TAG-Bench (Luo et al., 2026; Ahia et al., 2025; Shao et al., 2026; Kwon et al., 2026; Sun et al., 2026; Kulkarni et al., 2026; Dai et al., 2026). TEROS instead focuses on the downstream dependency in which a grounded occurrence is consumed as a reference for another relational operation.

We introduce TEROS across four audio sources: VoxConverse and MUStARD for speech, DataSED for environmental sound, and MusicNet for music (Chung et al., 2020; Castro et al., 2019; Fredianelli et al., 2025; Thickstun et al., 2017). Task 1 provides broad coverage of temporal event-reference queries, while Task 2 is a matched diagnostic over 591 short MusicNet configurations, with five questions sharing the same pair of temporal intervals: two constituent counts, overlap, intersection counting, and union counting with shared events counted once.

We organize the study around three questions. **RQ1:** Does temporal event-reference reasoning generalize across speech, environmental sound, and music? **RQ2:** How much does answer format affect measured accuracy, and does source duration explain additional variation? **RQ3:** Can models compose multiple grounded temporal constraints into exact downstream operations such as intersection and union counting?

Across these analyses, the benchmark reveals a distinction between recognizing temporal relations and using grounded temporal information compositionally. Task 1 shows substantial cross-domain and answer-format variation, while duration alone does not explain the observed behavior. Task 2 provides the clearest diagnostic: models answer the coarse overlap relation substantially more often than they produce the exact counts required by the same temporal constraints, with intersection and union particularly difficult.

Our contributions are threefold: (1) we formalize temporal event-reference reasoning as a class of audio queries in which a grounded event occurrence becomes an operand for downstream computation; (2) we introduce TEROS across speech, environmental sound, and music, combining broad Task 1 coverage with a matched multi-constraint Task 2 diagnostic; and (3) we analyze cross-domain behavior, answer-format and duration effects, and controlled temporal composition. Together, these results characterize the transition from temporal recognition and grounding to temporally grounded relational computation.

2 Related Work

2.1 Audio reasoning benchmarks

Audio evaluation has expanded from recognition and captioning toward structured reasoning. CompA established that compositional ordering and attribute binding can be difficult for ALMs even when the constituent events are recognizable (Ghosh et al., 2024). TREA subsequently provided a temporal reasoning suite covering ordering, duration, and counting, together with consistency analysis under semantically equivalent perturbations (Bhattacharya et al., 2025). MMAU-Pro evaluates a much broader collection of audio skills across speech, sound, music, long-form, and multi-audio settings (Kumar et al., 2025). TEROS is narrower by design: it asks whether one temporally grounded occurrence can serve as a reference for a subsequent computation.

2.2 Temporal grounding and long-form reasoning

Recent work increasingly evaluates whether audio-language models can preserve and localize temporal information. ChronosAudio evaluates temporal understanding across short, middle, and long-form audio (Luo et al., 2026), while BLAB and Listening with Time emphasize long-form conversational and temporal grounding settings (Ahia et al., 2025; Shao et al., 2026). MusTBENCH evaluates temporally grounded question answering for music (Kwon et al., 2026), and SpotSound-Bench targets fine-grained localization of short events in dense audio (Sun et al., 2026). More recently, TEMPO studies unified timestamping across audio, speech, and music (Kulkarni et al., 2026). TAG-Bench evaluates temporal audio grounding by requiring models to return all intervals matching a query, including occurrence enumeration over recordings ranging from seconds to minutes (Dai et al., 2026). These studies make clear that locating events and enumerating their temporal occurrences are

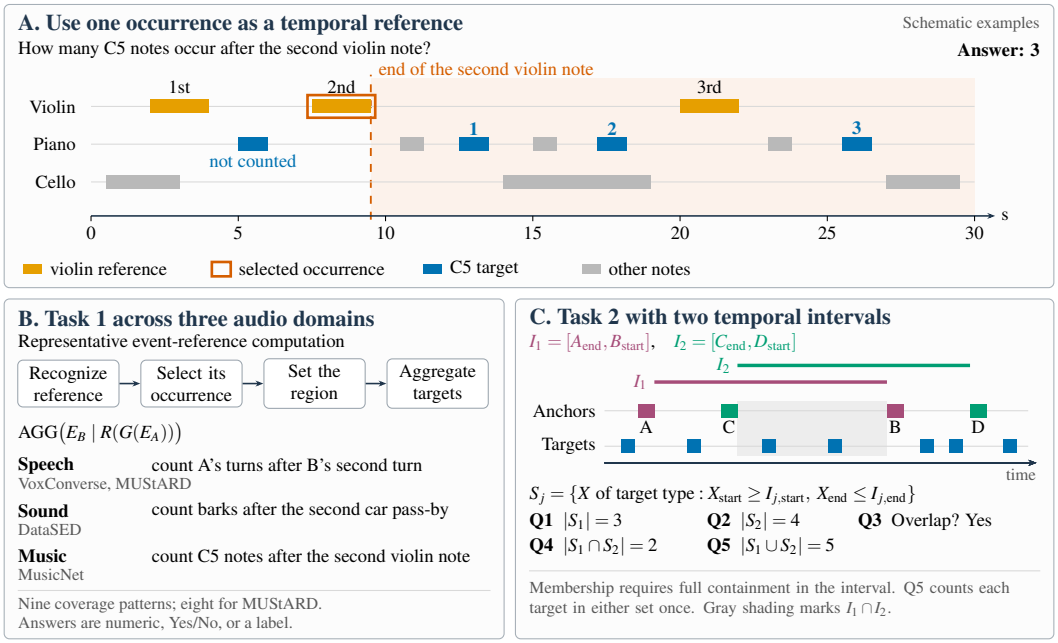

Figure 1: **Overview of TEROS.** The motivating "second violin note" example illustrates the event-reference computation. The same structure is instantiated with speaker occurrences, environmental sound events, and musical notes. Task 1 evaluates single-constraint relational temporal reasoning across four datasets; Task 2 composes two temporal intervals and evaluates overlap, intersection, and union.

important capabilities. TEROS targets a downstream dependency: the localized occurrence is not itself the final answer, but a reference that conditions another temporal computation.

2.3 Speaker- and Domain-Specific Temporal Reasoning

VoxConverse provides naturally occurring multi-speaker speech with diarization annotations (Chung et al., 2020); MSU-Bench evaluates speaker-centric conversational reasoning (Sun et al., 2026). In music, MusTBENCH examines temporal grounding of musical events (Kwon et al., 2026). TEROS uses speech, environmental sound, and music as interchangeable substrates for a common event-reference abstraction. The event may be a speaker turn, environmental sound instance, or musical note, but the logical dependency is the same: ground one occurrence and use it to constrain a subsequent computation.

Taken together, prior work establishes strong primitives for temporal recognition, localization, ordering, counting, and long-context understanding. TEROS studies their composition: *ground an event occurrence → use its temporal position as a reference → perform a downstream relational computation*. This distinction is particularly important for exact counting, where recognizing a temporal relation is insufficient unless the model preserves the identity and boundary of the grounded occurrence.

3 TEROS Benchmark

3.1 Benchmark Scope and Data Sources

Task 1 contains 67,466 questions across four datasets and three audio domains. Table 2 summarizes the construction. The source recordings are not all of the same duration: VoxConverse spans 0–30s through beyond 20 minutes, MUStARD is concentrated below two minutes, DataSED spans 0–6 minutes with a small long tail, and MusicNet is explicitly constructed in 0–30s, 30s–2min, and 2–6min regimes.

Table 1: **Positioning of TEROS among related audio benchmarks.** Checkmarks indicate that the benchmark explicitly evaluates the corresponding capability; they do not imply that the capability is its exclusive focus. The key distinction is whether a grounded occurrence is subsequently consumed as a reference for another computation.

Benchmark	Temporal reasoning	Grounding /localization	Occurrence enumeration	Reference→ downstream computation	Exact multi-constraint computation	Domains
CompA	✓	–	–	–	–	Speech
TREA	✓	–	–	–	–	Audio
MMAU-Pro	✓	–	–	–	–	Speech/sound/music
MusTBENCH	✓	✓	–	–	–	Music
SpotSound-Bench	–	✓	–	–	–	Audio
TEMPO	–	✓	–	–	–	Audio/speech/music
TAG-Bench	–	✓	✓	–	–	Audio
TEROS	✓	✓	✓	✓	✓	**Speech/sound/music**

Table 2: Task 1 construction in the current question exports. MUStARD has eight categories because no genuine concurrent-overlap question is supported by its available annotations. Question counts are taken from one complete model export per dataset; partial model exports do not change the benchmark question count.

Dataset	Domain	Audio units	Categories	Questions
VoxConverse	Speech	446	9	9,516
MUStARD	Speech	690	8	15,318
DataSED	Environment	717	9	20,191
MusicNet	Music	868	9	22,441
Total		**2,723**		**67,466**

For VoxConverse, we use the source RTTM speaker annotations (Chung et al., 2020). MUStARD contains conversational clips (Castro et al., 2019); we derive the speaker temporal intervals needed for occurrence-based questions during benchmark construction. DataSED supplies temporally localized environmental-event annotations (Fredianelli et al., 2025). MusicNet supplies aligned note and instrument annotations (Thickstun et al., 2017).

3.2 Temporal event representation

An *event* is a semantically identifiable phenomenon localized in time; an *occurrence* is one instance of that event. Temporal relations are defined from event boundaries:

$$X \prec Y \iff X_{\text{end}} \leq Y_{\text{start}}, \tag{2}$$

$$X \succ Y \iff X_{\text{start}} \geq Y_{\text{end}}, \tag{3}$$

$$X \bowtie Y \iff X_{\text{start}} < Y_{\text{end}} \wedge Y_{\text{start}} < X_{\text{end}}. \tag{4}$$

For X to belong to $\text{BETWEEN}(A, B)$, we require full containment:

$$X_{\text{start}} \geq A_{\text{end}} \quad \wedge \quad X_{\text{end}} \leq B_{\text{start}}. \tag{5}$$

Thus an event that merely crosses a boundary is excluded.

Speech occurrences are continuous episodes by a speaker. Consecutive portions by the same speaker are merged, and a new occurrence is counted only after another speaker intervenes; silence alone does not create a new turn. Speakers are indexed by first appearance. For environmental sound, we retain individual annotation instances rather than importing the speech-turn convention. For MusicNet, every annotated note is one occurrence; adjacent or overlapping annotations remain separate. This distinction is important because the benchmark's unit of reasoning is the annotated event occurrence, not an inferred perceptual segment.

3.3 TASK 1: RELATIONAL TEMPORAL REASONING

Task 1 contains nine reasoning patterns when supported by a dataset. The taxonomy provides concise definitions. The patterns are deliberately not ordered by presumed cognitive difficulty. They instead provide coverage of common temporal computations.

The nine patterns are Direct Recognition; Between Event A and Event B; Before/After and Conditional Counting; Sequence Patterns; Nested Temporal; Repetition and Source Tracking; First/Last Occurrence; Comparative Counting; and Concurrent/Overlapping Sound. These patterns are a coverage taxonomy rather than a presumed difficulty ordering.

A representative computation is

$$\mathrm{COUNT}(B \mid \mathrm{BEFORE}(\mathrm{OCC}_1(C))), \tag{6}$$

where the first occurrence of C must be grounded before the occurrences of B can be counted. Multiple natural-language questions can instantiate the same underlying computation. We retain the program representation during construction so that later analyses can separate linguistic realization from logical structure.

3.4 TASK 2: MULTI-CONSTRAINT TEMPORAL REASONING

Task 2 is intentionally smaller and more controlled. It contains 591 valid configurations drawn from 200 short MusicNet clips, producing 2,955 questions. Each configuration defines

$$I_1 = \mathrm{BETWEEN}(A, B), \qquad I_2 = \mathrm{BETWEEN}(C, D). \tag{7}$$

Five questions are asked from the same configuration: count target events in I_1; count target events in I_2; determine whether I_1 and I_2 overlap; count target events in $I_1 \cap I_2$; and count target events in $I_1 \cup I_2$, counting shared events once.

The interval-overlap question is highly imbalanced: 590 of 591 configurations have a positive answer and one has a negative answer. Its raw accuracy is therefore reported with the majority baseline and is used as a within-case diagnostic rather than as a standalone overlap-classification score.

3.5 QUESTION GENERATION AND VALIDATION

Questions are generated from structured annotations rather than from unconstrained inspection of the waveform. Dataset-specific prompts specify admissible event references, occurrence semantics, temporal relations, and answer formats. Unsupported or ambiguous questions are discarded. For MusicNet, questions use human-readable pitch and instrument names rather than raw annotation identifiers.

For benchmark validation, we use independent language-model validators including Gemini Flash 3.8, a Qwen 200B reasoning model, and GPT Astra 6. We retain question-answer constructions supported by agreement from at least two validators and independently verify a stratified subset with expert human review. The released benchmark records the normalized answer and the structured computation; the validator models are not used during ALM evaluation.

3.6 EVALUATION PROTOCOL

Models receive audio and the natural-language question, without annotation tables, timestamps, event identifiers, or the underlying temporal program. Responses are normalized to exact-match answers. The primary metric is

$$\mathrm{Acc} = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[\hat{y}_q = y_q]. \tag{8}$$

We report overall, dataset, category, answer-type, and duration results. Because Task 1 mixes answer types, answer-type-stratified results are essential.

For Task 2, we report overall accuracy, accuracy on integer questions, accuracy on Yes/No questions, and the five matched question types separately. Task 2 contains no ordinal ground-truth answers, so ordinal accuracy is not reported. Evaluation accepts unambiguous semantic variants of the required

Table 3: Task 1 accuracy (%) on the available questions. The final column is the macro-average of the four dataset accuracies; for Kimi-Audio, the mean is over its available dataset scores because its VoxConverse and MUStARD exports are partial.

Model	Vox	MUS	DataSED	MusicNet	Mean
DeSTA2.5-Audio	27.5	33.6	34.6	22.3	29.5
Gemma 3n E2B	45.6	42.8	46.4	35.0	42.4
Gemma 3n E4B	52.5	41.7	43.9	38.1	44.0
Kimi-Audio[†]	43.4	44.4	41.4	20.7	37.5
MiMo-Audio	48.0	43.9	42.7	33.6	42.0
OmniVinci	55.8	53.4	52.9	33.1	48.8
Voxtral-Small-24B	57.8	50.3	45.0	28.4	45.4

[†] Kimi-Audio has 8,130 VoxConverse and 7,443 MUStARD predictions, versus 9,516 and 15,318 questions respectively.

answer format (for example, "three" or "there are three" for an integer answer, and "yes, ..." for a Yes/No answer). Empty responses and responses whose intended answer cannot be determined unambiguously are counted as incorrect. Outputs explicitly stating that the audio file or recording is unavailable are excluded from the denominator because they are neither correct nor incorrect with respect to the ground truth. We do not use conditional measures such as $P(Q_4 \text{ correct} \mid Q_1, Q_2 \text{ correct})$ as primary metrics because the number of cases in which models answer both constituent counts correctly is too small for stable estimates.

3.7 Evaluation suite

The supplied Task 2 export contains 13 model outputs: 11 models evaluated on all 591 configurations (2,955 questions each), Gemini 3.8 Flash on 249 configurations (1,244 questions), and GPT-Audio 1.5 on 60 configurations (300 questions), for 34,049 model-question evaluations in total. The Task 2 tables use all available model outputs; partial models are marked and are not treated as directly comparable to the full 591-case evaluations.

4 Experiments and Results

4.1 RQ1: Cross-domain generality

Table 3 reports Task 1 accuracy for the seven models with outputs in all four datasets. Six models have complete coverage; Kimi-Audio has partial coverage in VoxConverse and MUStARD. The dataset-specific question counts in Table 2 are used as the denominator for each complete export. The main pattern is cross-domain variation rather than a universal ordering. MusicNet is lower for most models, but this should not be interpreted as evidence that music is intrinsically harder: its question mix contains substantially more exact numerical and label answers.

4.2 RQ2: Answer format and duration

A central methodological result is that Task 1 categories should not be interpreted as an ordered difficulty scale. The four datasets contain different proportions of numeric, Yes/No, and label answers. VoxConverse is 66.6% Yes/No, whereas MusicNet is 52.0% numeric and 13.3% label. More importantly, model accuracy differs sharply by answer type.

Across the seven common models, after averaging within each dataset before averaging across datasets, mean accuracy is 51.2% for Yes/No, 31.8% for numeric answers, and 11.6% for labels. Figure 2 shows the per-model pattern. Every model except DeSTA2.5-Audio exhibits a sizable Yes/No advantage over exact numerical answers.

This matters directly for interpreting the nine categories. In MusicNet, for example, Repetition and Source Tracking and Sequence Patterns are dominated by Yes/No questions and achieve around 52% mean accuracy, while Direct Recognition and First/Last Occurrence contain mostly numeric or label outputs and are much lower. The apparent category ordering therefore mixes reasoning structure with answer format.

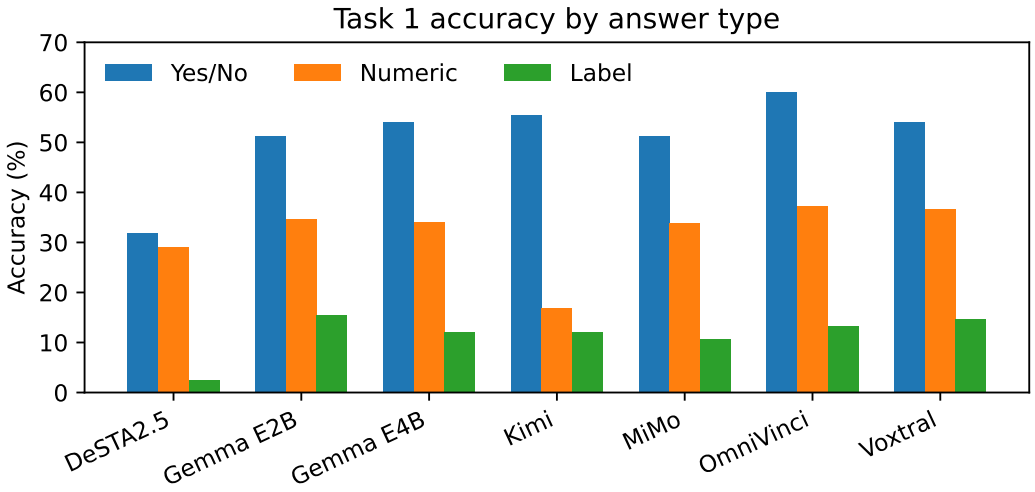

Figure 2: Task 1 answer-type accuracy, macro-averaged over the four datasets. The separation between binary and numerical answers makes raw category accuracy unsuitable as a difficulty hierarchy.

We consequently treat the taxonomy as *coverage*, not a ladder. A more informative analysis therefore asks whether the observed differences persist after accounting for answer type. The results show a consistent qualitative distinction: models are more reliable on binary verification and existence-style temporal judgments than on exact cardinality queries that require preserving an event reference and returning a precise integer.

4.3 Duration as a potential confound

Long audio is a known stressor (Luo et al., 2026; Ahia et al., 2025; Shao et al., 2026), so we evaluate MusicNet over 0–30s, 30s–2min, and 2–6min. Contrary to a simple long-context-collapse narrative, the seven fully evaluated models show no monotonic degradation over these bins. Across models, mean accuracy is approximately 30.7% in the 0–30s regime, 30.4% in 30s–2min, and 30.9% in 2–6min. This result suggests that duration is not the primary explanation for the observed failures. Indeed, Task 2 deliberately removes much of the long-context burden by using clips of at most 30 seconds, yet exact temporal composition remains extremely difficult. Duration is therefore a useful analysis dimension, but it is not sufficient to explain the relational reasoning failures.

4.4 RQ3: Multi-constraint temporal composition

Task 2 provides the controlled diagnostic for multi-constraint reasoning. Table 4 reports overall, answer-type, and question-level accuracy for all model outputs in the updated export. Exact integer accuracy remains low across the full-coverage models: it ranges from 0.0% to 5.7%, with Qwen2-Audio obtaining the highest integer accuracy at 5.7%. MiMo-Audio has the highest overall score at 22.1%, driven by 94.1% accuracy on the overlap question; its integer accuracy is only 3.9%. Intersection and union counting remain difficult across the suite.

The contrast between Q3 and the numerical questions must be interpreted against the class distribution. The overlap ground truth is 590 Yes and one No, giving an always-Yes majority baseline of 99.83%. Thus even 94.1% overlap accuracy does not by itself demonstrate temporal reasoning. The informative comparison is within the matched cases: models that answer the coarse overlap relation relatively often still perform poorly on the exact interval counts and set operations that use the same temporal constraints.

The intersection and union questions also test duplicate handling. If a target event lies in both I_1 and I_2, it contributes once to the union. Thus Q5 cannot be solved by independently counting Q1 and

Table 4: Task 2 accuracy (%). "Overall" pools all valid questions after the output normalization described in Section 3.6. "Integer" covers Q1, Q2, Q4, and Q5; "Yes/No" covers Q3. Outputs explicitly stating that the audio was unavailable are excluded. Empty or semantically unresolved outputs are incorrect. †: 249 configurations; ‡: 60 configurations.

Model	Overall	Integer	Yes/No	Q1 I_1	Q2 I_2	Q3 overlap	Q4 intersection	Q5 union
MiMo-Audio-7B	22.10	3.93	94.08	0.68	0.85	94.08	13.72	0.52
Qwen2-Audio-7B	17.02	5.67	62.44	8.12	7.28	62.44	5.58	1.69
Qwen3-Omni-Instruct	15.74	1.52	72.59	0.51	1.35	72.59	3.55	0.68
Gemini 3.8 Flash†	13.67	2.11	59.84	1.61	2.41	59.84	2.81	1.61
GPT-Audio 1.5‡	12.33	0.42	60.00	0.00	0.00	60.00	1.67	0.00
Phi-4-Multimodal	10.96	3.81	39.59	0.17	0.51	39.59	14.21	0.34
Qwen3-Omni-Thinking	3.71	1.30	14.73	1.79	1.41	14.73	1.47	0.42
Kimi-Audio-7B	3.32	2.54	6.43	1.02	1.02	6.43	7.78	0.34
Audio-Flamingo-3	1.96	2.16	1.18	0.51	0.51	1.18	7.45	0.17
AudSemThinker-QA-GRPO	1.62	0.68	5.41	0.51	0.34	5.41	1.52	0.34
Qwen2.5-Omni-7B	0.98	0.59	2.54	0.51	0.34	2.54	1.35	0.17
DeSTA2.5-Audio	0.41	0.47	0.17	0.00	0.17	0.17	1.52	0.17
Audio-Reasoner	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

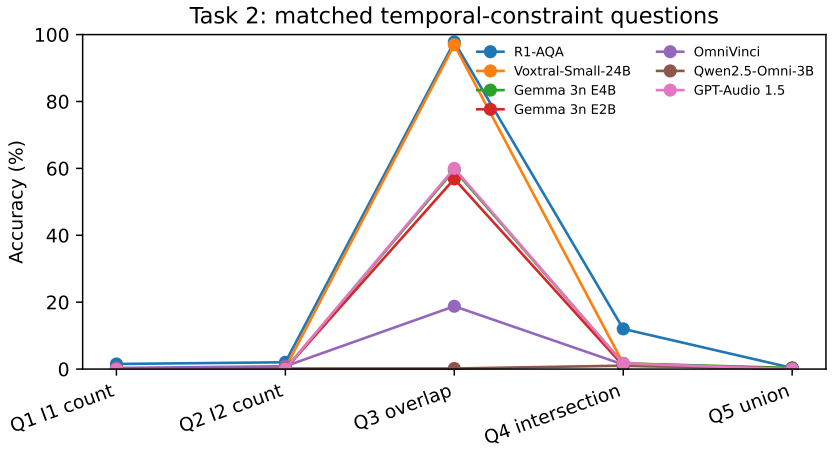

Figure 3: Task 2 performance. The large gap between interval-overlap recognition and exact counting is visible across models. The overlap question is highly imbalanced (590 Yes/1 No), so its raw accuracy is not interpreted as a standalone reasoning score.

Q2 and adding the results. The near-zero union accuracy suggests that models rarely maintain the identity of shared events through this composition.

4.5 Examples of temporal event-reference questions

The examples below are representative questions from the benchmark. They illustrate that the benchmark does not depend on one acoustic phenomenon. The reference can be a speaker occurrence, a sound occurrence, or a musical event; the downstream operation can be counting, comparison, or verification. Representative instances include speech questions such as "Does the second distinct speaker speak between the first time the first distinct speaker speaks and the first time the third distinct speaker speaks?", environmental questions such as "Does the sound of bells overlap with the sound of birds?", and music questions such as "After the first contrabass note played after the first violin note, how many cello notes are played?" Additional examples are provided in the supplement. The examples also clarify why exact integer answers are diagnostically useful. A binary question can be satisfied by identifying a relation; an integer question requires the model to preserve a temporal boundary and enumerate all qualifying occurrences under that boundary. The benchmark therefore exposes errors that can be hidden by a broad binary QA score.

4.6 Failure patterns

The results support three recurring failure patterns. First, *binary–numeric asymmetry*: models frequently answer Yes/No temporal checks correctly while failing exact counts. Second, *reference propagation failure*: performance collapses when the occurrence identified in one clause determines the region used by a later clause. Third, *set-composition failure*: even when models recognize two intervals, they rarely compute intersection or union counts reliably. We do not claim that these errors prove a single internal mechanism. They are behavioral signatures consistent with a failure to carry grounded temporal information through a downstream computation. The distinction is important because a model can fail at several points: event recognition, occurrence indexing, temporal localization, enumeration, or final aggregation. TEROS is designed to make these dependencies observable through matched question structures rather than through a single aggregate score.

5 Discussion

5.1 What TEROS establishes

Our results support a narrower conclusion than "ALMs cannot reason about time." Models solve many temporal questions, but performance depends strongly on how temporal information must be used. Binary verification is substantially more successful than exact numerical counting, and this gap persists in short clips where long-context retention is less likely to be the primary bottleneck. Task 2 makes the distinction explicit: relatively high overlap accuracy does not translate into reliable exact interval counts, intersection, or union counting. Temporal reasoning should therefore be evaluated as a composition of grounded operations rather than as a single classification skill.

5.2 Why cross-domain evaluation matters

TEROS applies the same event-reference abstraction to speakers, environmental sounds, and musical events. Although the source domains have different occurrence semantics, the underlying computation is shared, allowing researchers to separate logical-operation effects from acoustic-domain effects. MusicNet is lower for the common model set, but its answer distribution is also different; we therefore avoid attributing the difference to an intrinsic domain hierarchy. The executable program representation further supports matched future studies over occurrence indices, answer types, and temporal operators. An anonymous repository containing the benchmark and evaluation code is available at `https://anonymous.4open.science/r/After-The-Second-Bell-06AF/`.

6 Limitations

Task 1 is heterogeneous and its answer-type mixture is a confound; raw category accuracy should therefore not be read as a difficulty ranking. Task 2 is strongly imbalanced toward overlapping intervals, and the updated model export has unequal coverage for Gemini 3.8 Flash and GPT-Audio 1.5. Source-specific occurrence conventions may not match every possible definition of an occurrence, and the benchmark inherits the granularity and quality of its source annotations. Finally, exact-match evaluation can penalize differently formatted but semantically correct outputs.

7 Conclusion

We introduced TEROS, a benchmark for temporal event-reference reasoning across speech, environmental sound, and music. The central result is a measurable gap between recognizing temporal relations and using a grounded occurrence as an operand in exact downstream computation. Task 1 shows strong answer-type effects that caution against category-based difficulty ladders, while Task 2 provides a controlled diagnostic in which exact intersection and union counting remain difficult despite higher coarse overlap performance.

AI Use Statement. Large language models were used during benchmark construction for question generation and validation, and for assistance with analysis and manuscript editing. All benchmark-

generation procedures, deterministic ground-truth computations, experimental results, and scientific claims were independently checked by the authors.

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.

Debarpan Bhattacharya, Apoorva Kulkarni, and Sriram Ganapathy. Benchmarking and confidence evaluation of lalms for temporal reasoning, 2025. URL https://arxiv.org/abs/2505.13115.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an *Obviously* perfect paper). In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619–4629, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1455. URL https://aclanthology.org/P19-1455/.

Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. Spot the Conversation: Speaker Diarisation in the Wild. In *Interspeech 2020*, pp. 299–303, 2020. doi: 10.21437/Interspeech.2020-2337.

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.

Luca Fredianelli, Francesco Artuso, Geremia Pompei, Gaetano Licitra, Gino Iannace, and Andac Akbaba. Environmental noise dataset for sound event classification and detection. *Scientific Data*, 12(1):1712, oct 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-05991-w.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Evuru, S. Ramaneswaran, S. Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Compa: Addressing the gap in compositional reasoning in audio-language models, 2024. URL https://arxiv.org/abs/2310.08753.

Apoorva Kulkarni, Kaousheik Jayakumar, Sreyan Ghosh, Utathya Aich, Ramani Duraiswami, and Dinesh Manocha. Tempo: Temporally-grounded multi-task post-training for large audio-language models, 2026. URL https://arxiv.org/abs/2608.29999.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, Siddhi Patil, Soham Deshmukh, Lasha Koroshinadze, Yao Liu, Leibny Paola Garcia Perera, Eleni Zanou, Themos Stafylakis, Joon Son Chung, David Harwath, Chao Zhang, Dinesh Manocha, Alicia Lozano-Diez, Santosh Kesiraju, Sreyan Ghosh, and Ramani Duraiswami. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence, 2025. URL https://arxiv.org/abs/2508.13992.

Daeyong Kwon, Qiyu Wu, Shinobu Kuriya, Junghyun Koo, Shuyang Cui, Zhi Zhong, Wei-Hsiang Liao, Hiromi Wakaki, and Yuki Mitsufuji. Mustbench: Benchmarking and advancing temporal grounding in music llms, 2026. URL https://arxiv.org/abs/2605.29300.

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.

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.

Luoyi Sun, Xiao Zhou, Zeqian Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spotsound: Enhancing large audio-language models with fine-grained temporal grounding, 2026. URL https://arxiv.org/abs/2604.13023.

John Thickstun, Zaid Harchaoui, and Sham M. Kakade. Learning features of music from scratch. In *International Conference on Learning Representations (ICLR)*, 2017.

A SUPPLEMENTARY MATERIAL

A.1 A. FULL TASK 1 QUESTION DISTRIBUTION

Table 5: Task 1 question counts by category and dataset in the current exports. A dash indicates that the category is not supported by the benchmark construction.

Category	VoxConverse	MUStARD	DataSED	MusicNet
Direct Recognition	1,160	1,993	2,502	2,442
Between Event A and Event B	903	2,007	2,156	2,428
Before / After & Conditional Counting	1,013	1,974	2,328	2,489
Sequence Patterns	1,078	1,743	2,468	2,412
Nested Temporal	1,124	2,062	1,923	2,337
Repetition & Source Tracking	1,215	2,270	2,315	2,402
First / Last Occurrence	1,076	1,708	2,506	2,149
Comparative Counting	1,013	1,545	1,820	2,558
Concurrent / Overlapping	923	–	2,166	2,503
Total	**9,516**	**15,318**	**20,191**	**22,441**

A.2 B. ANSWER-TYPE DISTRIBUTION

Across the four current Task 1 exports, there are 32,775 Yes/No questions, 31,438 integer questions, 3,164 label questions, and 89 ordinal questions, for 67,466 questions in total. Ordinal questions occur only in MUStARD. The ordinal stratum is therefore reported separately and interpreted descriptively rather than as a general Task 1 performance estimate.

A.3 C. ADDITIONAL EXAMPLES

A.4 D. MUSICNET CATEGORY RESULTS

The category means in Figure 5 are: Direct Recognition 17.0%, Between Event A/B 20.7%, Before/After + Conditional Counting 22.8%, Sequence Patterns 51.0%, Nested Temporal 24.3%, Repetition/Source Tracking 51.6%, First/Last 17.0%, Comparative Counting 34.8%, and Concurrent/Overlap 32.7%. These values are included for completeness; the main paper uses answer-type-stratified analysis for interpretation.

A.5 D. TASK 2 CONSTRUCTION AND SANITY CHECKS

The 591 valid configurations are distributed across 200 short MusicNet clips. Each case produces five questions. The overlap question has 590 Yes and one No. This construction is therefore not suitable for evaluating binary overlap classification as a standalone task; a future balanced release should include approximately equal positive and negative cases.

The full-containment rule is applied to all interval membership tests. For an event X and interval $I = [s, e]$, membership requires $X_{\text{start}} \geq s$ and $X_{\text{end}} \leq e$. Intersection membership requires both I_1 and I_2 membership conditions. Union membership requires at least one condition, and a shared event is counted once.

A.6 E. TASK 2 ANSWER NORMALIZATION

The updated evaluation treats answer formatting separately from answer correctness. Integer answers are accepted when the response gives an unambiguous integer or an unambiguous number word; Yes/No answers are accepted when the response clearly resolves to one of the two labels. Four empty responses are therefore counted as incorrect. We exclude 447 responses that explicitly state that the audio was unavailable (23 from MiMo-Audio-7B and 424 from Qwen3-Omni-Thinking). No Task 2 ground-truth item requires an ordinal answer.

Table 6: Number of Task 1 questions by reasoning category and dataset.

Category	Vox	MUS	DataSED	MusicNet
Direct Recognition	1,160	1,993	2,502	2,442
Between Event A and Event B	903	2,007	2,156	2,428
Before / After & Conditional Counting	1,013	1,974	2,328	2,489
Sequence Patterns	1,078	1,743	2,468	2,412
Nested Temporal	1,124	2,062	1,923	2,337
Repetition & Source Tracking	1,215	2,270	2,315	2,402
First / Last Occurrence	1,076	1,708	2,506	2,149
Comparative Counting	1,013	1,545	1,820	2,558
Concurrent / Overlapping	923	–	2,166	2,503
Total	**9,516**	**15,318**	**20,191**	**22,441**

Table 7: Additional examples sampled from the benchmark.

Pattern	Example question
VoxConverse / repetition	"Does the second distinct speaker return to speak after the third distinct speaker begins speaking?"
VoxConverse / comparative	"How many more speaking turns does the second distinct speaker have than the third distinct speaker?"
MUStARD / first-last	"Does the first distinct speaker speak after the second distinct speaker finishes speaking for the last time?"
DataSED / repetition	"How many distinct sounds are heard more than once in the recording?"
DataSED / overlap	"Does the sound of bells overlap with the sound of birds?"
MusicNet / repetition	"Is note E3 played more than once in the recording?"
MusicNet / first occurrence	"Which note is played first by the violin?"
MusicNet / comparison	"Is the cello played more often than the contrabass?"
MusicNet / overlap	"How many contrabass notes overlap with at least one violin note?"

A.7 G. Model coverage

A.8 H. Validation and release protocol

The benchmark generator operates on structured annotation records. Natural-language questions are produced only when the required event occurrences exist and the answer is uniquely determined. Candidate questions are independently checked by multiple language-model validators. We retain constructions supported by at least two validators and send a stratified subset and validator disagreements to human review. The release will include the normalized answer, dataset identifier, question category, temporal program, and relevant construction metadata. The evaluation interface exposes only the audio and question under the standard condition.

A.9 I. Recommended future controlled subsets

The strongest next step is to release matched question pairs with the same underlying computation and answer type but different linguistic realization. For example, a numerical query and an equivalent Yes/No query can be generated from the same event-reference structure. Additional balancing should ensure approximately equal positive and negative overlap cases in Task 2. Finally, matched occurrence indices (first, second, third, and later) would permit a direct test of whether models lose temporal precision as the reference occurrence becomes less salient.

Table 8: Model coverage in the updated Task 2 export. Full coverage means 591 configurations (2,955 questions); partial coverage is reported in the table.

Model	Task 1	Task 2
DeSTA2.5-Audio	Yes	Yes
Gemma 3n E2B	Yes	No
Gemma 3n E4B	Yes	No
Kimi-Audio	Yes	Yes
MiMo-Audio	Yes	Yes
OmniVinci	Yes	No
Voxtral-Small-24B	Yes	No
Gemini Flash 3.8	Partial	Partial (249 cases)
GPT-Audio 1.5	Partial	Partial (60 cases)
AudSemThinker-QA-GRPO	No	Yes
Audio-Flamingo-3	No	Yes
Audio-Reasoner	No	Yes
Phi-4-multimodal-instruct	No	Yes
Qwen2-Audio-7B	No	Yes
Qwen2.5-Omni-7B	No	Yes
Qwen3-Omni-30B-A3B-Instruct	No	Yes
Qwen3-Omni-30B-A3B-Thinking	No	Yes