MuLA-Bench: A Multilingual Long-Form Audio Understanding Benchmark via Multi-Tier Auditing

Anonymous authors
Paper under double-blind review

Abstract

Long-form audio performance is often summarized by context length and aggregate accuracy, obscuring how language, evidence, and task jointly shape difficulty. We introduce **MuLA-Bench**, a challenging benchmark of 5,038 open-ended questions over 1,732 in-the-wild recordings totaling 1,341 hours, covering 16 languages and eight domains. A balanced Language$\times$Domain semantic track supports controlled comparisons, while a complementary acoustic track preserves naturally occurring non-speech evidence. Evidence-grounded generation, shortcut checks, and language-expert review provide auditable questions without translating a shared source set or injecting target sounds. We evaluate eleven audio-language models and conduct targeted diagnostics of language, evidence, and temporal performance. Language rankings change across domains and tasks; acoustic–semantic performance gaps vary with the requested operation; and later evidence is associated with lower temporal accuracy but nearly flat long-range accuracy in a four-hosted-model cohort. Long-range retrieval is comparatively strong, while precise clock alignment and factual grounding of natural acoustic events remain fragile. MuLA-Bench thus exposes conditional failure patterns that a single long-context score does not capture. Anonymous code and benchmark artifacts are included in the supplementary material.

1 Introduction

Large audio-language models (LALMs) increasingly process lectures, interviews, podcasts, and other recordings that extend far beyond the short clips used in earlier audio benchmarks. Evaluation has followed this shift from broad perception toward reasoning and long-context comprehension (Yang et al., 2024; Wang et al., 2025; Sakshi et al., 2025; Ma et al., 2025; Ahia et al., 2025; Ye et al., 2026; Huang et al., 2026). Yet admissible context length is not the same as reliable understanding. A model may retrieve a distant spoken fact, miss a nearby natural sound, or fail to locate a specified event on the recording clock. Long-form audio competence is therefore not a single scalar property: what a model can recover depends on the evidence it must use and the operation the question requires. This distinction matters more as recordings grow longer, because evidence type, evidence position, and requested operation vary simultaneously. A benchmark that reports only a context limit or aggregate score can therefore hide qualitatively different failure modes behind the same number.

Multilingual evaluation adds a second difficulty. Benchmark design faces a tension between *control* and *authenticity*. Translated variants, constructed task suites, or injected events can make conditions easier to compare, but they standardize part of the linguistic or acoustic variation present in native recordings. In-the-wild benchmarks preserve that variation, yet language, domain, task mix, and evidence source can become entangled (Gao et al., 2025; Iyer et al., 2026; Ye et al., 2026; Huang et al., 2026). The resulting design question is not simply how to make audio longer, but how to control the comparisons that matter without replacing the original linguistic content or manufacturing the acoustic evidence. In particular, multilingual analysis needs enough structure to separate language from domain and task mixture, while acoustic evaluation should not obtain that structure by inserting the very sounds it aims to measure.

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.

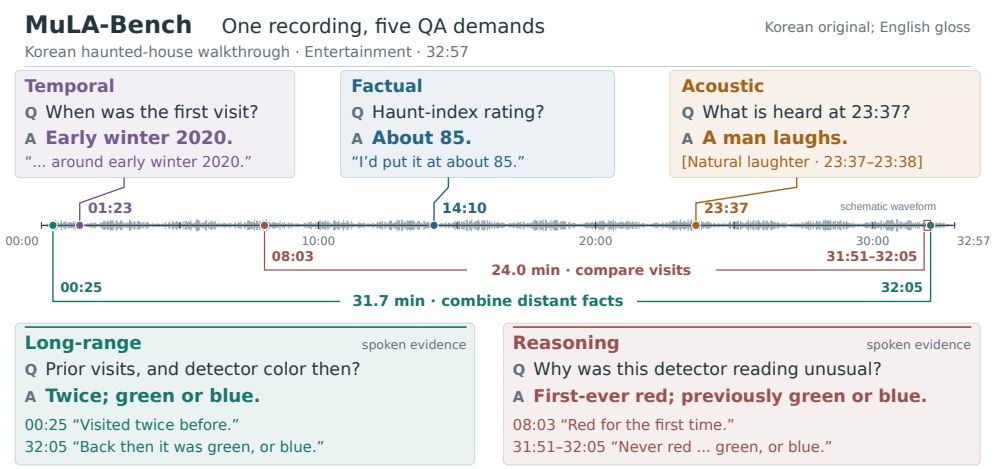


Figure 1: One recording supports multiple QA demands. English glosses of a Korean example show spoken facts, temporal references, distant evidence, and a natural acoustic event. Colors link questions to evidence positions; the waveform is schematic. Models receive audio and source-language questions only.

MuLA-Bench addresses this tension by controlling *question allocation rather than source content*. It contains 5,038 open-ended questions over 1,732 native, in-the-wild recordings spanning 16 languages and eight domains (Figure 1). The semantic track fixes both the number and task composition of questions in every Language×Domain cell, enabling controlled conditional comparisons while leaving the underlying recordings unchanged. The complementary acoustic track takes the opposite approach where forced balance would distort the phenomenon: it follows verified naturally occurring events rather than translating questions or inserting sounds. Both tracks use localized evidence, decoupled question and answer generation, automatic shortcut and grounding checks, and final language-expert review. The asymmetry between the two tracks is deliberate: semantic questions are balanced where allocation can be controlled without changing source content, whereas acoustic questions follow verified natural availability where forcing equal coverage would require fabrication or replacement.

This design exposes a consistent picture: long-form audio performance is *conditional rather than scalar*. Relative language performance changes with domain and task; the acoustic–semantic gap changes with the requested operation; and event ordering and precise clock localization show model-dependent performance gaps. The benchmark remains challenging: the best overall and acoustic accuracies are 73.48% and 55.59%, respectively. These findings argue for reporting conditional competence rather than reducing long-audio understanding to one language ranking, one modality gap, or one length effect. Our contributions are therefore (i) a native multilingual long-form benchmark that combines controlled semantic allocation with natural acoustic evidence; (ii) an evidence-grounded, multi-tier construction and auditing pipeline; and (iii) an eleven-model study that characterizes conditional language, evidence, and temporal failure patterns.

2 Related Work

From general audio models to long-form evaluation. General-purpose LALMs have progressively expanded from speech-centric interfaces toward open-ended reasoning over speech, environmental sound, and music. Representative systems include LTU and SALMONN, which established early general-audio reasoning and instruction-following capabilities, followed by Qwen-Audio/Qwen2-Audio and Audio Flamingo, which broadened unified audio understanding and conversational interaction (Gong et al., 2024; Tang et al., 2024; Chu et al., 2023; 2024; Kong et al., 2024). Benchmarking evolved in parallel: AIR-Bench, AudioBench, Dynamic-SUPERB Phase-2, MMAU, and MMAR expand task breadth from perception to open-ended comprehension and reasoning across heterogeneous audio content (Yang et al., 2024; Wang et al., 2025; Huang et al., 2025; Sakshi et al., 2025; Ma et al., 2025). More recent benchmarks extend evaluation from short clips toward document-scale audio.

Table 1: **MuLA-Bench** vs. representative long-form and multilingual audio benchmarks.

	Scale		Source		Evidence		Evaluation	
Benchmark	**Dur.**	**Lang.**	**Native**	**Wild**	**Ground.**	**Natural Aco.**	**Open**	**L×D**
ADU-Bench (Gao et al., 2025)	Dialogue	9	Trans.	●	QA	●	✓	✗
SCENEBench (Iyer et al., 2026)	Short	4+	Mixed	✗	Task	●	●	✗
ChronosAudio (Luo et al., 2026)	≤20m	En.-centric	–	✗	Task	✗	●	✗
BLAB (Ahia et al., 2025)	∼51m	1	–	✓	QA	✓	●	✗
VoiceGiraffe (Ye et al., 2026)	∼55m	2	✓	✓	Explicit	✓	✗	✗
AudioSpan (Huang et al., 2026)	10m–2h^{+}	2	✓	✓	Explicit	●	●	✗
MuLA-Bench	**∼47m**	**16**	✓	✓	**Dual**	✓	✓	✓

Dur.: reported duration range or mean (MuLA-Bench: source mean). *Ground.*: evidence grounding; *Natural Aco.*: naturally occurring acoustic evidence. ✓ fully supported; ● partially/mixed; ✗ not supported. *Dual* denotes localized evidence from both speech and non-speech sound.

LongSpeech targets long-form speech tasks, ChronosAudio studies length generalization across multiple audio tasks, BLAB evaluates naturally long recordings averaging roughly 51 minutes, and VoiceGiraffe and AudioSpan move into hour-scale in-the-wild understanding (Yang et al., 2026b; Luo et al., 2026; Ahia et al., 2025; Ye et al., 2026; Huang et al., 2026). These efforts establish that duration itself is challenging, but longer inputs do not by themselves determine which sources of variation are controlled when comparing models or languages. This leaves a complementary benchmark-design question: once hour-scale context is available, can failures be localized to language, domain, evidence type, or requested operation rather than absorbed into a single long-context score?

Controlling benchmark variation without erasing it. Existing benchmarks obtain control in different ways, and these choices imply different trade-offs between comparability and natural variation. ADU-Bench constructs a nine-language dialogue evaluation with multilingual variants derived from English queries (Gao et al., 2025); SCENEBench targets controlled cross-linguistic and environmental audio-understanding settings (Iyer et al., 2026). AHa-Bench separately probes semantic, acoustic, and semantic–acoustic hallucinations, highlighting that spoken and non-speech evidence need not fail in the same way (Cheng et al., 2025). AudioSpan combines questions derived from native audio content with a complementary path that injects acoustic anchors, whereas VoiceGiraffe emphasizes curated real-world long-form recordings (Huang et al., 2026; Ye et al., 2026). **MuLA-Bench** occupies a different point in this design space: it keeps source recordings native and in the wild, controls *question allocation over Language×Domain cells*, and leaves acoustic coverage governed by verified natural event availability. This makes the benchmark deliberately non-parallel at the source level but controlled at the semantic question level, which is the comparison needed for the conditional analyses in this paper. Table 1 summarizes these design differences.

Open-ended and temporally grounded evaluation. Generative audio benchmarks increasingly complement fixed-choice accuracy with open-ended scoring. AIR-Bench, for example, evaluates open-ended audio interaction with LLM-based semantic grading, while AudioSpan uses rubric-based open-ended evaluation alongside multiple-choice questions (Yang et al., 2024; Huang et al., 2026). More broadly, LLM-based judging has become a common mechanism for evaluating free-form model outputs, motivating explicit attention to judge agreement and scoring contracts (Zheng et al., 2023). Temporal grounding is also becoming an explicit modeling and evaluation target. TimeAudio introduces temporal markers and absolute time-aware encoding for fine-grained localization (Wang et al., 2026); Listening with Time introduces long-form temporal data and LAT-Bench/LAT-Audio for precise temporal awareness (Shao et al., 2026); MOSS-Audio uses explicit time markers for time-aware question answering (Yang et al., 2026a); and Audio Flamingo Next grounds intermediate reasoning steps to timestamps through Temporal Audio Chain-of-Thought (Ghosh et al., 2026). **MuLA-Bench** brings these concerns together with route-specific open-ended scoring that distinguishes semantic agreement, completeness, event order, and recording-clock localization, allowing *event ordering to be analyzed separately from clock grounding*.

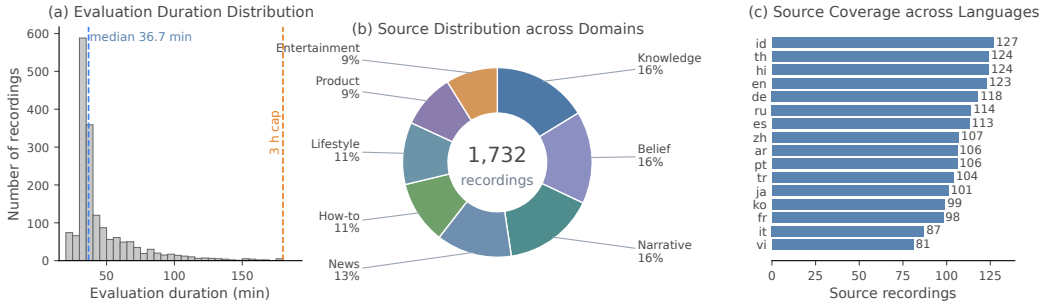


Figure 2: Dataset composition of MuLA-Bench. (a) Distribution of evaluation-audio duration, capped at three hours. (b) Distribution of source recordings across the eight content domains. (c) Source-recording coverage across the 16 target languages. Domain percentages use within-domain recording counts (Appendix B.2).

3 MuLA-Bench

MuLA-Bench crosses native languages and domains through two evidence tracks and four requested operations. Evidence-grounded generation and multi-tier auditing connect each answer to its source recording. Figures 2 and 3 summarize the released collection and construction workflow, respectively.

3.1 Audio Collection

Sources require usable original-language audio and matching-language subtitles. Recording-level language labels are supported by the source catalog and subtitle content, then inherited by the questions. Metadata flags possible dubbing, simultaneous interpretation, or language mismatch for inspection; confirmed invalid sources and known overlap are excluded. Occasional code-switching is not itself an exclusion criterion. Here, *native multilingual* denotes original source content, not translated benchmark variants.

GPT-5.6 Sol classifies each transcript by its dominant source-level function using eight domains: knowledge, how-to, narrative, entertainment, lifestyle, news, product, and belief. These domains and 16 languages define the target grid. Appendices A and B detail the language metadata, taxonomy, curation rules, and the distinction between source and question-level domain assignments.

3.2 Task Taxonomy

The evidence tracks distinguish *semantic* evidence—spoken facts and relations—from *acoustic* evidence, including non-speech and paralinguistic events such as laughter, applause, and music. Both tracks support four task families. **Factual/completeness** questions request explicitly recoverable information, including single facts, exhaustive lists, and multiple required slots. **Reasoning** requires inference over recording evidence rather than generic world knowledge. **Temporal localization** includes event ordering and physical time-point localization. **Long-range retrieval** uses late or widely separated evidence: semantic items require evidence beginning at least 10 minutes into the recording or, for multiple evidence points, a span of at least 10 minutes. Multiple points must be topically coherent. Acoustic long-range items use separately verified event evidence rather than the semantic packing rule.

Long-range can therefore include a single late fact rather than multi-hop reasoning. Event ordering also differs from metric clock localization; success on one does not establish the other. Formal definitions and edge cases are in Appendix A.

3.3 QA Construction

Semantic mining converts timestamped subtitle cues into atomic facts and rebases each fact to its supporting cue; weak alignments are discarded. TimeAudio (Wang et al., 2026) proposes acoustic

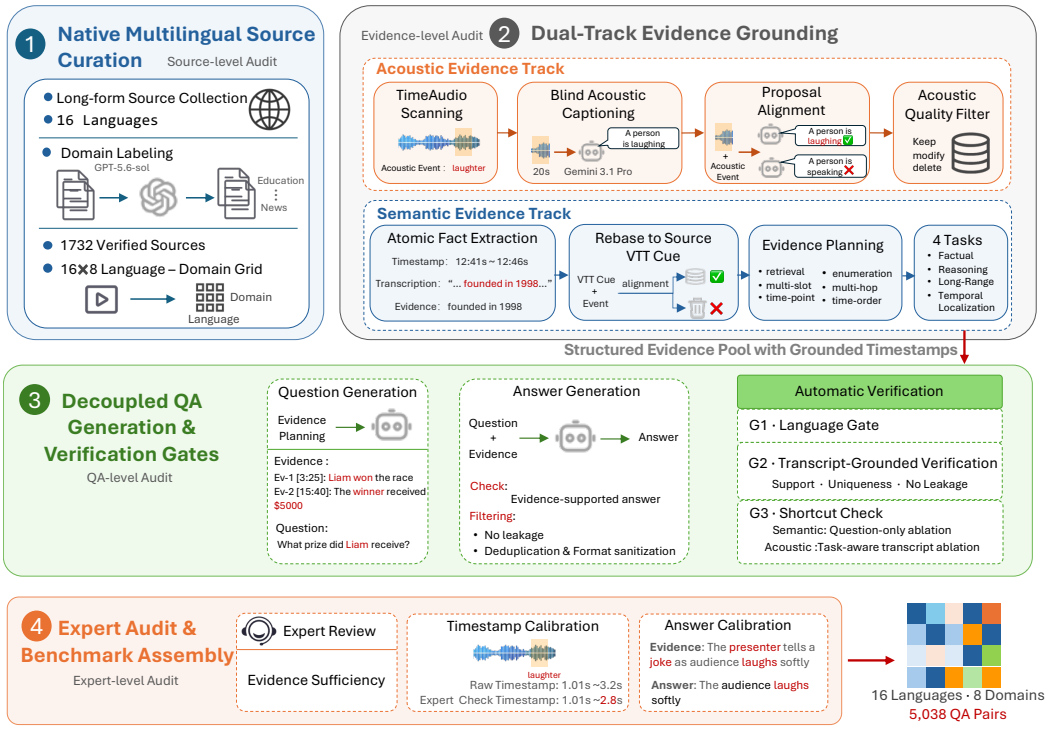


Figure 3: Construction pipeline of MuLA-Bench. Native multilingual sources are organized over the Language–Domain grid before semantic and acoustic evidence are grounded in parallel. Candidate QA pairs are generated from structured evidence, screened by automatic verification and shortcut tests, and finally audited against localized source evidence before release.

events across the full recording, with selected windows verified as described in Section 3.4. The evidence planner organizes retained semantic facts into retrieval, enumeration, multi-slot, reasoning, time-point, or event-order plans. Each plan determines the supporting evidence and requested operation before question wording is generated. Reasoning plans require at least two facts linked by a concrete causal, comparative, temporal, or arithmetic relation; semantic long-range plans additionally satisfy the evidence-position or span rule.

One call writes a question from its evidence plan without a prewritten answer; a separate fresh-context call derives the answer from the question and evidence. Localized evidence supports verification, expert replay, and analysis of position and separation. Models receive audio and questions only, without transcripts or evidence annotations. Figure 3 shows the workflow; Appendices B and E provide frozen settings and prompt templates.

3.4 Multi-Tier Quality Auditing

Auditing operates at the evidence, QA, and final expert-review stages. **Evidence-level checks.** Up to 12 selected windows per source undergo two-pass Gemini 3.1 Pro verification: proposal-hidden observation, then proposal alignment with keep, modify, or delete decisions. Visual context may assist verification, but retained events must be audible; cited acoustic events receive a further localized audibility check. Metric temporal items require precise timing. Verification focuses on selected windows following full-recording proposals.

QA-level checks. Candidates undergo language, grounding, uniqueness, leakage, question-only, and timestamp checks. For acoustic items, a solver attempts to answer using only the question and full transcript, allowing both sound clues named in the question and evidence from distant passages. An item is rejected when the solver produces a supported answer and a separate comparison call confirms a match to the reference under these shortcut rules. For acoustic reasoning, the cited sound must additionally contribute evidence needed for the target inference rather than merely accompany a

conclusion recoverable from speech. Repairable candidates receive one correction and reverification; fatal grounding errors are dropped.

Expert review. Every released question is reviewed by one paid expert for its language, with 16 experts in total. Reviewers inspect the question, answer, evidence, and localized audio to accept, repair, or replace candidates. They correct temporal references against playback; automatic timestamps are not final human ground truth. Exhaustive questions additionally receive a global transcript check. This full-set quality review complements the sampled evaluator audit in Section 4; procedures are in Appendix B.

3.5 Benchmark Statistics

MuLA-Bench contains **5,038 questions** over **1,732 recordings** totaling **1,341 hours**. The median source duration is 36.7 minutes and the mean is 46.5 minutes; 91.9% of recordings last at least 30 minutes. Figure 2 shows the duration distribution and source-recording coverage by domain and language.

The **semantic track** contains 3,840 questions across all 128 Language×Domain cells: exactly 30 per cell, comprising eight factual, eight reasoning, seven temporal, and seven long-range questions. Question allocation, rather than recording counts, is balanced. Equal task composition reduces task-mixture confounding and gives languages equal weight within a domain, and domains equal weight within a language. It supports the conditional analysis in Section 5.2 while retaining the original topics, speakers, and recording conditions.

The **acoustic track** contains 1,198 questions: 867 factual, 122 reasoning, 112 temporal, and 97 long-range. Verified event availability determines its uneven coverage; two cells have no usable acoustic items. Sparse cells are not filled with translated questions or inserted sounds. Its factual-heavy mixture motivates operation-specific comparisons rather than one aggregate semantic–acoustic gap.

Earliest evidence falls within the first 30 minutes for 3,758 questions (74.6%) and at 60 minutes or later for 267 (5.3%). These question counts include all tasks and differ from source-duration statistics. Appendix A gives detailed composition and coverage; Section 5.4 analyzes performance by evidence position.

4 Evaluation Protocol

Models and inputs. Table 2 evaluates eight hosted and three open-weight systems. Models receive the source-language question and canonical audio, without transcripts, video, or gold evidence. Four predefined post-3h items use a fixed later-stage three-hour segment. Local models receive the longest prefix permitted by their native context limits, starting from the canonical input. Missing, invalid, or noncompliant responses score zero under the fixed 5,038-item denominator. Scores therefore reflect complete-system performance including native context limits. Exact model settings, context limits, and coverage are reported in Appendix C.

Scoring and reliability. A binary rubric routes answers to single-fact, exhaustive-list, multi-slot, reasoning, time-point, or time-order criteria. GPT-5.6 Sol checks semantic agreement; deterministic guards enforce structural and temporal constraints. Time-point scoring requires both the correct event and clock alignment within ± 3 seconds, with explicit short-interval rules. Only final committed answers are scored. An expert audit spans all 16 languages and all six judge types, with expert–evaluator agreement on 795/844 responses (94.2%). Scoring and audit protocols appear in Appendix C.

Exhaustive lists require all members, multi-slot answers require every mandatory slot, and reasoning answers require the target conclusion and relation. Time-order scoring checks sequence rather than a clock value.

Diagnostic cohorts. The main results and evidence–task and language–domain diagnostics use all eleven models. Evidence-position profiles use Gemini 3.8 Flash, Gemini 3.7 Flash, Gemini 3.1 Pro, and Qwen3.5-Omni-Plus, separately from local models with fixed prefix limits. Covariate-adjusted and relative-position analyses are specified in Appendix C.9.

5 Results and Diagnostic Findings

5.1 Overall Results

Table 2: Accuracy (%) on the fixed benchmark and slice denominators. Missing responses score zero. Best and second-best results are bold and underlined.

		Evidence		Task			
Model	**Overall**	**Sem.**	**Aco.**	**Factual**	**Reason.**	**Temporal**	**Long-range**
Hosted / API models							
Gemini 3.8 Flash	**73.48**	**79.06**	**55.59**	**71.44**	<u>72.60</u>	**66.67**	<u>85.30</u>
Gemini 3.7 Flash	<u>73.14</u>	<u>78.96</u>	<u>54.51</u>	<u>70.91</u>	72.51	<u>66.17</u>	85.20
Qwen3.8-Omni-Flash	72.57	78.93	52.17	67.95	**74.26**	<u>66.17</u>	**85.90**
Gemini 3.1 Pro	61.39	72.27	26.54	56.58	68.85	39.68	83.99
Qwen3.5-Omni-Plus	60.92	66.77	42.15	61.92	60.38	40.67	80.16
Qwen3.5-Omni-Flash	46.47	52.34	27.63	48.02	44.24	27.28	65.56
Doubao seed 2.0 Lite	57.01	62.92	38.06	60.07	60.21	26.98	77.95
Muse Spark 1.2	33.78	36.67	24.54	37.28	33.60	13.19	48.24
Open-weight models							
MOSS-Audio-8B-Thinking	13.42	13.49	13.19	14.60	14.05	6.75	17.22
Qwen3-Omni-30B-A3B-Instruct	24.77	28.91	11.52	28.29	23.39	10.91	33.74
Audio Flamingo Next 8B	13.46	15.16	8.01	16.66	11.95	5.16	17.52

Gemini 3.8 Flash and 3.7 Flash lead with 73.48% and 73.14% overall accuracy, with Qwen3.8-Omni-Flash close behind at 72.57%. Qwen3.8 leads reasoning (74.26%) and long-range retrieval (85.90%), while Gemini 3.8 leads overall, semantic, acoustic, factual, and temporal accuracy. Similar totals can still mask different profiles: Gemini 3.1 Pro and Qwen3.5-Omni-Plus score 61.39% and 60.92% overall despite different temporal and acoustic-factual behavior. Aggregate accuracy therefore captures strength but not failure structure; complete evidence–task results are in Appendix D.

5.2 Language–Domain Performance Interactions

On the balanced semantic grid, we remove each model's language and domain marginals from its cell accuracy:

$$R_{m,\ell,d} = A_{m,\ell,d} - A_{m,\ell} - A_{m,d} + A_m. \tag{1}$$

Figure 4(a) averages these residuals over the eleven models. For example, Chinese has residuals of $+10.6$ pp in Entertainment and -16.0 pp in Product. Both deviations remain in the same direction under adjustment for task, scoring route, and evidence metadata ($+13.7$ and -16.1 pp, respectively), with the specified source-clustered simultaneous intervals excluding zero in both analyses. Appendix C.9 reports recording counts and the composition-adjusted sensitivity analysis. On this fixed benchmark, these departures show substantial Language×Domain variation in relative performance. Because both language and domain marginals have been removed, a large residual does not merely say that one language or one domain is globally harder. It identifies combinations whose observed accuracy departs from what those two marginals would predict additively on the balanced semantic grid.

The effect extends to relative rankings (Figure 4(b)). Against each model's semantic language ranking, Gemini 3.1 Pro still shows a negative Entertainment correlation ($\rho = -0.14$), while Qwen3.8-Omni-Flash ranges from $\rho = 0.28$ for Narrative to 0.73 for How-to and 0.72 for temporal questions. For Qwen3.8-Omni-Flash, the leading language in the semantic slice is Indonesian, whereas the Narrative slice is headed by French and the How-to slice by Japanese, illustrating that the reordering can be substantial. These are benchmark-conditioned rankings, not an intrinsic ordering of languages.

The two panels are complementary: residuals measure departures from the additive Language×Domain expectation, while rank correlations measure whether conditional slices preserve each model's semantic language ordering. They need not move together, and averaging residuals does not imply a shared interaction pattern across models (Appendix D.6). Likewise, MOSS's stable rankings do not imply strong absolute performance (13.42% overall). Together, the panels separate

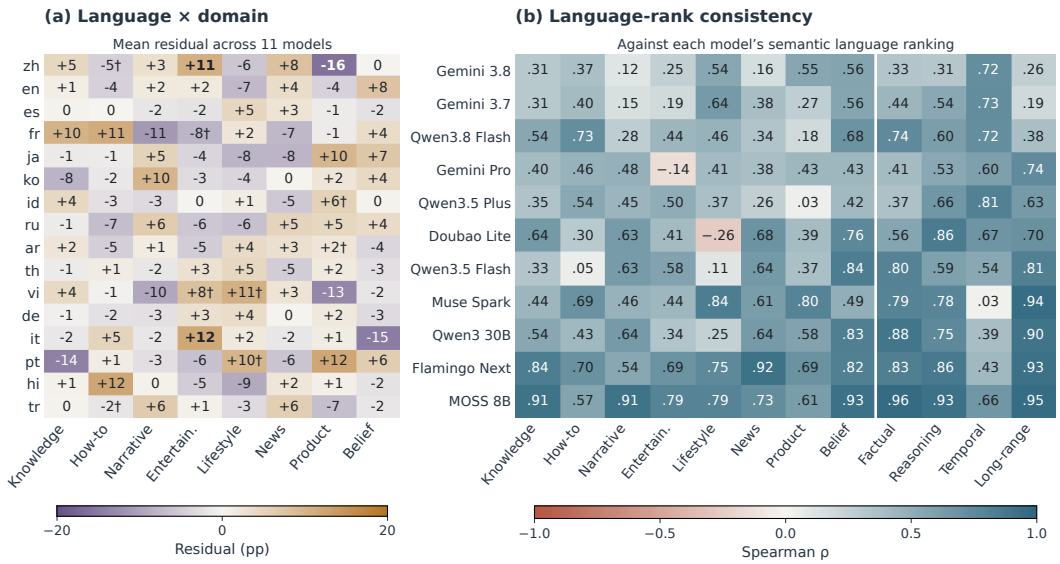


Figure 4: Language–Domain performance variation across eleven models. (a) Mean semantic Language×Domain residual (pp); bold cells have source-multiplier-bootstrap simultaneous 95% intervals excluding zero among the 120 eligible cells. Daggers mark cells with fewer than five recordings. (b) Spearman correlation between conditional and semantic language rankings; higher ρ denotes more stable rankings, not higher accuracy.

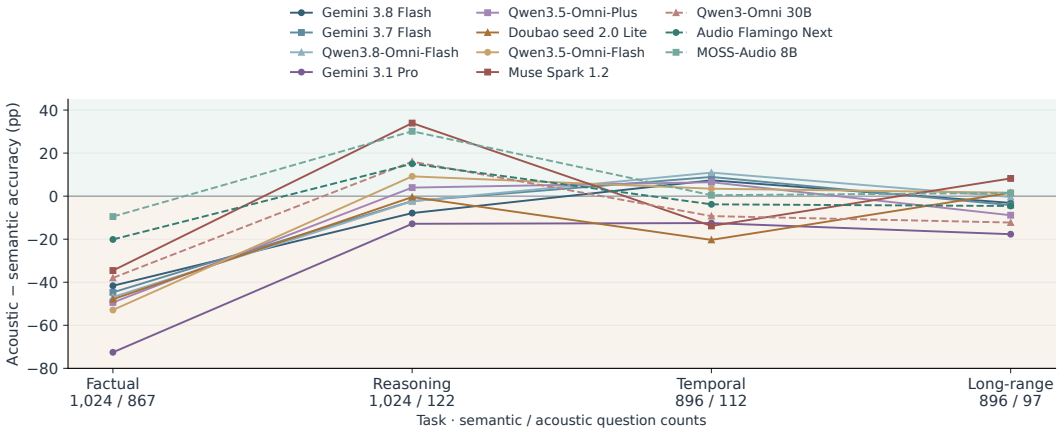


Figure 5: Evidence gaps across operations for eleven models. $\Delta = 100(\mathrm{Acc}_{\mathrm{aco}} - \mathrm{Acc}_{\mathrm{sem}})$; negative values favor semantic evidence. Counts denote semantic/acoustic items; lines connect task categories within each model.

absolute model strength, benchmark-conditioned language ordering, and local Language×Domain interaction.

5.3 Operation-Dependent Evidence Gaps

Figure 5 shows acoustic-minus-semantic accuracy for all eleven models. Factual gaps are uniformly negative (-9.5 to -72.5 pp), whereas reasoning gaps reverse for six models, including MOSS ($+30.1$ pp) and Audio Flamingo Next ($+15.1$ pp). Temporal gaps are mixed; long-range gaps are generally smaller. Each gap compares the corresponding question sets, with counts in the figure. Gemini 3.1 Pro scores 89.84% on semantic factual items but 17.30% on acoustic ones; MOSS scores 10.84% on semantic reasoning and 40.98% on acoustic reasoning. This operation-dependent pattern

is inconsistent with a single additive "acoustic penalty", motivating an evidence×operation profile alongside overall accuracy.

Gold-localized crops further probe acoustic-factual difficulty. On 200 questions, three hosted models score 38.0–44.5% with crops (median 31.1 s); their mean changes by -0.33 pp from full input (95% CI $[-5.10, 4.41]$). Of full-input errors, 80.1% persist; all three models fail on 64/200 questions under both conditions. Substantial difficulty thus remains after gold localization (Appendix D.9).

5.4 Evidence Position

Evidence position matters, but not uniformly across operations. In the four hosted models specified in Appendix D.7, later evidence is associated with lower temporal accuracy but nearly flat long-range accuracy. Across the 0–10 through 60–120 minute bins, temporal accuracy falls from 60.4% to 41.3%, whereas long-range accuracy stays between 81.8% and 85.0%. Using all 2,001 temporal and long-range questions, including the sparse ≥ 120-minute tail, the Spearman correlations with earliest evidence position are -0.179 and -0.008, respectively. Their difference is -0.171 (recording-cluster bootstrap 95% CI $[-0.260, -0.082]$).

The task-by-position interaction remains significant under covariate adjustment ($p = 0.00345$, recording-clustered standard errors). Late content retrieval can thus remain comparatively robust even as temporal-task accuracy declines. These are descriptive absolute-position associations; profiles and alternative specifications appear in Appendix D.7 and Figure A5. The contrast persists for the mean of all eight hosted models (adjusted $p = 0.00223$; Table A18). In the four-model cohort, adjusted slopes are -9.88 pp/30 min for event ordering (95% CI excludes zero) and $+0.28$ for clock localization (CI includes zero). Thus, the temporal decline is concentrated in ordering, distinguishing clock-localization difficulty from sensitivity to later evidence.

The separate eleven-model adjusted analysis also leaves acoustic factual and semantic temporal questions well below semantic factual questions (OR $= 0.093$ and 0.150). A +1 SD shift in log duration or normalized earliest position has an OR of 0.787 or 0.866, respectively. Relative evidence span is negatively associated with success (OR $= 0.918$, 95% CI $[0.870, 0.969]$), whereas evidence count is close to neutral (OR $= 1.032$, 95% CI $[0.987, 1.078]$). These conditional contrasts include native context limits and differ from the hosted comparison. Distance should be interpreted jointly with the requested operation; full coefficients are in Table A17.

5.5 Event Ordering versus Time-Point Localization

Temporal questions separate event ordering from alignment to the physical recording clock. On the 387 time-order and 568 time-point slices, Gemini 3.1 Pro, Doubao seed 2.0 Lite, and Muse Spark 1.2 show gaps of 36.6, 41.3, and 25.6 pp, respectively (Table 3). Nine of eleven models score lower on time-point localization, but the three leading systems largely close or reverse the gap (-3.8 to $+3.9$ pp; Table A14). These are distinct question sets, not paired tests conditioned on recognition of the same event. A Turkish case illustrates the clock errors: when asked to locate a specified 63-to-69 stitch-count change, models report 15:57, 16:26, or 14:32 instead of 16:06. The question supplies the event description (Appendix F).

Table 3: Order vs. time-point accuracy (%).

Model	Order	Point
Gemini 3.1 Pro	59.7	23.1
Doubao seed 2.0 Lite	49.4	8.1
Muse Spark 1.2	26.6	1.1

6 Conclusion

MuLA-Bench is a challenging benchmark for multilingual long-form audio understanding, combining balanced semantic questions with verified natural acoustic evidence across 16 languages and eight domains. Eleven-model results show that strong distant-fact retrieval can coexist with poor acoustic factual grounding and clock localization. Language rankings and evidence gaps also change with the evaluation slice. The benchmark's evidence annotations, auditing pipeline, and routed scoring support diagnosing these differences, beyond ranking systems by a single aggregate or context-length score.

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.

Alibaba Cloud Model Studio. qwen3.8-omni-flash model info. https://www.alibabacloud.com/help/en/model-studio/qwen3-8-omni-flash, September 2026. Official model documentation, updated September 18, 2026.

ByteDance Seed. Seed2.0. https://seed.bytedance.com/zh/seed2/, 2026. Official model release page; includes Seed2.0 Lite.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. Aha-bench: Benchmarking audio hallucinations in large audio-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1169afa8ad13c98239f0b666b099e159-Abstract-Datasets_and_Benchmarks_Track.html.

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

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

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4763–4784, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.237. URL https://aclanthology.org/2025.acl-long.237/.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, et al. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026. URL https://arxiv.org/abs/2604.10905.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. Listen, think, and understand. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/510d0935b543a29d686f93fa52d1c288-Abstract-Conference.html.

Google DeepMind. Gemini 3.1 pro: Model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026a. Published 19 February 2026.

Google DeepMind. Gemini 3.7 flash: Model card. https://deepmind.google/models/model-cards/gemini-3-7-flash/, August 2026b. Published 13 August 2026.

Google DeepMind. Gemini 3.8 flash: Model card. https://deepmind.google/models/model-cards/gemini-3-8-flash/, September 2026c. Published 2 September 2026.

Chien-yu Huang et al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce1a7774429d5c06cc744ff4260a8e9f-Abstract-Conference.html.

Wen Huang, Yunfei Chu, Meng Gao, Haolin He, and Jin Xu. AudioSpan: Spanning the duration and depth of audio comprehension, 2026. URL https://arxiv.org/abs/2608.26431.

Laya Iyer, Angelina Wang, and Sanmi Koyejo. SCENEBench: An audio understanding benchmark grounded in assistive and industrial use cases. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7123–7137, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.335. URL https://aclanthology.org/2026.eacl-long.335/.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*, 2024. URL https://arxiv.org/abs/2402.01831.

Kaiwen Luo, Liang Lin, Yibo Zhang, Moayad Aloqaily, Jialiang Tao, Dexian Wang, Zhenhong Zhou, Junwei Zhang, Kun Wang, Li Sun, and Qingsong Wen. ChronosAudio: A comprehensive long-audio benchmark for evaluating audio-large language models, 2026. URL https://arxiv.org/abs/2601.04876.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/610a7d6507d55be70c6d13d0b663227d-Abstract-Datasets_and_Benchmarks_Track.html.

Meta AI. The multimodal intelligence of muse spark 1.2. https://research.meta.ai/blog/multimodal-intelligence-of-muse-spark-1-2/, August 2026. Published 20 August 2026.

OpenAI. Gpt-5.6 system card. https://deploymentsafety.openai.com/gpt-5-6, July 2026. Published 9 July 2026.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026. URL https://arxiv.org/abs/2604.15804.

Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d36f208919582785db965fe648b9fe59-Abstract-Conference.html.

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

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/476ab8f369e489c04187ba84f68cfa68-Abstract-Conference.html.

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

Hualei Wang, Yiming Li, Shuo Ma, Hong Liu, and Xiangdong Wang. Listening between the frames: Bridging temporal gaps in large audio-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26233–26241, 2026. doi: 10.1609/aaai.v40i31.39827. URL https://ojs.aaai.org/index.php/AAAI/article/view/39827.

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. URL https://arxiv.org/abs/2509.17765.

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

Fei Yang, Xuanfan Ni, Renyi Yang, Jiahui Geng, Qing Li, Chenyang Lyu, Yichao Du, Longyue Wang, Weihua Luo, and Kaifu Zhang. LongSpeech: A scalable benchmark for transcription, translation and understanding in long speech, 2026b. URL https://arxiv.org/abs/2601.13539.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.109. URL https://aclanthology.org/2024.acl-long.109/.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Reproducibility Statement

Appendix A specifies the task definitions and allocation; Appendix B records construction and expert-review procedures. Appendix C gives model identifiers, input limits, scoring contracts, and statistical specifications. Full model breakdowns and sensitivity analyses are in Appendix D; prompt specifications and source templates are in Appendix E and the supplementary artifact.

AI Use Statement

Generative AI systems were used as components of benchmark construction and evaluation, including evidence mining and planning, question and reference-answer generation, acoustic-event verification, and answer judging, as described in the main text and appendices. Generative AI tools were also used to assist manuscript editing. All AI-assisted benchmark artifacts used in the released evaluation are subject to the automatic and/or expert verification procedures described in this paper. The authors reviewed the resulting benchmark, analyses, and manuscript and take responsibility for their contents.

Appendix Overview

The appendix documents the benchmark, construction process, evaluation, and full results. Section titles and page references are clickable.

Section	Contents	Page
A. Dataset	Languages and domains; task definitions; annotated examples; composition and coverage	13
B. Construction	Source curation; evidence mining; QA and shortcut checks; expert review; frozen settings	17
C. Evaluation	Model settings and input limits; six scoring routes; timing rules; human audits; statistical methods	20
D. Results	Eleven-model breakdowns; accuracy intervals; residual variation; position and gold-crop diagnostics	25
E. Prompts	Construction specifications; verification templates; inference wrapper; six judge instructions	34
F. Cases and release	Illustrative model outputs; source provenance and release scope; limitations	44

A Dataset Specification and Composition

MuLA-Bench includes 16 languages, eight domains, and two evidence tracks. The semantic allocation is balanced at the question level; source counts and natural acoustic event availability need not be balanced.

A.1 Language Coverage and Metadata

Table A1 summarizes linguistic and script coverage for the 16 target languages.

Table A1: Linguistic and script coverage of the 16 target languages.

Language	ISO	Language family	Primary script(s)
English	en	Indo-European (Germanic)	Latin
Chinese	zh	Sino-Tibetan (Sinitic)	Han
Hindi	hi	Indo-European (Indo-Aryan)	Devanagari
Spanish	es	Indo-European (Romance)	Latin
French	fr	Indo-European (Romance)	Latin
Arabic	ar	Afro-Asiatic (Semitic)	Arabic
Portuguese	pt	Indo-European (Romance)	Latin
Russian	ru	Indo-European (Slavic)	Cyrillic
Indonesian	id	Austronesian	Latin
German	de	Indo-European (Germanic)	Latin
Japanese	ja	Japonic	Kanji + kana
Vietnamese	vi	Austroasiatic	Latin
Turkish	tr	Turkic	Latin
Korean	ko	Koreanic	Hangul
Italian	it	Indo-European (Romance)	Latin
Thai	th	Kra-Dai	Thai

A.2 Domain Taxonomy and Allocation

The eight domains are Knowledge & Education, Skill & Tutorial, Narrative & Storytelling, Entertainment & Performance, Lifestyle & Physical Activities, News & Commentary, Product & Review, and Belief & Spiritual Culture. The controlled semantic allocation is

$$n_{\text{sem}}(\ell, d) = 30, \qquad |\mathcal{Q}_{\text{sem}}| = \sum_{(\ell,d)\in\mathcal{G}} n_{\text{sem}}(\ell, d) = 16 \times 8 \times 30 = 3,840. \tag{2}$$

Each cell contains eight factual, eight reasoning, seven temporal, and seven long-range questions. The acoustic track contains 1,198 verified questions without an equal-cell constraint.

A.3 Task Definitions and Edge Cases

Factual/completeness. This task family covers requests for explicitly recoverable information, including entities, values, events, and requested sets of facts. The family is heterogeneous by design: most items require a single grounded fact, while a smaller subset requires an exhaustive list or multiple mandatory slots. We therefore use *factual/completeness* as a task-family label rather than as a claim that all items test exhaustive recall.

Reasoning. Reasoning questions require inference over evidence in the recording, including causal, comparative, procedural, and multi-evidence relations. Questions that can be answered from generic world knowledge without the recording are rejected during construction.

Temporal localization. Temporal questions include recording-clock localization and event ordering. Clock localization requires both the correct event and its physical position in the recording. Ordering distinguishes *mention order* (when events are mentioned in the audio) from *event chronology* (when they occur in the story or world). The requested reference frame governs the answer; a calendar year in the narrative is not a recording timestamp. Task-family labels and judge routes are stored separately in the frozen release. Scoring follows the stored route consistently across model outputs.

Long-range retrieval. For semantic questions, long-range status follows Eq. 3 and therefore requires either late evidence or a sufficiently wide separation among topically coherent evidence points. This definition is intentionally based on evidence position rather than nominal recording duration. Acoustic long-range questions use verified acoustic-event evidence and are analyzed separately from the semantic packing rule.

Formal semantic long-range rule. For evidence timestamps $\mathcal{E}_q = \{t_1, \ldots, t_m\}$ in seconds,

$$\mathrm{LR}(q) = \begin{cases} \mathbf{1}[t_1 \geq 600], & m = 1, \\ \mathbf{1}[(\max_i t_i - \min_i t_i \geq 600) \ \vee \ (\min_i t_i \geq 600)], & m > 1. \end{cases} \tag{3}$$

This rule covers late single-evidence and multi-evidence retrieval. Among the 896 semantic long-range questions, 614 (68.5%) contain a single evidence point occurring at or after 10 minutes, while 282 (31.5%) involve multiple annotated evidence points satisfying the long-range criterion. Multiple evidence points must be topically coherent. Acoustic long-range items are generated separately from verified acoustic evidence.

A.4 Task Subtypes and Answer Requirements

The four families describe the requested operation, while a scoring route specifies how the submitted answer is checked. These are distinct annotations: for example, long-range questions may use single-fact or multi-slot scoring. The following distinctions explain task structure without assuming a universal ordering of difficulty.

Table A2: Task structures and the information a successful response must recover.

Family	Structure	Required answer
Factual	Single fact	The requested entity, value, event, or relation.
	Multiple slots	Every requested slot, with its correct entity–value correspondence.
	Enumeration	The requested item set; completeness is enforced when the question explicitly requires it.
Reasoning	Causal, comparative, procedural, or cross-evidence	The conclusion and any explanations or relations explicitly required by the question.
Temporal	Recording-clock localization	The identified event and its point or interval on the original recording clock.
	Mention order / event chronology	All requested events in the specified reference order; timestamps are required only when requested.
Long-range	Late single evidence	A source-specific fact supported at a late recording position.
	Separated evidence	Required facts recovered across separated positions and assigned to the correct slots or relation.

Boundaries. A single explicit fact is sufficient for direct retrieval, whereas reasoning requires the relation asked for in the stem. The presence of several annotated evidence points alone does not establish multi-hop reasoning. Likewise, the semantic long-range rule in Eq. 3 includes late single-point retrieval; it does not assert that every such question combines distant facts. Acoustic evidence is an evidence source shared by the four task families, rather than a fifth operation. Evaluators compare final committed answers; additional explanation is required only to the extent demanded by the question.

A.5 Representative Official Questions

Table A2 is illustrated below using actual English questions from the frozen release. Questions and gold answers are retained; formatting is normalized for typesetting. Evidence is the stored supporting excerpt, not a new transcription. Examples illustrate structures and are not used to estimate their frequencies. Source IDs and item suffixes identify the corresponding full records in the accompanying example file.

Example 1. Factual: single fact. Narrative
Q: What weight did the nurse record for the newborn?
A: 8 pounds 8.9 ounces.
Evidence: 00:26:06 - [Nurse] Eight pounds 8.9 ounces.
Source: `AxmWt7rYEXQ`; item: `62b1fc0b82b9`; route: `single_fact`.

Example 2. Factual: enumeration. Knowledge
Q: According to the recording, what are all the forms in which the angular-momentum relation is written, including the notation or reference frame associated with each?
A: - Vector notation: h = Iω.
- Matrix notation: $\tilde{h}$ = I$\tilde{\omega}$.
- In the primed frame: $\tilde{h}'$ = I'$\tilde{\omega}'$.
Evidence: 00:00:43 us write this h which is the angular velocity vector . This we have written as I times omega, — 00:00:52 ok. This is in the vector notation. And the same thing also in the matrix notation, we — 00:01:23 and in this frame, this is my h vector. So, now, if I took a, took a another frame say
Source: `2Ein0Yejqbo`; item: `5a4bfcd95247`; route: `exhaustive_list`.

Example 3. Reasoning: linking an action to an outcome. Entertainment
Q: What earlier action explains the later comment, “It shoots ink”?
A: Someone had put ink inside it.
Evidence: 00:13:08 I’m putting ink in there. — 00:30:30 - It shoots ink. - That is crazy.
Source: `Pj-3zshnyMs`; item: `992a3858acc8`; route: `reasoning`.

Example 4. Reasoning: applying a condition across evidence. Lifestyle
Q: Would the presenter have been eligible for the restaurant's free-delivery offer at their stated location, and why?
A: Yes, the presenter would have been eligible. They were in Kabul's Quwai Markaz area, and the restaurant provides free delivery within that area.
Evidence: 00:04:12 Right now, I am in the Quwai Markaz area — 00:19:29 yes, we do free delivery to the Quwai Markaz area
Source: `KNpozn__pmU`; item: `3c81fda03e17`; route: `reasoning`.

Example 5. Temporal: physical recording clock. Howto
Q: At what time in the recording is the chest circumference reported as 96.52 centimeters?
A: 00:00:29
Evidence: 00:00:29 Measure chest circumference (96.52 cm).
Source: `7aWKcbFrYkE`; item: `1bbd99593091`; route: `time_point_program`.

Example 6. Temporal: mention order. Lifestyle
Q: What is the chronological mention order of the complete set of Verizon mobile-hotspot averages and the concluding upload result from the DSL speed test?
A: 1. 00:15:00 — Verizon mobile-hotspot average: about 24 Mbps download.
2. 00:15:03 — Verizon mobile-hotspot average: 6.7 Mbps upload.
3. 00:16:55 — DSL speed test concluding result: 0.86 Mbps upload.
Evidence: 00:15:00 it averages about 24 megabits per second download — 00:15:03 and 6.7 megabits per second upload — 00:16:55 0.86 megabits per second
Source: `g6xz3ihm5AY`; item: `f3e74a20daff`; route: `time_order`.

Example 7. Long-range: late single evidence. Product
Q: What monthly cost did the presenter mention?
A: 1600
Evidence: 00:21:59 for a monthly cost of 1600
Source: `Dvzv4tHnc1I`; item: `0876b1a26e47`; route: `single_fact`.

Example 8. Long-range: separated slots. Entertainment
Q: During the Dream Phone playthrough in the recording, what phone number did the boys share, and who did the players ultimately declare had won the crush?
A: The boys shared the phone number 5-555-5555; the players ultimately declared that Chanse had won the crush.
Evidence: 00:00:26 and yes, their number is 5-555-5555. — 00:28:45 - It looks like Chanse wins the crush
The two cited positions are 28 min 19 s apart.
Source: `FlXfc09MQ7k`; item: `33ad8dc12dc0`; route: `multi_slot`.

Example 9. Acoustic evidence: factual recognition. Product
Q: What is heard at 17:39 in the recording?
A: At 00:17:39, a man laughs.
Evidence: [1059.4-1059.4] [acoustic] A man laughs. 1059-1060s [precise]
Source: `-aUZaYVAWDA`; item: `p2`; route: `single_fact`.

A.6 Source Distributions

The collection contains 1,732 source recordings totaling 1,341 hours, with median duration 36.7 minutes and mean duration 46.5 minutes. The proportion lasting at least 30 minutes is 91.9%. Figure 2 shows the duration and source-coverage distributions.

A.7 Task by Evidence Type

A.8 Language–Domain Coverage

The semantic track follows the same 30-question allocation in every one of the 128 Language–Domain cells. The acoustic track is intentionally less uniform: 39 cells contain fewer than 10 acoustic questions, 14 contain more than 10, and two cells are empty because no usable naturally occurring acoustic evidence was available under the release criteria. We preserve these differences rather than fill them with cross-language substitutions or synthetic events.

Table A3: Question composition by task and evidence source.

Task	Semantic	Acoustic	Total
Factual / Completeness	1,024	867	1,891
Reasoning	1,024	122	1,146
Temporal Localization	896	112	1,008
Long-range Retrieval	896	97	993
Total	**3,840**	**1,198**	**5,038**

B Construction and Quality Assurance

This appendix expands the construction stages summarized in Figure 3. The goal is to make the released benchmark traceable from source selection to final expert audit for the frozen 5,038-item pipeline.

B.1 Pipeline Summary

Table A4: Operational summary of the frozen MuLA-Bench construction pipeline.

Stage	Model / tool	Primary input	Main acceptance criterion
Source curation	rules + GPT-5.6 Sol	source metadata, matching VTT	usable audio/subtitles, valid language, no known leakage or confirmed dub/overlay
Semantic evidence	GPT-5.6 Sol + VTT rebase	timestamped subtitle cues	atomic fact supported by a real source cue; weak rebase rejected
Acoustic evidence	TimeAudio + Gemini 3.1 Pro	full recording + selected local clips	event is audibly present after blind-listen/alignment verification
QA generation	GPT-5.6 Sol	structured evidence plan	question has one evidence-supported target; fresh-context answer agrees
Automatic audit	programmatic gates + GPT-5.6 Sol	candidate QA + transcript/evidence	language, grounding, uniqueness, no leakage, no unintended shortcut
Expert audit	language experts	localized audio, transcript, QA, timestamps	answer/evidence correct; temporal reference corrected when necessary

The shared GPT-5.6 Sol calls used by semantic evidence mining, evidence planning, question writing, fresh-context answer derivation, and textual verification use temperature 0, high reasoning effort, and a maximum completion budget of 8,192 tokens in the frozen construction configuration.

B.2 Source Curation and Domain Structuring

A source enters construction with a catalog language code and a matching language VTT file. Language is a recording-level label inherited by its questions. Metadata terms suggesting dubbing or simultaneous interpretation are used only to trigger inspection; hard exclusion requires direct evidence in the subtitle/audio content, such as bilingual overlap or a mismatch between the declared language and dominant subtitle content. Occasional code-switching is not itself an exclusion criterion.

Domain assignment is performed over the transcript with GPT-5.6 Sol using an eight-way closed taxonomy: knowledge, how-to, narrative, entertainment, lifestyle, news, product, and belief/spiritual culture. The classifier chooses a dominant source-level function rather than segment-level topics. Known predecessor sources and leakage lists are excluded before QA generation.

Released question-level analyses use the frozen `domain` field attached to each item. This field defines semantic quotas and the Language–Domain comparisons; the source classification above describes the curation stage. Source-coverage statistics count distinct `source_id` values within each domain, and Figure 2(b) normalizes these counts by their sum. The collection total of 1,732

counts distinct recordings across the complete release. Statistical resampling groups all questions from the same recording together.

B.3 Semantic Evidence Mining and Planning

Semantic evidence is mined from VTT text in long slices and represented as atomic, self-contained facts with timestamps and information types. The frozen configuration processes up to 40 minutes per mining slice and targets roughly 8–15 high- or medium-information facts per 10 minutes, while allowing denser segments to contribute more. Low-information filler is not counted toward the usable-fact pool.

Each mined fact is rebased to the real VTT cue that supports it. A standard semantic fact requires a rebase score of at least 2.5; temporal plans use the stricter threshold 3.5. Facts that cannot be tied back to a sufficiently confident source cue are discarded. This step is the source of semantic timing in the released benchmark; MuLA-Bench does not use forced alignment to create the semantic timestamps in the official 5,038-item release.

The evidence planner then creates structured plans over the retained facts. Supported structures include retrieval, enumeration, multi-slot, multi-hop, time-point, and time-order. Multi-hop plans must use at least two facts linked by a concrete causal, comparative, temporal, or arithmetic relation; plans spanning distant but unrelated statements are rejected. Long-range semantic plans follow Eq. 3.

B.4 Natural Acoustic Evidence Construction

Full-recording proposal stage. TimeAudio scans the full audio using 20-s windows with a 20-s hop. The frozen configuration uses the TimeAudio checkpoint built on Whisper-large-v2, BEATs, and Vicuna-7B components and requests event descriptions with start and end timestamps. This stage is exhaustive over the recording windows; the later Gemini stage is not.

High-value window selection. Because multimodal re-listening is substantially more expensive, a heuristic selector retains at most 12 candidate windows per recording. Event classes are assigned fixed informativeness weights, with short distinctive effects, applause, laughter, cheering, music, and other salient non-speech events ranked above weak breath or background-noise categories. At the source level, the selection score is

$$S = W + 4N_{\mathrm{div}} + 2N_{\mathrm{high}}, \tag{4}$$

where W is the summed event weight, N_{div} counts distinct nontrivial event classes, and N_{high} counts high-weight events. The selector first promotes class diversity, then fills remaining slots with high-value windows while avoiding redundant windows where possible.

Two-pass multimodal verification. Gemini 3.1 Pro verifies selected source clips in two passes. Pass 1 is blind to the TimeAudio proposal: the model watches and listens to the clip and lists non-speech events it can actually perceive. Visual context is used only to support or rule out a possible source; the event must remain audible. Pass 2 then reveals the TimeAudio proposal and aligns each proposal to the blind observation with *keep*, *modify*, or *delete*. The alignment prompt is deliberately conservative for laughter, applause, crying, and crowd reactions, while requiring stronger evidence before retaining specific ambient-source labels such as water, wind, or background music.

Post-processing converts the two-pass decisions into event records. Only events with sufficient evidence strength enter the QA pool. Coarse in-window events remain usable for non-temporal questions, whereas temporal questions require precise timing metadata. Expert-rejected event IDs and confirmed invalid acoustic sources are removed before harvesting.

B.5 Decoupled Question and Answer Generation

The semantic question-writing stage receives a structured evidence plan and is explicitly instructed to output the question only. It may see the intended evidence plan but not a prewritten reference answer. Its prompt prohibits answer keywords, source timestamps, explicit reasoning operations, and location hints, and requires a single defensible answer grounded in the recording rather than world knowledge.

A separate fresh-context call then receives the generated question together with the supporting evidence and derives the answer. For semantic questions, this answer is derived from transcript evidence; for acoustic questions, the released gold is tied to the verified event-grounded evidence record rather than to a Gemini full-recording teacher answer. The answer call and question writer use the same model family in separate calls.

B.6 Automatic Verification Gates

The main semantic verifier is organized into three layers. **G1** is a programmatic language gate over the generated question and answer. **G2** checks the candidate against transcript context for unsupported claims, false premises, multiple valid answers, answer or method leakage, omissions, and QA mismatch. When the transcript fits the verifier budget it is supplied in full; for longer transcripts, deterministic beginning/middle/end coverage is used instead of claiming complete transcript visibility. Repairable failures such as answer leakage or a narrow QA mismatch may be minimally repaired once and reverified; fatal grounding defects are dropped. **G3** is a question-only ablation: a solver is asked to answer without audio, transcript, or cited evidence, and a second check removes the item if the guess materially matches the reference.

After LLM verification, programmatic packer rules enforce additional invariants. Questions must retain nonempty evidence; semantic numeric answers must be supported by the source cue rather than only by a generated fact summary; reasoning items require at least two evidence pieces; long-range items must satisfy the corresponding temporal geometry; and semantic time-point golds must remain close to their stored listening point. Standard semantic evidence uses a minimum rebase score of 2.5, while temporal items use 3.5. These checks are construction-time consistency guards and are distinct from the ± 3 s evaluation tolerance.

B.7 Task-Aware Acoustic Shortcut Tests

Acoustic items receive a no-audio shortcut check across all task families. The solver receives the question q and full transcript T_{full}, without audio or acoustic captions. It attempts two paths: using sound clues explicitly named in the question, alone or with the transcript; and using the full transcript, including evidence from distant passages. For reasoning, both locally stated conclusions and inferences supported by other transcript passages count. A sound mentioned in the question is a usable clue even if the transcript does not name it. Unsupported guesses do not establish a shortcut.

Let c_q indicate that the solver reports a supported answer $\hat{a}_q$ through either path, and let M_q indicate that the comparison stage confirms a material match to the reference a_q under these shortcut rules. The filter is

$$\text{Reject}_{\text{aco}}(q) = \mathbf{1}[c_q \wedge M_q]. \tag{5}$$

If the solver declares the item unanswerable, the candidate proceeds to the remaining checks; otherwise, the comparison stage determines whether to reject it. This procedure removes detected shortcuts for factual/completeness, reasoning, temporal, and long-range items. Acoustic reasoning additionally requires the cited sound to contribute evidence needed for the target inference.

In addition, every cited acoustic event is re-presented in its localized source window for an independent audibility check. A rejected or uncertain cited event invalidates the corresponding evidence plan before expert review.

B.8 Human Annotation Protocol

Annotators and assignment. We use one paid language expert for each of the 16 target languages. Annotators have relevant annotation experience and language credentials. Each released item receives a final quality-control review by the expert assigned to its language.

Review interface and evidence access. For each candidate, reviewers have access to the question, reference answer, transcript evidence, evidence timestamp(s), source link, and the corresponding localized audio segment. Reviewers are required to listen to the localized evidence region rather than judge from text alone. For acoustic questions, the cited non-speech or paralinguistic event must be audibly present in that region. For completeness-style questions that require exhaustive coverage, reviewers additionally search the full VTT to check for relevant evidence elsewhere in the recording.

Decision categories and resolution. The annotation interface records whether the item is correct or exhibits a question problem, incorrect answer, incorrect sound, incorrect time, unintelligible audio, or another issue. Reviewers check evidence sufficiency, answer correctness, and, where applicable, acoustic-event and timestamp correctness. Correct items are accepted. Repairable errors in the question, answer, evidence, or timestamp are corrected; irreparable candidates are replaced. Every released question passes this final expert review.

Temporal calibration. Temporal review explicitly separates construction-time consistency, human annotation, and evaluation tolerance. Semantic construction first applies a coarse consistency gate between the proposed clock and its localized evidence point. The expert then replays the localized region, identifies when the referenced sound or spoken event actually occurs, and corrects the reference time when necessary. Evaluation finally applies the deterministic temporal matcher in Eq. 7; its ± 3 s tolerance is therefore a scoring rule, not a substitute for timestamp annotation.

B.9 Frozen Construction Configuration

Table A5: Key frozen configuration choices used to construct MuLA-Bench.

Component	Frozen setting	Role
Shared LLM	GPT-5.6 Sol, temperature 0, high reasoning effort	evidence mining, planning, QA, textual verification
Semantic slice	40 min maximum per mining slice	atomic-fact extraction
Semantic rebase	score ≥ 2.5; temporal ≥ 3.5	source-cue grounding
TimeAudio sweep	20-s window / 20-s hop	full-recording event proposals
Gemini verification	Gemini 3.1 Pro, at most 12 selected windows/source	blind listen + proposal alignment
Acoustic candidate harvest	configured caps: 10 per L$\times$D cell; 3 per source	candidate-stage diversity controls; final counts are in Appendix A
Temporal evaluation	± 3 s	final scoring tolerance only

Appendix E specifies the information supplied to each model call and the output decisions used by the pipeline. Programmatic gates remain separate from the semantic decisions made by the models.

B.10 Audit and Rejection Statistics

Event-level expert auditing removes 56 rejected acoustic event IDs before final question harvesting, and eight confirmed invalid acoustic sources are excluded. The released benchmark contains 1,732 unique source recordings and 5,038 questions after all automatic gates and expert verification.

C Evaluation and Statistical Protocol

C.1 Frozen Model Pool, Snapshots, and Decoding

Table A6 records the frozen model pool, exact model identifiers, and inference settings. For hosted APIs, we use provider defaults unless otherwise noted and do not explicitly set temperature, thinking mode, or a reasoning budget. Local open-weight models are decoded greedily with the limits shown in the table. Model-family documentation and release descriptions are available for Gemini 3.8 Flash, Gemini 3.7 Flash, and Gemini 3.1 Pro (Google DeepMind, 2026c;b;a); Qwen3.8-Omni-Flash, Qwen3.5-Omni, and Qwen3-Omni (Alibaba Cloud Model Studio, 2026; Qwen Team, 2026; Xu et al., 2025); Seed2.0 Lite and Muse Spark 1.2 (ByteDance Seed, 2026; Meta AI, 2026); and MOSS-Audio and Audio Flamingo Next (Yang et al., 2026a; Ghosh et al., 2026).

C.2 Question Prompt

All systems receive the original benchmark question preceded by a fixed language-specific wrapper equivalent to "Please answer the following question based on the full audio content." No additional system instruction, chain-of-thought request, or JSON schema is added. This wrapper is translated

Table A6: Frozen evaluation model snapshots and decoding settings.

Model	Model identifier	Sampling / thinking	Output limit
Gemini 3.8 Flash	`gemini-3.8-flash`	provider default	provider default
Gemini 3.7 Flash	`gemini-3.7-flash`	provider default	provider default
Qwen3.8-Omni-Flash	`qwen3.8-omni-flash`	provider default	provider default
Gemini 3.1 Pro	`gemini-3.1-pro-preview`	provider default	provider default
Qwen3.5-Omni-Plus	`qwen3.5-omni-plus`	provider default	provider default
Qwen3.5-Omni-Flash	`qwen3.5-omni-flash`	provider default	provider default
Doubao seed 2.0 Lite	`doubao-seed-2-0-lite-260428`	provider default	provider default
Muse Spark 1.2	`muse-spark-1.2`	provider default	8,192 tokens
MOSS-Audio-8B-Thinking	OpenMOSS-Team/MOSS-Audio-8B-Thinking	greedy	1,024 tokens
Qwen3-Omni-30B-A3B-Instruct	Qwen/Qwen3-Omni-30B-A3B-Instruct	greedy	thinker: 512 tokens
Audio Flamingo Next 8B	nvidia/audio-flamingo-next-hf	greedy; repetition penalty 1.2	512 tokens

into each of the 16 target languages and is identical across models for a given item. The source transcript, construction evidence, and video frames are never provided at evaluation time.

C.3 Canonical Input and Context Limits

Hosted models receive the canonical audio for each question. For the four predefined items whose evidence extends beyond the three-hour mark, this is a fixed later-stage three-hour segment containing the annotated evidence. Evidence timestamps refer to the original source clock.

Local models receive the longest prefix permitted by their native context limits, starting at the beginning of the canonical input. The frozen limits are 3,276.8 s for MOSS-Audio-8B-Thinking, 2,400 s for Qwen3-Omni-30B-A3B-Instruct, and 1,800 s for Audio Flamingo Next 8B. Content-preserving re-encoding or bitrate reduction is allowed for hosted APIs; persistent failures after retries score zero.

C.4 Fixed-Denominator Scoring and Coverage

All headline and slice accuracies use the official number of benchmark items as the denominator. For the full benchmark this is 5,038, regardless of whether a model returns a valid response. After retrying transient network or server errors, persistent input-limit failures, empty responses, and policy refusals are scored as incorrect. We separately report *coverage*—the fraction of items that yield a non-empty valid response under the prescribed input—as an operational diagnostic.

Table A7: Valid-response coverage for all eleven models. Coverage does not change the 5,038-item accuracy denominator.

Model	Valid / total	Coverage (%)
Gemini 3.8 Flash	5,036 / 5,038	99.96
Gemini 3.7 Flash	5,037 / 5,038	99.98
Qwen3.8-Omni-Flash	5,030 / 5,038	99.84
Gemini 3.1 Pro	5,038 / 5,038	100.00
Qwen3.5-Omni-Plus	5,035 / 5,038	99.94
Qwen3.5-Omni-Flash	4,770 / 5,038	94.68
Doubao seed 2.0 Lite	4,923 / 5,038	97.72
Muse Spark 1.2	5,030 / 5,038	99.84
MOSS-Audio-8B-Thinking	5,038 / 5,038	100.00
Qwen3-Omni-30B-A3B-Instruct	5,038 / 5,038	100.00
Audio Flamingo Next 8B	5,037 / 5,038	99.98

C.5 Evaluation Rubric Design

Following the broader LLM-as-a-judge paradigm (Zheng et al., 2023), open-ended evaluation is organized around *answer structure* rather than a single holistic similarity score. Each item is mapped to one of six routes with an explicit binary acceptance rubric. This separation matters because the same benchmark task can require very different answer contracts: a factual item may request one entity, an exhaustive set, or several mandatory slots, while a temporal item may require either a physical clock value or only an ordering. The routing therefore prevents a lenient single-fact criterion from being reused for intrinsically stricter outputs.

We follow four design principles. First, **semantic tolerance**: harmless paraphrases, translations, unit-equivalent expressions, and benign transcription variants are accepted when identity remains unambiguous. Second, **structural strictness**: exhaustive and multi-slot questions receive no credit when a required element is missing. Third, **claim-level reasoning**: reasoning responses are judged on the primary claims required by the reference, not stylistic overlap or the presence of a particular chain of thought. Fourth, **deterministic temporal grounding**: the LLM judge only matches events to candidate times; final clock acceptance is decided by code.

Table A8: Route-specific binary rubrics used for open-ended evaluation.

Judge route	Count	Binary acceptance criterion
`single_fact`	2,454	The final answer commits to the correct core grounded fact; benign paraphrase or transcription variants are allowed unless the question explicitly requires an exact string.
`exhaustive_list`	81	Every required member of the requested set must be present; extra contradictory members fail.
`multi_slot`	394	Every mandatory slot must be filled and attached to the correct entity or value.
`reasoning`	1,145	The answer must reach the target conclusion and preserve the required relation or explanation.
`time_point_program`	575	The referenced event must match semantically and the predicted point/interval must pass the deterministic temporal matcher.
`time_order`	389	The explicit event sequence must be complete, correctly ordered, and internally consistent; time checks follow the item configuration.
Total	**5,038**	

Judge configuration. All semantic judge routes use GPT-5.6 Sol (OpenAI, 2026). The frozen evaluator uses low reasoning effort for `single_fact` and medium effort for `exhaustive_list`, `multi_slot`, `reasoning`, `time_point_program`, and `time_order`. The judge receives the question, reference answer, and model prediction under the route-specific rubric. Programmatic guards then enforce route-specific hard constraints.

C.6 Deterministic Temporal Matching

For metric temporal questions, the LLM judge identifies the event–time pairing but does not decide the numeric tolerance. The programmatic matcher uses a 3.0 s tolerance. Point–point predictions pass when the absolute difference is at most 3.0 s. If the reference is an interval and the model predicts a point, the point must fall inside the reference interval after expanding each endpoint by 3.0 s. If the model predicts an interval, its width must be at most 10.0 s; such a short interval passes only when it contains a reference point (under the same tolerance) or overlaps a reference interval after expansion. Wider predicted windows do not receive credit merely for covering the answer. Gold spans of at least 0.5 s, or items explicitly typed as intervals, are treated as reference intervals. For *time-order*, the judge checks sequence completeness, the requested ordering mode, and internal consistency. All 387 semantic temporal ordering items use audio mention order (`mention_order`) and `require_associated_times=false`, with no per-item overrides; timestamp accuracy is not required for credit. The optional associated-time branch in Appendix E is inactive for this slice.

C.7 Final-Answer Policy

Only the final committed answer is scored. Intermediate reasoning, self-correction, or discarded candidates are ignored. If the final response contains mutually competing unresolved answers, the item is marked incorrect. A response that claims it cannot access the audio but then guesses from external knowledge receives no credit under the explicit audio-denial guard. Missing responses and persistent inference failures are incorrect under the fixed-denominator protocol.

C.8 Expert–Evaluator Agreement Audit

We audit the frozen evaluator on **844 distinct questions**, with one frozen model response per reviewed question. The completed audit spans all 16 benchmark languages and all six judge types: single-fact, exhaustive-list, multi-slot, reasoning, time-point, and time-order. Sampling is organized by language and judge route, with balanced evaluator-positive and negative labels where feasible; cells with fewer eligible items contribute all available items. Model identity is not used as a stratification variable.

For each sampled response, experts independently assign a binary correctness label using the question, gold answer, evidence, and model response, without access to the evaluator's label or rationale. The independently recorded human labels are then compared with the frozen evaluator's judgments. The compared evaluator label is the frozen `judge_v3` end-to-end score, including programmatic temporal checks where applicable. All reviewed responses remain in a single audit denominator; audit observations are not retrospectively removed from the sample.

Experts agree on **795/844 responses (94.2%)**. Because sampling is stratified and label-balanced where feasible, this percentage describes the reviewed sample; it is not a population-weighted estimate of judge accuracy. Table A9 reports agreement across all 16 languages.

Table A9: Expert–evaluator agreement across all 16 benchmark languages.

Lang.	n	Agreed	%	Lang.	n	Agreed	%
ar	50	47	94.0	de	49	46	93.9
en	51	48	94.1	es	49	45	91.8
fr	58	55	94.8	hi	50	50	100.0
id	47	43	91.5	it	52	52	100.0
ja	51	48	94.1	ko	60	56	93.3
pt	55	47	85.5	ru	51	49	96.1
th	58	54	93.1	tr	52	49	94.2
vi	59	57	96.6	zh	52	49	94.2
Total		**795**	**94.2**				$n = 844$

Human inter-reviewer agreement. Two human reviewers independently assess 150 responses drawn from the 844-response audit sample, without access to the evaluator's judgments or each other's labels. Their judgments agree on **145/150 responses (96.7% raw agreement)**. This inter-reviewer measure complements the expert–evaluator agreement reported above.

C.9 Scores and Statistical Analyses

For a fixed evaluation subset $\mathcal{S}$, with binary item scores s_q,

$$\mathrm{Acc}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{q \in \mathcal{S}} s_q. \tag{6}$$

Missing and invalid outputs remain in the denominator. Wilson 95% intervals summarize individual model proportions under an item-independent binomial approximation. They are descriptive intervals; recording-cluster methods are used for the inferential diagnostics below.

Temporal score. Semantic event agreement and physical-clock matching are both required:

$$s_q^{\mathrm{time}} = \mathbf{1}[\mathrm{event}(\hat{a}_q) = \mathrm{event}(a_q)] \cdot M_{\mathrm{time}}(\hat{t}_q, t_q^\star). \tag{7}$$

Here M_{time} applies the point/interval rules above. Task families and judge routes are distinct taxonomies; a task-conditioned slice need not have the same denominator as an all-task judge-route count.

Language–Domain residuals and rankings. Each model's residual is computed from Eq. 1 on the balanced semantic track and averaged across all eleven models. Spearman correlations compare conditional domain/task rankings with overall semantic language rankings. The semantic cells contain a median of 11 recordings (range 2–21). We retain all cells in descriptive residuals and report their recording counts uniformly in the appendix. Inferential highlighting is suppressed for cells supported by fewer than five recordings. For raw residuals, 10,000 source-level exponential multiplier replicates share one weight across all questions and model responses from each recording, across the complete release. Each replicate recomputes cell means and equal-weight language and domain marginals. Positive weights avoid empty cells in the resampled estimates. Simultaneous max-$|t|$ intervals cover the 120 eligible cells.

The composition-adjusted sensitivity analysis fits the fixed eleven-model item mean using saturated cell intercepts, task, scoring route, standardized log duration, relative earliest position, relative span, and evidence count. Double-centered cell coefficients use a source-clustered sandwich covariance and Gaussian max-$|t|$ simultaneous intervals from 10,000 simulations. This descriptive adjustment assumes common covariate slopes across cells.

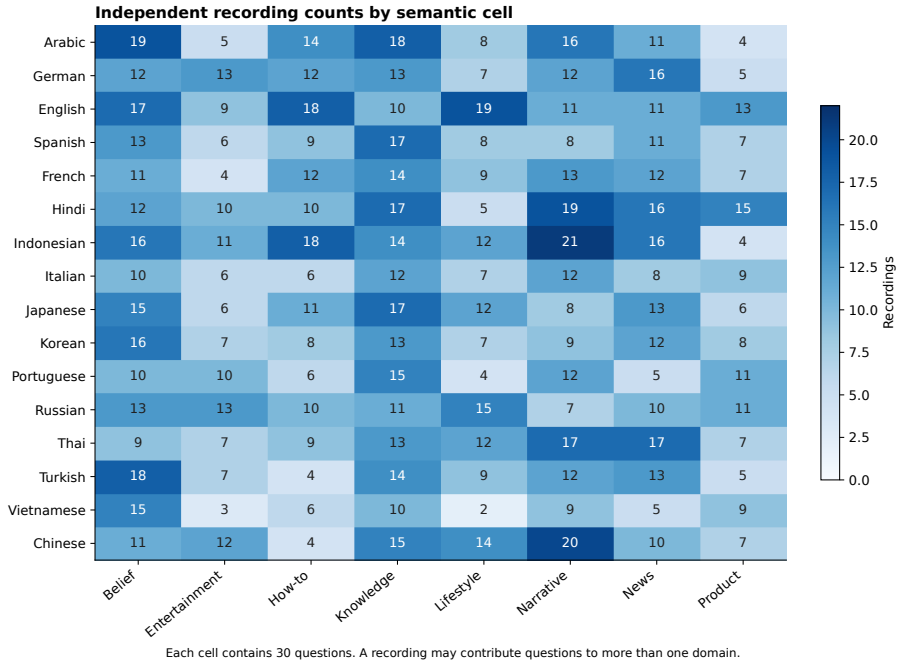


Figure A1: Distinct recordings within each semantic Language–Domain cell. Each cell contains 30 questions.

Hosted evidence-position analysis. Figure A5 uses all 2,001 temporal and long-range questions across both evidence tracks (1,008 temporal and 993 long-range), from 1,018 recordings. Each question receives the mean binary score of the four specified hosted models; failed or invalid responses retain score zero. Absolute position is `earliest_evidence_sec`/60, measured from the source-recording origin. No cutoff is applied at 30 minutes or at any local model's input limit. The $\geq$ 120-minute tail contains eight temporal and ten long-range questions. Every plotted bin reports its own task denominators.

We estimate each task's Spearman correlation between position and the item mean, and their difference $\Delta\rho = \rho_{\text{temporal}} - \rho_{\text{long-range}}$. Percentile 95% intervals use 3,000 recording-level pairs-bootstrap replicates (seed 20260924). Each replicate samples recordings jointly across tasks, keeping all

associated questions and their four model responses together. The OLS check fits $\bar{Y}_q = \beta_0 + \beta_T T_q + \beta_x x_q + \beta_{Tx} T_q x_q + \epsilon_q$, where T_q indicates temporal questions and x_q is absolute position in minutes. Standard errors cluster by recording, with the finite-sample sandwich correction; the two-sided interaction test is targeted to this pair of tasks. The four evaluated models are treated as a fixed cohort.

The four predefined post-3h cases belong to factual or reasoning tasks; none is included in this temporal–long-range analysis. The adjusted absolute-position OLS additionally includes language, domain, evidence type, scoring route, log source duration, relative evidence span, and evidence count. A separate relative-position OLS replaces minutes with earliest position divided by source duration and allows task-specific log-duration slopes. Exact formulas accompany the analysis code.

Identifiable logistic specification. The descriptive model uses 55,418 item–model observations (5,038 questions across eleven models). Its linear predictor is

$$\begin{aligned}\operatorname{logit}\Pr(Y_{qm}=1) = \beta_0 + \alpha_m + \lambda_{\ell(q)} + \delta_{d(q)} + \gamma_{e(q),t(q)} \\ + \boldsymbol{\theta}^\top \mathbf{z}_q,\end{aligned} \tag{8}$$

where $\mathbf{z}_q$ contains standardized log duration, normalized earliest position, normalized evidence span, and evidence count. Joint evidence–task categories use semantic factual as reference; model, language, and domain references are Gemini 3.8 Flash, English, and Knowledge. No additional evidence or task main-effect indicators are included. The fit is unpenalized, with recording-clustered 95% Wald intervals (1,732 source clusters, with a finite-sample sandwich correction). Continuous variables are standardized once over the 5,038 questions. Category contrasts and +1 SD continuous shifts describe different comparisons; neither is interpreted as causal variable importance.

D Additional Results

All language, domain, evidence–task, and headline results below use the same eleven models. Evidence-position diagnostics use the four-model hosted cohort. Compact-table abbreviations are G3.8/G3.7: Gemini Flash; Q3.8: Qwen3.8-Omni-Flash; Pro: Gemini 3.1 Pro; Plus/Q3.5: Qwen3.5-Omni-Plus/Flash; Q30B: Qwen3-Omni-30B-A3B-Instruct; AFN: Audio Flamingo Next; and MOSS: MOSS-Audio-8B-Thinking.

D.1 Language and Domain Breakdowns

Table A10: Overall language accuracy (%) for all eleven models. Counts cover both evidence tracks.

Lang.	n	G3.8	G3.7	Q3.8	Pro	Plus	Doubao	Q3.5	Spark	Q30B	AFN	MOSS
zh	337	73.6	75.4	78.0	56.1	66.5	67.1	57.0	21.1	32.9	20.8	25.5
en	347	76.1	75.2	77.5	62.8	66.0	65.7	53.3	51.9	35.7	29.7	24.8
es	316	70.9	70.3	66.1	58.2	59.2	51.6	42.4	35.8	27.8	16.1	16.8
fr	338	76.3	73.1	73.7	63.3	63.9	58.6	47.3	39.1	26.0	18.9	15.4
ja	320	77.5	78.8	78.8	68.8	66.6	64.4	53.1	35.6	30.6	18.1	21.6
ko	333	67.0	68.8	68.2	57.1	55.3	52.3	44.1	25.5	21.6	6.0	16.5
id	324	73.8	74.7	73.1	64.8	58.3	59.3	45.7	35.2	25.3	13.9	9.9
ru	308	72.1	71.4	76.3	59.7	63.0	59.7	49.4	31.5	26.9	15.3	17.5
ar	313	71.6	72.8	70.9	63.9	60.4	53.4	41.9	31.9	11.8	7.3	3.5
th	309	72.5	72.2	65.7	58.6	57.9	50.2	43.0	31.1	19.1	3.6	12.3
vi	288	76.7	77.8	70.8	61.5	63.5	51.7	42.7	26.4	24.7	3.8	4.2
de	304	73.7	73.4	76.3	62.5	57.6	56.2	45.4	38.8	23.0	16.1	9.2
it	304	71.1	64.8	65.8	55.6	56.2	49.7	38.2	33.2	22.0	14.1	9.9
pt	306	73.2	73.5	71.9	62.7	60.1	56.9	49.3	38.2	26.8	16.0	14.7
hi	290	73.1	71.4	72.4	61.0	57.2	52.1	44.1	30.3	21.0	1.0	6.9
tr	301	76.7	76.7	74.4	65.8	61.8	60.8	44.2	33.2	18.3	10.3	1.7

Table A11: Semantic domain accuracy (%) for all eleven models; each domain has 480 questions.

Domain	G3.8	G3.7	Q3.8	Pro	Plus	Doubao	Q3.5	Spark	Q30B	AFN	MOSS
Knowledge	80.2	78.3	80.4	74.8	70.2	67.7	55.8	39.8	31.5	15.0	16.0
How-to	77.1	77.7	78.3	70.2	69.6	64.4	55.0	36.2	36.2	16.2	15.0
Narrative	78.3	76.7	78.8	68.5	63.3	58.1	48.8	28.1	25.2	13.5	11.5
Entertain.	80.4	81.9	79.4	73.3	63.5	63.3	48.5	35.6	27.1	19.0	15.6
Lifestyle	81.2	81.2	79.4	71.9	65.2	67.5	53.8	36.5	25.6	15.2	16.0
News	76.7	77.3	77.1	74.0	66.9	61.7	50.8	36.2	25.6	12.1	9.8
Product	77.7	78.8	77.7	71.9	63.3	61.3	51.0	38.5	28.5	15.6	12.5
Belief	80.8	79.8	80.4	73.5	72.1	59.4	55.0	42.3	31.5	14.6	11.5

D.2 Evidence by Task

Table A12: Semantic accuracy (%) for all eleven models. Denominators are 1024, 1024, 896, 896 in column order.

Model	Factual	Reasoning	Temporal	Long-range
Gemini 3.8 Flash	90.53	73.44	65.85	85.60
Gemini 3.7 Flash	91.41	72.75	65.18	85.60
Qwen3.8-Omni-Flash	89.45	74.51	64.96	85.94
Gemini 3.1 Pro	89.84	70.21	41.07	85.71
Qwen3.5-Omni-Plus	84.57	59.96	39.96	81.03
Qwen3.5-Omni-Flash	72.27	43.26	26.90	65.40
Doubao seed 2.0 Lite	82.03	60.25	29.24	77.79
Muse Spark 1.2	53.12	29.98	14.73	47.43
MOSS-Audio-8B Thinking	18.95	10.84	6.70	17.08
Qwen3-Omni 30B-A3B-Instruct	45.70	21.68	11.94	34.93
Audio Flamingo Next 8B	25.88	10.35	5.58	17.97

Table A13: Acoustic accuracy (%) for all eleven models. Denominators are 867, 122, 112, 97 in column order.

Model	Factual	Reasoning	Temporal	Long-range
Gemini 3.8 Flash	48.90	65.57	73.21	82.47
Gemini 3.7 Flash	46.71	70.49	74.11	81.44
Qwen3.8-Omni-Flash	42.56	72.13	75.89	85.57
Gemini 3.1 Pro	17.30	57.38	28.57	68.04
Qwen3.5-Omni-Plus	35.18	63.93	46.43	72.16
Qwen3.5-Omni-Flash	19.38	52.46	30.36	67.01
Doubao seed 2.0 Lite	34.14	59.84	8.93	79.38
Muse Spark 1.2	18.57	63.93	0.89	55.67
MOSS-Audio-8B Thinking	9.46	40.98	7.14	18.56
Qwen3-Omni 30B-A3B-Instruct	7.73	37.70	2.68	22.68
Audio Flamingo Next 8B	5.77	25.41	1.79	13.40

Table A14 reports both temporal routes for all eleven models. Its 387 order and 568 time-point items account for 955 of the 1,008 temporal questions; the remaining 53 use other stored judge routes.

D.3 Event Ordering and Time-Point Localization

Table A14 expands the representative examples in Table 3 to all eleven models. The two scoring routes use distinct question sets. Nine systems have lower accuracy on time-point localization, while the three highest-scoring systems have small or reversed gaps. The comparison diagnoses operation-dependent performance; it does not condition on successful recognition of the same event.

Table A14: Accuracy on distinct temporal slices (%). Gaps use unrounded scores.

Model	Order ($n = 387$)	Point ($n = 568$)	Gap (pp)
Gemini 3.8 Flash	65.6	66.4	-0.7
Gemini 3.7 Flash	68.0	64.1	3.9
Qwen3.8-Omni-Flash	63.8	67.6	-3.8
Gemini 3.1 Pro	59.7	23.1	36.6
Qwen3.5-Omni-Plus	40.1	38.2	1.8
Doubao seed 2.0 Lite	49.4	8.1	41.3
Qwen3.5-Omni-Flash	31.5	21.5	10.0
Muse Spark 1.2	26.6	1.1	25.6
Qwen3-Omni-30B	17.8	3.2	14.7
Audio Flamingo Next 8B	10.1	0.5	9.5
MOSS-Audio-8B-Thinking	10.9	3.3	7.5

D.4 Domain and Task Profiles

Figures A2 and A3 visualize descriptive accuracy profiles for all eleven models. Models are split into groups of six and five only to keep curves legible. Every axis uses a 0–100% scale; polygon area is not an aggregate score. Missing or invalid responses remain incorrect under the official denominators.

Domain profiles use semantic questions only (480 per domain), preserving the balanced Language–Domain design. Task profiles separate semantic and acoustic evidence; their denominators are given in Tables A12 and A13. Figure 4 uses all eleven models; Figure A5 uses the specified four-hosted-model cohort.

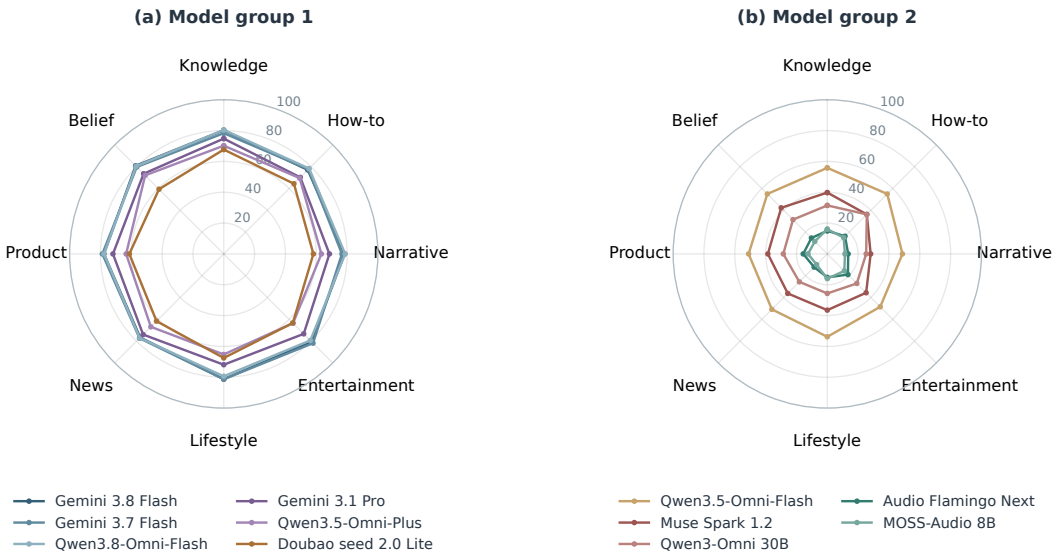


Figure A2: Semantic domain accuracy for all eleven models. Each domain contains 480 questions. The two panels share the same scale; the model split is for readability.



Figure A3: Task accuracy separated by evidence type for all eleven models. The top and bottom rows use the two model groups in Figure A2; columns show semantic and acoustic evidence. All panels share a 0–100% scale.

D.5 Headline Accuracy Intervals

Table A15: Overall accuracy and Wilson 95% intervals over the fixed question set.

Model	Correct / total	Accuracy (%)	95% CI
Gemini 3.8 Flash	3,702 / 5,038	73.48	[72.25, 74.68]
Gemini 3.7 Flash	3,685 / 5,038	73.14	[71.90, 74.35]
Qwen3.8-Omni-Flash	3,656 / 5,038	72.57	[71.32, 73.78]
Gemini 3.1 Pro	3,093 / 5,038	61.39	[60.04, 62.73]
Qwen3.5-Omni-Plus	3,069 / 5,038	60.92	[59.56, 62.26]
Qwen3.5-Omni-Flash	2,341 / 5,038	46.47	[45.09, 47.85]
Doubao seed 2.0 Lite	2,872 / 5,038	57.01	[55.63, 58.37]
Muse Spark 1.2	1,702 / 5,038	33.78	[32.49, 35.10]
MOSS-Audio-8B Thinking	676 / 5,038	13.42	[12.50, 14.39]
Qwen3-Omni 30B-A3B-Instruct	1,248 / 5,038	24.77	[23.60, 25.98]
Audio Flamingo Next 8B	678 / 5,038	13.46	[12.54, 14.43]

D.6 Residual Variation across Models

Figure A4 complements the cross-model mean residuals in Figure 4(a) with their cross-model standard deviations. The interactions are not uniformly shared across models: Vietnamese–Product has a residual standard deviation of 12.9 pp, Vietnamese–Entertainment 5.9 pp, and Italian–News 11.1 pp. All values use the same frozen eleven-model scores as the main analysis.

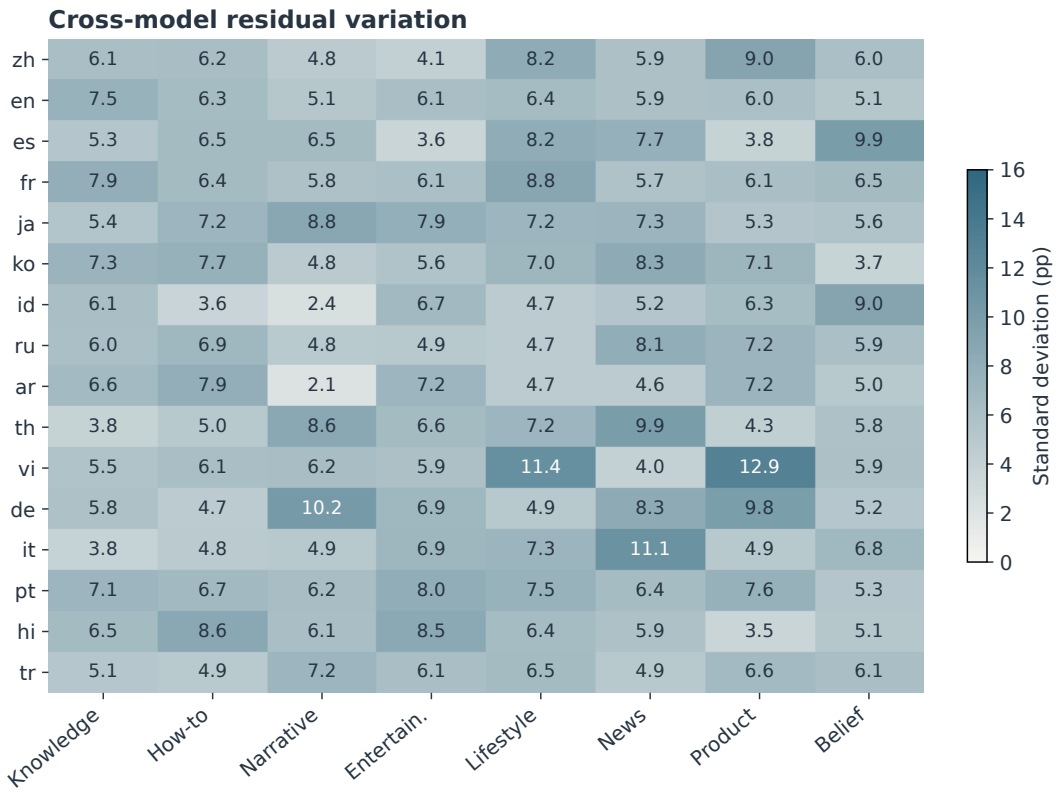


Figure A4: Cross-model standard deviation of Language–Domain additive residuals on the balanced semantic track, in percentage points (pp). The eleven-model evaluation set and the residual definition are the same as in Figure 4(a). Larger values indicate greater variation in the interaction residual across models.

D.7 Hosted Position Profiles and Adjusted Contrasts

Figure A5 reports the absolute-position analysis referenced in Section 5.4. The left panel displays four-model mean accuracies with task-specific question counts; the right panel reports task correlations and their difference with recording-cluster bootstrap intervals. The shaded final bin is sparse but remains in all tests.

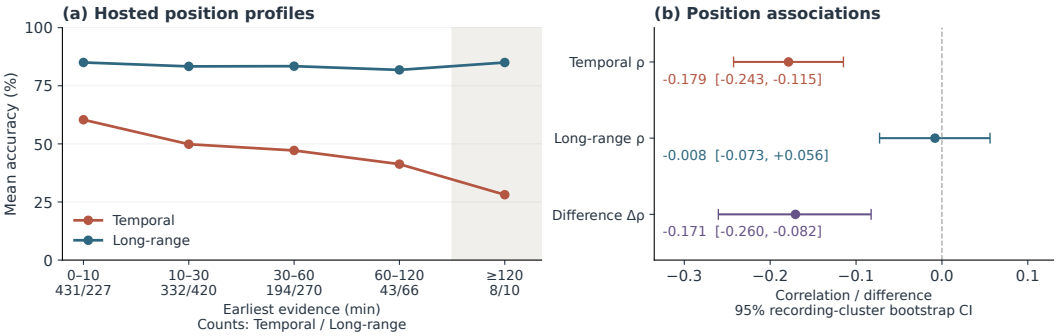


Figure A5: Absolute evidence position in the four-hosted-model cohort: Gemini 3.8 Flash, Gemini 3.7 Flash, Gemini 3.1 Pro, and Qwen3.5-Omni-Plus. (a) Mean accuracy by earliest evidence position; counts below each bin are Temporal / Long-range. The shaded $\geq$ 120-minute bin contains 8/10 questions. (b) Spearman correlations and their difference, with 95% intervals from 3,000 joint recording-cluster bootstrap replicates. All positions are retained.

Table A16: Temporal–long-range position-interaction tests on the four-model item mean over all 2,001 questions. Standard errors cluster by recording. These specifications estimate different associations.

OLS specification	Interaction p
Absolute position, unadjusted	0.00060
Absolute position, covariate-adjusted	0.00345
Relative position, covariate-adjusted	0.269

The absolute-position association remains under the specified covariate adjustment. The relative-position interaction does not reach significance when task-specific duration slopes are included; this addresses position as a fraction of total duration, a different comparison from elapsed minutes. Individual-model absolute-position interaction tests also vary (Gemini 3.8: $p = 0.173$; Gemini 3.7: $p = 0.334$; Gemini Pro: $p < 0.0001$; Qwen Plus: $p = 0.0040$). The pooled estimates summarize the mean performance of these four models.

Table A17 reports the separate eleven-model logistic specification in Appendix C.9. Its category and standardized continuous contrasts summarize the complete system pool and are not the four-model position-interaction estimates.

Table A17: Adjusted odds ratios from the eleven-model fit with recording-clustered 95% CIs. The first seven rows compare against semantic factual; remaining rows represent +1 SD shifts.

Contrast	OR	Source-clustered 95% CI
Acoustic factual	0.093	[0.080, 0.109]
Acoustic temporal	0.113	[0.092, 0.139]
Semantic temporal	0.150	[0.131, 0.171]
Semantic reasoning	0.332	[0.292, 0.377]
Acoustic reasoning	0.408	[0.300, 0.556]
Acoustic long-range	0.502	[0.374, 0.672]
Semantic long-range	0.804	[0.710, 0.911]
Log duration (+1 SD)	0.787	[0.749, 0.828]
Relative earliest position (+1 SD)	0.866	[0.823, 0.911]
Relative evidence span (+1 SD)	0.918	[0.870, 0.969]
Evidence count (+1 SD)	1.032	[0.987, 1.078]

D.8 Position Diagnostics by Route, Evidence Structure, and Model Pool

We extend the position analysis using the same frozen predictions and recording-clustered inference. Table A18 separates within-task slopes from the temporal–long-range slope contrast. All slopes are reported in accuracy percentage points per 30 minutes of earliest evidence position. Adjusted fits include language, domain, evidence type, log source duration, relative evidence span, and evidence count; long-range and pooled-task fits also include judge route. Covariates constant within a slice are omitted.

Temporal routes. Of the 1,008 temporal questions, 568 use the time-point route, 387 use time-order, and 53 use other routes. Panel A fits separate regressions to the two principal routes. The adjusted negative association is concentrated in time-order questions. A pooled regression with a route-by-position interaction also distinguishes their slopes: the time-point minus time-order contrast is 11.94 pp per 30 minutes (95% CI $[5.20, 18.68]$; $p = 0.00051$). This pooled interaction shares nuisance slopes across routes, whereas Panel A estimates each route separately. The decomposition identifies which temporal items drive the aggregate position pattern.

Long-range evidence structure. The semantic long-range set contains 614 single-evidence and 282 multiple-evidence questions. Their observed position ranges intersect at 10.07–82.73 minutes, retaining 576 and 83 questions, respectively. Within this shared range, mean accuracy is 84.5% and 82.8%; both groups have small estimated slopes (Panel B). A pooled adjusted single-minus-multiple slope contrast is -3.95 pp per 30 minutes (95% CI $[-14.00, 6.10]$). The subgroup estimates are

Table A18: Position slopes and slope contrasts with 95% recording-clustered intervals. Panels A–B use the original four-model item mean. Panel C compares temporal minus long-range slopes over all 2,001 questions. Each row reports its own question count; adjusted and unadjusted columns use the same items.

Slice / model pool	n	Unadjusted slope [95% CI]	Adjusted slope [95% CI]
A. Temporal questions, separated by judge route			
Time-point	568	-2.37 $[-5.41, \ 0.68]$	0.28 $[-3.92, \ 4.47]$
Time-order	387	-12.96 $[-18.07, -7.85]$	-9.88 $[-17.34, -2.43]$
B. Semantic long-range, common position range: 10.07–82.73 min			
Single evidence	576	-2.14 $[-7.00, \ 2.73]$	1.05 $[-5.59, \ 7.69]$
Multiple evidence	83	-0.14 $[-9.16, \ 8.89]$	0.52 $[-16.52, 17.56]$
C. Temporal minus long-range slope, all evidence positions			
Four hosted	2,001	-6.44 $[-10.12, -2.76]$	-5.83 $[-9.73, -1.92]$
Eight hosted	2,001	-6.91 $[-10.09, -3.73]$	-5.19 $[-8.52, -1.86]$

consistent with the aggregate long-range profile; the wider multiple-evidence interval reflects lower precision.

Expanded hosted pool. We additionally include Qwen3.8-Omni-Flash, Qwen3.5-Omni-Flash, Doubao Seed 2.0 Lite, and Muse Spark 1.2, covering all eight hosted systems with their frozen evaluation settings. The temporal–long-range position contrast persists (Panel C; adjusted $p = 0.00223$). Spearman correlations are -0.242 for temporal and 0.011 for long-range, with difference -0.253 (95% recording-cluster bootstrap CI $[-0.348, -0.162]$; 3,000 replicates, seed 20260925). Thus, the pooled contrast extends beyond the original four-model set. The supplied `review_sensitivity.py` reproduces these analyses without additional inference or rescoring.

D.9 Gold-Localized Acoustic Factual Diagnosis

We compare full-recording answers with answers to gold-localized audio crops on the same 200 acoustic factual questions. This diagnostic asks how much difficulty persists when the search region is greatly reduced using the annotated event location.

Question selection and inputs. The subset spans 176 recordings, all 16 languages, and all eight content domains. Before crop inference, the sampler shuffles eligible questions within language-by-sound-category cells (seed 5038) and visits the cells in round-robin order until 200 questions are selected. Eligibility requires an available QA record, finite event timestamps, successful full responses with binary scores for all three models, and a crop longer than one second, at most 180 seconds, and shorter than 80% of the source. The four special post-three-hour items are excluded. Selection does not depend on whether a full response is correct. The resulting subset provides broad category coverage rather than reproducing the natural category frequencies of the 867 acoustic factual questions.

For an annotated event interval $[s, e]$ in a recording of duration T, we extract $[\max(0, s - 15), \min(T, e + 15)]$ seconds, using the earliest and latest evidence timestamps if event endpoints are unavailable. Crops are encoded as 16-kHz mono MP3 at 32 kbps. The source-duration median is 36.5 minutes; the encoded crop-duration median is 31.1 seconds (range 15.1–170.1 seconds). All 200 files decode successfully, and their decoded durations match the requested windows. Each annotated event interval is contained in its crop. The unchanged official question is preceded by an English excerpt notice giving the original-clock start and end and stating that question times refer to the original recording, not clip time zero. No reference answer is supplied to the answering model.

Models and scoring. We use Gemini 3.8 Flash, Gemini 3.7 Flash, and Qwen3.5-Omni-Plus. Full answers reuse the frozen benchmark runs; crop answers were collected on 24 September 2026 using the same nominal model IDs and default decoding settings. The official `judge_v3` scores both conditions against the same references; 199 questions use the single-fact route and one uses time-point

scoring. All 600 full–crop pairs contain successful, nonempty responses. Full and crop use separate requests; crop also introduces the excerpt-specific prefix.

Table A19: Full recording versus gold-localized crop on 200 acoustic factual questions per model. Accuracy is in percent; changes are crop minus full in percentage points. Intervals use 10,000 joint recording-cluster bootstrap replicates. Panel B reports paired correctness counts, not independent samples.

A. Accuracy and paired change

Model	Full	Crop	Change	95% CI
Gemini 3.8 Flash	42.50	44.50	$+2.00$	$[-5.08, \ 9.18]$
Gemini 3.7 Flash	44.00	43.00	-1.00	$[-7.69, \ 5.77]$
Qwen3.5-Omni-Plus	40.00	38.00	-2.00	$[-8.96, \ 4.79]$
Fixed-model mean	42.17	41.83	-0.33	$[-5.10, \ 4.41]$

B. Correctness transitions

Model	Both correct	Crop only	Full only	Both incorrect
Gemini 3.8 Flash	63	26	22	89
Gemini 3.7 Flash	64	22	24	90
Qwen3.5-Omni-Plus	55	21	25	99

Paired inference and persistent errors. For the fixed-model mean, we average the three binary outcomes per question before averaging questions. We resample recordings 10,000 times (seed 20260925), retaining all questions, models, and paired conditions from each sampled recording. Percentile intervals therefore preserve within-recording and cross-model dependence; they describe variation across the diagnostic sample, not repeated-generation uncertainty. Table A19 shows no consistent aggregate improvement, while 23–24% of questions per model change correctness in either direction. Separate requests without repeated generations do not identify which flips arise from input changes or response variation.

Across the three models, 278 of 347 incorrect full responses remain incorrect under cropping: 80.1% (95% CI $[74.6, 85.3]$). The per-model fractions are 77.4%, 80.4%, and 82.5%. Moreover, 64/200 questions are incorrect for all three models under both conditions, spanning 60 recordings, all 16 languages, and all eight stored sound categories. This shared set is an observed difficulty profile of the evaluated systems. The result shows that substantial acoustic-factual difficulty persists under gold-localized input.

Window and route sensitivities. Although every annotated sound event is inside its crop, five questions have textual anchors or annotated evidence extents reaching beyond a crop boundary (one by only 0.85 seconds). Removing all five leaves 195 questions and nearly the same mean change (-0.34 pp), with 63 questions still incorrect in all six responses. Excluding the single time-point item or the two crops longer than 60 seconds also leaves the result essentially unchanged (Table A20). These are metadata-based checks of coverage and scoring-route composition.

Table A20: Sensitivity of the fixed three-model mean to question filtering. The last column counts questions incorrect for all three models under both conditions. All rows retain the original scores.

Subset	n	Full (%)	Crop (%)	Change [95% CI]	All six wrong
All questions	200	42.17	41.83	$-0.33 \ [-5.10, 4.41]$	64
Single-fact route	199	42.04	41.71	$-0.34 \ [-5.14, 4.42]$	64
Crop ≤ 60 s	198	42.09	41.75	$-0.34 \ [-5.15, 4.40]$	64
All evidence inside	195	41.88	41.54	$-0.34 \ [-5.06, 4.36]$	63

Exploratory category variation. Table A21 reports all eight stored sound groups. Music questions improve by 17.5 pp on average, whereas laughter questions decline by 11.6 pp; each of the three models follows these respective directions. The category intervals are pointwise and exploratory. Music and laughter crops also differ in median duration (50 versus 31 seconds), and question construction is not matched across categories. These results describe heterogeneity without isolating

its mechanism. Reweighting the same questions equally by language gives a mean change of -0.35 pp; equal weighting of the eight sound groups gives $+4.01$ pp. Thus, the overall mean depends on category composition, while substantial residual difficulty remains in either weighting.

Table A21: Exploratory full–crop results by stored sound category, averaged over the same three models. Intervals are pointwise 95% recording-cluster intervals without multiplicity adjustment. The six animal questions come from four recordings.

Category	n	Full (%)	Crop (%)	Change (pp)	95% CI
Laughter	43	58.14	46.51	-11.63	$[-22.50, -1.59]$
Cough	41	51.22	43.09	-8.13	$[-20.72, 4.17]$
Applause	38	51.75	51.75	0.00	$[-9.30, 9.09]$
Music	21	38.10	55.56	$+17.46$	$[1.52, 32.14]$
Other acoustic	21	22.22	26.98	$+4.76$	$[-4.00, 15.00]$
Alarm	17	15.69	23.53	$+7.84$	$[-6.25, 24.45]$
Speech overlap	13	17.95	23.08	$+5.13$	$[-6.06, 19.05]$
Animal	6	16.67	33.33	$+16.67$	$[0.00, 66.67]$

Traceable response examples. In a German question asking what is heard at 26:38, the reference is a man's laughter. With the 26:23–26:54 crop, both Gemini models explicitly map the target to clip second 15 but answer yawning or breathing; Plus answers with spoken hotel-room details. All six full/crop answers are scored incorrect. Conversely, a Vietnamese question asks what follows a quoted drill price: all three full answers are incorrect, while all three crop answers identify the reference bell sound. These comparisons illustrate persistent and recovered answers under the existing scoring. The supplementary `fullcrop_diagnostic/` directory contains the frozen question list, paired responses, prompts, crop metadata, case identifiers, and analysis code.

E Prompt Templates and Verification Rules

This appendix presents the benchmark-relevant prompt specifications used in construction and evaluation. Line breaks and typographic formatting are normalized for readability, while the operational constraints are unchanged. Complete source templates, calibration examples, and implementation-level formatting are provided in the accompanying `prompts/` directory, with a manifest mapping each presented stage to its source symbol.

E.1 Semantic Evidence and QA Construction

The semantic construction calls use GPT-5.6-Sol with temperature 0, high reasoning effort, and an 8,192-token completion budget in the shared caller. Large transcripts are processed in 40-minute slices. These construction inputs are privileged annotations; evaluated models receive only the audio and the question.

Atomic evidence mining. The mining prompt asks for atomic, self-contained, checkable statements that are explicitly supported by timestamped transcript cues. Each fact remains in the source language, retains its supporting cue time, and is assigned a coarse information type and information-density label. The prompt explicitly rejects inference, unsupported merging of distant statements, filler, and facts that cannot be grounded to the supplied transcript.

Evidence-plan construction. For released items, grounded facts are organized into plans for direct retrieval, enumeration, reasoning, temporal localization, and long-range retrieval. Multi-evidence reasoning plans must connect facts through a concrete causal, comparative, temporal, or arithmetic relation rather than mere topical co-occurrence. Long-range semantic plans follow the evidence position rule in Eq. 3. At this stage the model proposes only the evidence structure; it does not write the final question or reference answer.

Question generation. A separate call converts one evidence plan into a single natural question in the source language. The operational constraints require one defensible answer, forbid timestamps

or answer keywords in the stem, and prevent explicit leakage of the intended reasoning operation or evidence locations. Questions must be audio-specific rather than answerable from generic world knowledge alone, and the target event or entity must be uniquely identifiable from the recording.

Task- and route-specific variants. The targeted variants preserve the same grounding constraints while adapting the requested answer form. Exhaustive questions make completeness explicit; multi-slot questions request distinct required fields; recording-clock items ask when a uniquely identified event occurs without revealing its position; and event-order questions uniquely scope the target events while asking only for their requested order. Table A2 summarizes the resulting answer requirements.

Reference-answer derivation. The reference answer is produced in a fresh context from the question and its grounded evidence. The answer must use only the supplied evidence, respond directly to what the question asks, preserve required slot or list completeness, and avoid unsupported background knowledge. Route-specific formatting follows the stored construction target; for metric temporal items, this includes an explicit recording time. Verbatim prompt templates and task-specific substitution dictionaries are included in the supplementary artifact.

E.2 QA Verification, Repair, and Shortcut Checks

Candidate verification and minimal repair. The verification stage checks whether the candidate question and reference answer are supported by the supplied transcript context and whether the item has a unique defensible answer. It also checks answer leakage, method leakage, missing required content, question–answer mismatch, unsupported additions, and temporal-format requirements where applicable. Repairable defects receive a minimal correction that preserves the intended task, followed by reverification; fatal grounding failures are removed during construction. The complete verifier and repair templates are provided verbatim in the supplementary prompt files.

Question-only shortcut check. The following two prompts implement the question-only ablation described in the main text. They are retained here because their acceptance condition directly affects released-item filtering.

Question-only solver

```
Answer this question using ONLY your prior knowledge -- you have NOT
  heard the audio. Do your best guess. Language: {lang}.

Question: {question}

Return ONLY JSON: {{"guess": "...", "confidence":
  "low|medium|high"}}
```

Question-only answer comparison

```
A model answered a question WITHOUT access to the audio. Decide
  whether its guess is materially CORRECT (same core answer as the
  reference).

question: {question}
reference: {answer}
model guess (no audio): {guess}

Return ONLY JSON: {{"correct_without_audio": true|false, "reason":
  "..."}}
```

Transcript-only acoustic checks. The solver uses the question and full transcript, permitting both sound clues in the question and cross-passage textual reasoning. An item is rejected when the solver reports a supported answer and the comparison stage confirms a material match to the reference (Eq. 5). The complete solver and comparison templates are provided in the supplementary `prompts/` directory as `P_ABLATE_AC.txt` and `P_ABLATE_JUDGE_AC.txt`, respectively.

E.3 Acoustic Evidence Verification

Full-recording event proposals. TimeAudio applies the following detection prompt to successive audio windows; local output timestamps are mapped back to the original recording clock.

Full-recording event proposal

```
<Speech><SpeechHere></Speech> What events can you detect in the
  sound with the start and end timestamps?
```

Source video can assist annotation, but an event must be audible to support an audio question. Proposal-hidden observation prevents the candidate label from being included in the first-pass prompt. The subsequent alignment pass does not add missed events. A later localized check rejects claims that remain uncertain. This annotation-stage use of video is separate from audio-only benchmark inference.

Proposal-hidden observation

```
The clip corresponds to original recording
  {window_start}-{window_end},
approximately 10 seconds. Clip time t maps to original time
  window_start+t.

Watch and listen. List the non-speech acoustic events actually
  present.
No candidate labels are provided. Visual context may confirm or
  exclude a
source, but the event must also be audible; do not infer a sound
  from vision.

Report laughter, crying, applause, coughing, sneezing, sniffing,
  breathing,
humming, and crowd noise, but not ordinary speech or narration. Also
  report
distinct impacts, friction, alerts, effects, running water, music,
  vehicles,
and similar events. Represent continuous ordinary recording noise or
  room
hum once as ambient noise; do not assign an unsupported object
  identity.

Name conservatively. Undifferentiated low hum, rustling, or
  whooshing is
ambient noise unless identifiable water, melody, fan rhythm, or horn
  cues
support a specific source. For a short ding/pop/click/beep, name its
  mechanism
only when identifiable; otherwise use a short sound effect or
  electronic
alert. Distinguish paper from plastic only when clear; otherwise say
  object
friction. If a high-frequency sound could be breathing or a bird,
  report that
ambiguity. A short added sound out of sync with the scene may be
  described
as a post-production effect.

Do not omit faint laughter, applause, crying, or crowd reactions
  just because
speech dominates; report low confidence when appropriate.

Return JSON only:
{"heard":[{"time":"00:25-00:33","desc":"...",
           "confidence":"high|low"}],
```

Proposal-hidden observation (continued)

```
 "note":"one sentence describing the non-speech layers"}
```

Candidate alignment

```
Original window: {window_start}-{window_end}. Local t maps to
  window_start+t.
Your proposal-hidden observations: {heard_json}
Candidate labels, now revealed, in original recording time:
  {candidate_labels}

Align each candidate to your observations. Watch and listen again,
  but do not
reinterpret the audio as a specific category that appears only in
  the label
and was absent from your observations. Use visual context to check
  sources;
visible actions can support an already heard event.

- If observations match the core category: keep and repeat the
  label.
- If observations identify another category: modify to the heard
  description.
- If only speech, silence, or ordinary noise is present and the
  claimed event
  is absent: delete and mark no such event.
- If alignment is uncertain or the sound is weak: keep rather than
  guess.

Asymmetric rules:
1. Laughter/applause/crying/crowds: retain conservatively if heard
  or plausibly
   present. Delete only when the entire clip clearly contains speech
     without
   the claimed independent event. Avoid swapping these categories
     unless sure.
2. Water/wind/background music/fan/car/microwave: if observations
  contain only
   ambient noise or lack identifying cues, do not retain the
     specific label.
   Modify to ambient noise or delete.
3. Paper/bird/phone/glass/plop/beep/shuffling: if observations say
  friction,
   high-frequency sound, or a generic/post-production effect, use
     that
   description instead of retaining an unsupported specific name.

Do not add new events or fill missing annotations. Return JSON only:
{"items":[{"time":"00:25-00:33","label":"original label",
 "heard_match":"matching observation or none",
 "verdict":"correct|incorrect|hallucinated|uncertain",
 "action":"keep|modify|delete","corrected":"...","reason":"..."}]}
```

Final localized audibility check

```
This is a real clip from the original recording (window
  {w0:.0f}-{w1:.0f}s).
Watch the picture and listen. Do not invent sounds.

Check each claimed acoustic evidence. The kind_hint is WEAK -- do
  not trust it.
Confirm ONLY if the claimed TEXT describes a sound that is actually
```

Final localized audibility check (continued)

```
  audible.
- Speech-only / "speaking indoors" is NOT a door / crowd / cheer.
- "train" is NOT rain. "cheerfully" is NOT cheer. "indoors" is NOT a
  door slam.
- Confirm music / laugh / wind / car / bird only if you actually
  hear that sound.
- Confusion pairs -- reject if the clip is the lookalike, not the
  claim:
  bird != child cry / wood knock / machine / tape fast-forward /
    BGM;
  water != writing on a board / frying / metal hit / BGM;
  crying != sniffle / spoken interjection / laugh / sigh;
  cough != speech stutter / shout / laugh / pant / edit glitch;
  laugh != beep / electronic scale / electrical hum / smile with no
    sound;
  applause != knocking; wind != splash / rain / dog.
- A visual smile or a "crying" face is NOT an audible event.
- If you cannot hear it, reject. Uncertain -> reject.

Return ONLY JSON:
{{"items":[{{"id":"a1","verdict":"confirm","heard":"what you
  heard","reason":"..."}}]}}

CLAIMS:
{claims}
```

E.4 Inference Wrapper and Scoring Inputs

Models receive audio under the input policy in Appendix C and the source-language wrapper followed by the original question. No extra system instruction, chain-of-thought request, transcript, evidence, or gold answer is appended. All 16 original language wrappers are preserved in `prompts/inference_wrappers.txt`; the English wrapper follows.

Inference wrapper

```
Please answer the following question based on the full audio
  content:
{question}
```

E.5 Six Routed Judge Instructions

The frozen evaluator uses GPT-5.6-Sol: low reasoning effort for single-fact scoring and medium for the other routes. The stored judge-type field takes precedence over fallback routing. The following instructions expose the acceptance and rejection conditions; complete source templates and few-shot calibration outputs are retained separately. The judge receives the question, reference answer, and model prediction; an optional evidence field defaults to "not provided" in the supplied semantic builders.

Across semantic routes, score the final committed answer, allowing translation and semantic equivalence rather than exact string matching. Do not penalize abandoned intermediate attempts by themselves. Unresolved competing answers, contradictions of the required answer, or explicit denial of audio information followed by an external-knowledge guess receive no credit. A review flag marks uncertainty and is not itself an automatic change to the frozen binary score.

Single-fact judge

```
You are a strict but semantically tolerant long-audio QA evaluator.
Determine whether the final committed answer gives the correct core
  fact.

1. Define the core fact from the question and gold. Allow synonyms,
   translations, abbreviations, equivalent units, reasonable
     approximations,
   homophones, and transcription variants when identity is
     unambiguous.
2. Do not demand explanatory gold details. Nonconflicting
  elaboration is not
   an error. Reject unresolved competing alternatives and changes to
     the core.
3. Distinguish entity identity, surface form, and whether the
  question requires
   an exact name/code/term/quotation. An identifiable equivalent
     entity passes
   when exact wording is not required. When exact wording is
     required, a
   merely similar-sounding form does not automatically pass;
     case-only
   differences are acceptable. Extra model/product suffixes belong
     to
   extra_claims unless they change the entity's identity.
4. If evidence is absent, do not guess whether added context is in
  the audio;
   check conflict with the core answer. If supplied evidence does
     not cover a
   core claim, flag review because the evidence may be incomplete.
5. Reject an explicit audio-information denial followed by an
  external guess.

Return JSON with result (0/1), matched_claims, missing_claims,
entity_identity_match, surface_form_match, exact_name_required,
contradictions, competing_answers, extra_claims,
abstained_or_denied_audio, review_flag, and reason (at most 50
  characters).
Result is 1 only if no required core fact is missing, no
  contradiction or
unresolved competitor exists, and the audio-denial condition is
  false.

Question: {question}
Gold: {gold}
Evidence: {evidence_or_not_provided}
Model answer: {model_answer}
```

Enumeration judge

```
Judge the final submitted item set, ignoring abandoned intermediate
  attempts.

1. Decompose gold into distinct required items; merge synonymous
  expressions.
   Match model items semantically; order alone does not matter.
2. Determine whether the question explicitly requests all items. If
  exhaustive,
   every required item must be covered. For an open-ended
     enumeration,
   unmentioned gold items are missing_minor_items and do not alone
     fail it.
   Do not count synonymous descriptions as separate items to meet a
```

Enumeration judge (continued)

```
     number.
3. Check requested generation/stage/role/entity/index
  correspondences.
   Even if all names appear, a mismatched correspondence fails the
     answer.
4. Extra gold-external items: if explicitly supported and in
  category, flag
   review without automatic failure; if contradicted, off-category,
     or wrong,
   reject. Without evidence, put unverifiable relevant additions in
   extra_claims and flag review rather than automatically rejecting
     them.
5. Reject unresolved competing candidates or denial of audio
  followed by an
   external guess. Flag core items outside supplied evidence for
     review.

Return JSON: result, exhaustive_required, matched_claims,
  missing_claims,
missing_minor_items, matched_relations, relation_mismatches,
  contradictions,
competing_answers, extra_claims, abstained_or_denied_audio,
  review_flag, reason.
Nonempty relation_mismatches, contradictions, or competing_answers,
  or audio
 denial, implies 0. Missing required items imply 0 when
   exhaustive_required.

Question: {question}
Gold: {gold}
Evidence: {evidence_or_not_provided}
Model answer: {model_answer}
```

Multi-slot judge

```
Decompose the question into independent required slots using
  question and gold.
Judge the final answer for every slot, not abandoned intermediate
  attempts.

Every required slot must have a correct, explicit corresponding
  answer.
Missing slots, wrong associations, and wrong values fail. Allow
  semantic,
translation, abbreviation, homophone, and nonessential modifier
  differences.
Search the entire response, including explanations, for each slot's
  answer.
Answering one slot negatively does not negate another slot: for
  example,
"the individual has no duty" can coexist with "the provider has the
  duty."

An unresolved set of candidates does not pass just because it
  contains the
correct one. Put nonconflicting explanations in extra_claims,
  competing slot
answers in competing_answers, and changed/wrong relations in
  contradictions.
With no evidence, do not judge whether background came from the
  recording.
With evidence, flag uncovered core slot claims for review. Reject
```

Multi-slot judge (continued)

```
  explicit
audio-information denial followed by an external-knowledge guess.

Return JSON: result, matched_claims, missing_claims, contradictions,
competing_answers, extra_claims, abstained_or_denied_audio,
  review_flag, reason.
Return 1 only when all required slots match and no rejection
  condition holds.

Question: {question}
Gold: {gold}
Evidence: {evidence_or_not_provided}
Model answer: {model_answer}
```

Reasoning judge

```
Judge the final conclusion and the key relations that determine it.
Ignore abandoned attempts and incidental intermediate imperfections
  if the
final conclusion and necessary reasoning nodes are correct.

1. Extract the conclusion and necessary reasoning nodes from
  question/gold.
   If only a conclusion is requested, do not demand a full
     explanation.
   If causes, a procedure, N grounds, or steps are requested,
     require the
   decisive reasons/relations/steps. N grounds must be genuinely
     distinct.
2. Preserve critical quantities, prior intent/timing, participants,
  locations,
   and mechanisms. Mentioning a node while weakening or changing its
     essential
   elements counts as missing_required_nodes. Pure paraphrase is
     acceptable.
3. Distinguish missing_required_nodes from missing_minor_details;
  only the
   former fail. Do not treat duplicated phrasing as independent
     grounds.
4. Added parallel core reasons that change the answer go to
   alternative_core_reasons and fail. Nonconflicting
     background/intermediate
   stages go to nonconflicting_context and do not fail if all key
     nodes remain.
   Wrong conclusions, reversed causality, and wrong core entities
     are
   contradictions. Incidental intermediate mistakes alone do not
     fail.
5. Without evidence, do not use common knowledge to certify added
  reasoning;
   check whether it replaces/distorts the required chain. Flag
     suspected gold
   incompleteness or core claims outside supplied evidence for
     review.
6. Reject audio-information denial followed by an external-knowledge
  answer.

Return JSON: result, matched_required_nodes, missing_required_nodes,
missing_minor_details, alternative_core_reasons,
  nonconflicting_context,
contradictions, abstained_or_denied_audio, review_flag, reason.
Return 1 only with no missing required nodes, alternative core
```

Reasoning judge (continued)

```
  reasons,
contradictions, or audio denial. The original template includes five
calibration examples distinguishing these fields (retained in the
  source file).

Question: {question}
Gold: {gold}
Evidence: {evidence_or_not_provided}
Model answer: {model_answer}
```

Event-order judge

```
Order mode: {order_mode}
Require associated audio timestamps: {require_associated_times}

mention_order means first mention in the recording; event_chronology
  means
actual date/time in the story or world. Use the requested mode.
1. Read the model's explicit numbering/arrows/list as its submitted
  order.
   Do not silently reorder it using accompanying timestamps.
2. Reject contradictions between explicit order and its own
  dates/times,
   marking internal_inconsistency=true.
3. Allow equivalent event names. All requested events must appear;
  missing,
   substituted duplicates, or incorrect order fail.
4. Do not substitute story chronology for mention order, or
  recording times
   for story dates. If timestamps are not required, small timing
     errors alone
   do not fail; explicit internal contradictions still do.
5. If associated timestamps are required, output event-time pairs
  and leave
   tolerance checks to the program rather than computing +/-3
     seconds.

Return JSON: result, gold_sequence, model_sequence, missing_events,
extra_events, order_correct, internal_inconsistency, time_pairs,
review_flag, reason. Each time pair has event, gold_time,
  model_time,
semantic_match. The LLM layer passes only complete, correct,
  consistent order;
when required, final success also requires the programmatic time
  check.

Question: {question}
Gold: {gold}
Model answer: {model_answer}
```

Event–time alignment judge

```
You pair events with time expressions; do not compute timing error.
Extract event-linked points/intervals from gold and model answers
  and match
by event semantics, never greedily by nearest time alone.

1. Accept HH:MM:SS, MM:SS, minutes/seconds, total seconds, and
  equivalent
   expressions in other languages; preserve the original
     expressions.
```

Event–time alignment judge (continued)

```
2. Distinguish points from intervals; retain both interval
  endpoints.
3. If an event lacks a model time, use an empty model_time and
  missing=true.
4. Set semantic_match=false when a time is assigned to the wrong
  event.
5. Do not decide the +/-3-second tolerance or output a computed
  error.
6. For a point question answered with a broad interval, preserve the
  interval;
   do not pick whichever endpoint is closest.

Return JSON:
{"question_mode":"point|interval|multi_point",
 "pairs":[{"event":"gold event","gold_time_type":"point|interval",
 "gold_time":"original expression",
 "model_time_type":"point|interval|missing",
 "model_time":"original expression or empty",
 "semantic_match":true,"missing":false}],
 "unmatched_model_times":[],"review_flag":false,"reason":"..."}

Question: {question}
Gold: {gold}
Model answer: {model_answer}
```

Programmatic timing and frozen routing. After semantic pairing, the frozen scorer applies the point/interval rules specified in Appendix C. Routing and timing rules are fixed before inspecting model answers and are applied consistently across models.

F Qualitative Cases, Release, and Limitations

F.1 Illustrative Response Cases

The following cases illustrate response differences among the evaluated models.

Timestamp errors for a specified event (Turkish). The question asks when a stitch count increases from 63 to 69 after an increase in the third pattern of a how-to recording. The reference timestamp is 16:06; model responses report 15:57, 16:26, or 14:32, corresponding to absolute errors of 9, 20, and 94 seconds. All exceed the scoring tolerance of ± 3 seconds. With the event description supplied in the question, this case illustrates errors in locating its position on the recording clock.

Natural event confusion (Portuguese and Italian). For a Portuguese factual question about an event near 33:48, the verified event is a man's laugh. Gemini 3.8 Flash, Gemini 3.7 Flash, Qwen3.5-Omni-Plus, and Doubao identify laughter; Gemini 3.1 Pro instead reports victory music from *Final Fantasy*. In an Italian how-to item targeting a short laugh, responses propose a mouse click, cough, or notification sound. These cases illustrate local event-grounding errors in natural recordings.

Entity recovery in long-range retrieval (Korean). The recording gives a past sentence of "2 years 6 months with 4 years suspended." Several models recover the pair, while one changes the prison term to 1 year 6 months and another selects a separate life-imprisonment conviction mentioned in the same recording. The item distinguishes precise entity recovery from grounding to a nearby competing event.

F.2 Release Scope and Provenance

The benchmark artifacts comprise questions, reference answers, localized evidence, task metadata, and evaluation metadata. Source provenance supports traceability. Original recordings were collected from publicly accessible online media; copyright remains with the respective rights holders. Redistribution of source audio is subject to source-specific permissions.

F.3 Limitations

MuLA-Bench controls semantic question allocation across languages and domains, but it does not hold recording content or speakers identical across languages. Acoustic availability and task mixtures remain natural and uneven; acoustic and semantic tracks are not matched interventions. Long-range scores should not be read as general multi-hop reasoning performance. Local context limits and response failures remain part of the end-to-end score.