CABINBENCH: BENCHMARKING WHAT LONG-AUDIO AGENTS ACTUALLY HEARD

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio understanding has advanced on clips of seconds, where the answer and its evidence arrive in the same model call. Long recordings break that pairing: the evidence for a question may last three seconds inside an hour, so a correct answer may have been heard or guessed, and benchmarks that hand the model the whole recording and score only the answer cannot say which. A vehicle cabin is a natural testbed: over a trip of tens of minutes, several occupants talk over road noise, phone calls and navigation prompts cut in, and plans change, so understanding the scene means finding, tracking, and attributing evidence. CabinBench is, to our knowledge, the first benchmark that scores the listening process itself rather than the answer alone: the agent decides where to listen under a budget, must cite audio for every claim, and is credited only for citations that the harness's delivery log confirms. Its recordings are synthesized from fictional scripts by design, which makes the ground truth fully controllable: every question carries an exact evidence contract, and any sentence can be removed, replaced, or moved to produce a counterfactual recording whose truth is known. On 200 recordings and 2,608 questions, a 30B-parameter open omni-modal model answers 16 to 20% of questions yet establishes its evidence in under 1% of runs; most of its correct answers never heard the sentence that settles the question, and a tenth of its submissions cite audio it was never given. Controllable data, an agent in the loop, and process-level scoring turn listening from an assumption into a measurement.

1 INTRODUCTION

On a long recording, a correct answer is weak evidence that the model listened. The sentence that settles a question may last a few seconds, be revised minutes later, or never be spoken, so a model can often reach the right answer from an early plan or from what sounds plausible. In Figure 1, a passenger's plan is corrected at 18:42–19:08 and confirmed at 27:15–27:40, and both agents answer *north entrance*. Agent A, *verified*, takes its evidence span from a listened audio view of the correction: evidence sufficient. Agent B, *guessed*, cites the early plan (cited span 04:10–04:38), which contradicts its answer: evidence insufficient. Scoring the answer credits both.

Answer-only scoring cannot separate these agents. Prior work measures long-audio understanding, temporal localization, active media access, and abstention in different settings (Section 2). The selected protocols in Table 1 do not jointly make a final claim's cited audio and its delivery in the same charged run a condition of credit. This motivates measuring the search trajectory alongside the answer rather than inferring evidence use from correctness.

CabinBench rewards listening rather than answers, through three design choices. The agent must *seek* its evidence: it receives the full recording without a transcript, chooses which intervals to hear under a budget of audio seconds and tool calls, and ends by answering or abstaining. It must *cite* an interval for every claim, and a hidden evidence contract per question states what the citations must cover: the sentences that settle the question with their speaker and source, the counter-evidence to resolve, and whether the question is answerable. Finally, the harness *verifies* each citation against its own delivery log, accepting a submission only if every cited interval reached the agent in that run. The primary metric, Evidence-Constrained Success (ECS), therefore holds only when answer, evidence, attribution, abstention, and budget all pass. We synthesize the audio from fictional scripts so intended speakers, sources, and statements can be registered before rendering; forced alignment

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.

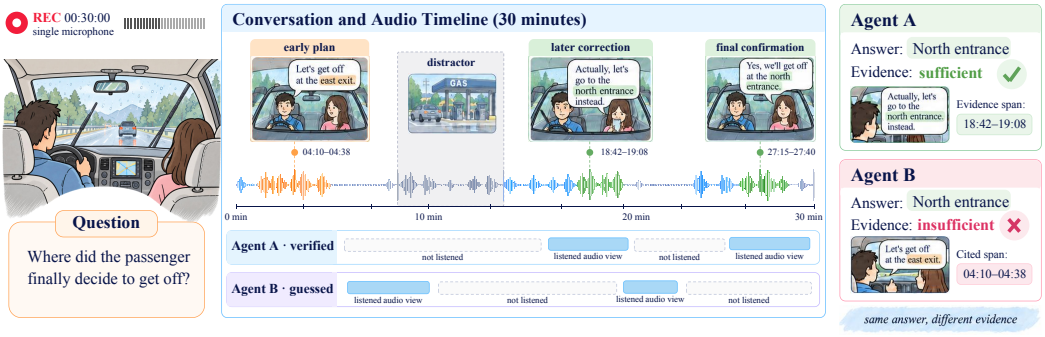


Figure 1: An illustrative CabinBench episode. Both agents answer *north entrance*. Agent A (verified) takes its evidence span, the later correction, from a listened audio view: evidence sufficient. The cited span of Agent B (guessed) is the early plan, which contradicts its answer: evidence insufficient. Answer-only scoring treats the two runs alike. CabinBench credits A and fails B.

Table 1: Audio and long-context understanding benchmarks, abridged from Appendix A. **Dur**: audio length per item, the longest reported or else the mean. **Tx**: transcript provided. **Seg**: pre-segmented before the model sees it. **Bud**: listening or retrieval budget as a controlled variable. **Int**: an evidence interval is required. **Ver**: the cited interval is verified to have been accessed. **Abs**: abstention is evaluated. †: arXiv preprint, not peer reviewed. A dash means the feature is not required by the cited default evaluation, not that a system is incapable of it. Parentheses: a budget imposed by the evaluator or fixed as a method hyperparameter. HARP is an agentic retrieval method included for context; * denotes accepted to SLT 2026 with proceedings forthcoming.

Benchmark	Venue	Dur	Tx	Seg	Bud	Int	Ver	Abs
Short-clip audio understanding								
MMAU (Sakshi et al., 2025)	ICLR'25	10 s	–	✓	–	–	–	–
MMAR (Ma et al., 2025)	NeurIPS'25 D&B	30 s	–	✓	–	–	–	–
AQUA-Bench (Kuan & Lee, 2026)	ICASSP'26	clip	–	✓	–	–	–	✓
Long-form audio								
CORAAL QA (Shankar et al., 2024)	ICASSP'24	60 min	✓	✓	–	–	–	–
AudioMarathon† (He et al., 2025)	arXiv'25	5 min	–	–	(✓)	–	–	–
TAG-Bench† (Dai et al., 2026)	arXiv'26	20 min	–	–	–	✓	–	–
LAT-Bench† (Shao et al., 2026)	arXiv'26	30 min	–	–	–	✓	–	–
VoiceGiraffe† (Ye et al., 2026)	arXiv'26	2 h	–	–	–	–	–	–
HARP (Li et al., 2026a)	SLT'26*	75 min	–	–	(K=3)	✓	–	–
In-car audio								
AISHELL-5 (Dai et al., 2025)	INTERSPEECH'25	ASR	–	✓	–	–	–	–
CabinBench	This work	2 h	–	–	✓	✓	✓	✓

estimates where those statements land in the waveform. Actual speech and identity still require audio-side audit. The recordings were generated after the evaluated model checkpoints, reducing exact-item exposure without ruling out content priors.

We show that at a quarter budget, most correct answers of an open omni-modal model, Qwen3-Omni, did not come from the evidence: 62% of them never heard any sentence that settles the question. With the full recording, the harness rejected 10.9% of its runs for citing audio it never delivered, and 107 of those rejected submissions had the right answer. The full budget changes what the model does more than what it gets right: abstention falls from 26.0% to 5.5% while accuracy rises from 16.4% to 19.6%.

We contribute (1) **a protocol** that turns "did the agent hear its evidence?" into a quantity read from a log (Section 3); (2) **a candidate corpus** of 200 synthetic Mandarin in-cabin recordings (138

hours) and 2,608 questions with sentence-exact evidence (Section 4); and (3) **a diagnosis** of open omni-modal models that separates correct answers from heard evidence (Section 5).

2 Related Work

2.1 Long-audio understanding and grounding

SD-Eval, MMAU, and MMAR evaluate seconds-scale audio (Ao et al., 2024; Sakshi et al., 2025; Ma et al., 2025); AIR-Bench, Dynamic-SUPERB, and SAKURA cover broader audio tasks (Yang et al., 2024; Huang et al., 2024; Yang et al., 2025). Audio Flamingo 2 and MMAU-Pro reach ten minutes at most (Ghosh et al., 2025; Kumar et al., 2026). Longer inputs appear in transcript-based meeting tasks (Zhong et al., 2021; Prasad et al., 2023), CORAAL QA's retrieval over transcribed 60-second segments (Shankar et al., 2024), constructed ten-minute LongSpeech items (Yang et al., 2026), and up-to-two-hour VoiceGiraffe recordings (Ye et al., 2026). AudioBench aggregates multiple audio tasks (Wang et al., 2025); Video-MME evaluates long video with optional audio and subtitles (Fu et al., 2025). LongAudioSpan extends the duration and depth of clip-style audio comprehension benchmarks (Huang et al., 2026). Fixed-input answer accuracy in these settings does not by itself reveal which acoustic evidence an interactive agent used.

2.2 Temporal localization and active access

SLUE named-entity localization (Shon et al., 2023), audio moment retrieval (Munakata et al., 2025; 2026), temporal grounding (Dai et al., 2026; Shao et al., 2026), and LongVALE's event boundaries (Geng et al., 2025) evaluate predicted intervals. Agentic systems make access a policy: HARP retrieves a fixed top three segments per evidence set (Li et al., 2026a), LAT-Audio iteratively crops audio (Shao et al., 2026), OmniGAIA permits on-demand media regions (Li et al., 2026b), and T* searches video frames under an inference budget (Ye et al., 2025). AudioMarathon studies compute efficiency (He et al., 2025), while HAL records agent costs across tasks (Kapoor et al., 2026). These works evaluate complementary aspects of access, localization, and cost. CabinBench additionally requires each final audio citation to overlap the right speaker/source evidence and to be present in that run's confirmed delivery log.

2.3 Abstention, attribution, and in-cabin audio

SQuAD 2.0 evaluates unanswerability in text (Rajpurkar et al., 2018); AQUA-Bench adapts audio multiple-choice questions with a "none of the above" option (Kuan & Lee, 2026). AbstentionBench studies refusal across text tasks (Kirichenko et al., 2025), and Trust-Score studies citations and refusal in text RAG (Song et al., 2025). CabinBench asks whether an abstention follows a bounded search of a long recording. Real in-cabin corpora such as AISHELL-5 and ICMC-ASR primarily target recognition and diarization (Dai et al., 2025; Wang et al., 2024); our setting instead evaluates claims about the scene, speaker, and source.

3 Task, Protocol, and Harness

3.1 Setup, actions, and cost

Task input (A). An episode of the loop in Figure 2 pairs a continuous, unsegmented recording A of length T with a question q and a listening budget B. The agent receives the question, the recording's duration and format, the speaker roster (identifier, role, source type), its budget, and a response schema that fixes the submission

$$\hat{z} = (\texttt{decision}, \texttt{answer}, \texttt{claims}, E^{+}, E^{-}), \tag{1}$$

whose *claim slots* name the propositions an answer must support without revealing their values. The harness passes nothing else: no transcript, contract, stems, or evidence location. All results use the `raw_listening` track, which forbids global semantic preprocessing (Appendix A.3).

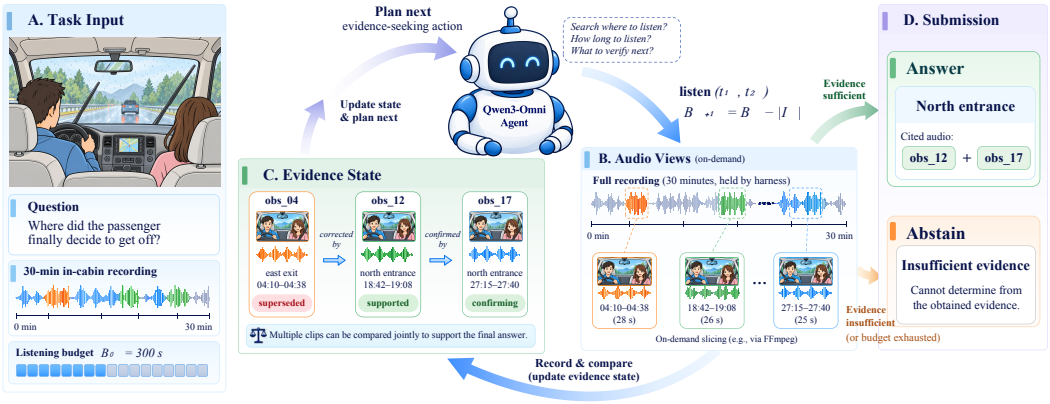


Figure 2: The evaluation loop: task input with a listening budget B_0 (A), audio views that the harness cuts on demand (B), the evidence state the agent records and compares (C), and the submission (D). The budget shown is illustrative.

Planning and audio views (B). The agent plans its next evidence-seeking action (where to listen, how long to listen, what to verify next) and calls `listen`(t_1, t_2) on a half-open interval of the final waveform. The harness cuts the interval with FFmpeg from the full recording it holds and delivers it as an *audio view*. Actions and observations form a trajectory $\tau = (a_1, o_1, \ldots, a_n, o_n)$ with cost vector

$$C(\tau) = (C_{\text{unique}},\ C_{\text{total}},\ N_{\text{tool}}), \tag{2}$$

the union length of intervals visited, the sum of every access including repeats, and the number of content-bearing calls. The budget $B = (B_{\text{audio}}, B_{\text{tool}}) = (\beta T, N_\beta)$ constrains C_{total}: its audio part is the listening budget $B_0 = \beta T$, and each listen to an interval I lowers it, $B_{t+1} = B_t - |I|$, so re-listening is never free. We use $\beta \in \{0.25, 1.00\}$ and set N_β in Section 5. An action that would exceed the budget is refused before any audio is delivered (`budget_exceeded`) and costs a tool call but no audio.

Evidence state (C). After each audio view the agent records and compares what it heard in its *evidence state*: each finding with its value, its audio view, and whether it supports or counters the answer. In Figure 2, the east exit (04:10–04:38) is *superseded*, corrected by the north entrance (18:42–19:08), which is *supported* and has a *confirming* entry (27:15–27:40). Our agent has only `listen`, `record`, and `submit`; lacking the `compare` of the full set (Table 6), it compares entries in its own reasoning.

Submission (D). The decision is `answer` or `abstain`. An answer declares the evidence sufficient and gives *cited audio* for every claim, an interval and the audio view that delivered it (`obs_12` and `obs_17` in the figure). Supported and confirming entries form the support E^+; superseded ones, the counter-evidence E^- that the answer resolves. An abstention states a reason, such as insufficient evidence or an exhausted budget, and is correct only for an unanswerable question (`absent`, `ambiguous`, or `conflicted`; Section 4.4). If an answer exists but goes unfound within the budget, the run fails.

3.2 Evidence Contracts and Scoring

Scoring credits a run only when its cited audio covers the sentences its contract names.

Evidence contracts. Every question has a private evidence contract, the scoring ground truth. It fixes the answer and the answerability, one or more legal minimal sufficient support sets, the counter-evidence that must be addressed and the resolution the contract accepts, and distractor intervals. Each evidence item, a *gold sentence*, carries a sentence-level interval, a speaker, and a source type.

Scoring. Let P be the union of intervals a submission cites and G the union of one legal support set. With temporal precision $|P \cap G|/|P|$ and recall $|P \cap G|/|G|$, write F_{time} for their harmonic mean, and A_{src} for the fraction of gold sentences the submission both overlaps and attributes to the right speaker and source type. Then $\text{EvidenceMatch}(P, G) = F_{\text{time}} \cdot A_{\text{src}}$: right seconds with the wrong speaker score zero, and so does right attribution without overlap. Several minimal sets may be legal, and every claim c that requires evidence must be supported in its own right:

$$G^{+} = \min\Big(\max_{M^{+}} \text{EvidenceMatch}(P, M^{+}),\ \min_{c \in \text{required}} \max_{M_c^{+}} \text{EvidenceMatch}(P_c, M_c^{+})\Big), \tag{3}$$

where M^{+} ranges over legal support sets, P_c and M_c^{+} restrict citations and support sets to claim c, and G^{-} is defined analogously over the required counter-evidence and is 1 when none is required. The primary metric requires every condition to hold at once:

$$\text{ECS@}B = \mathbb{1}\left[\begin{array}{l} \text{decision correct, and answer correct when answerable} \\ \text{every required claim value matches its gold value} \\ G^{+} \geq \eta_{\text{support}} \ \wedge\ G^{-} \geq \eta_{\text{counter}} \\ C_{\text{total}} \leq B_{\text{audio}} \ \wedge\ N_{\text{tool}} \leq B_{\text{tool}} \\ \text{abstention certificate valid when unanswerable} \end{array}\right], \tag{4}$$

with thresholds $\eta_{\text{support}} = 0.8$ and $\eta_{\text{counter}} = 1.0$, fixed before any result was seen and changed only with a new protocol version. Every cited interval must also overlap a gold sentence with its speaker and source type and lie inside an audio view the harness delivered. For an unanswerable question, the decision is correct only with the reason the contract accepts. The abstention certificate can also demand a minimum search coverage C_{unique}/T or named intervals; the v0.1 contracts demand neither, so here it reduces to the reason. ECS is all-or-nothing, so we report its components beside it: answer accuracy, evidence precision and recall, abstention calibration, cost, and the first failing stage. That stage is the first condition a run fails, in order: *budget* (over the limit), *decide* (wrong decision or wrong abstain reason), *reason* (wrong answer or claim value), *locate* (certificate not met), and *submit* (evidence not established).

3.3 Harness and reference agent

The harness, not the agent, decides when cited audio counts as heard.

Environment and delivery. The harness owns the budget, the delivery receipts, and the official trajectory record; the agent never touches the waveform. One rule underlies the central measurement: *a well-formed action is never taken as proof that the model received the audio*. Audio that entered a successful model request is confirmed and charged, even if the resulting action is unusable. Audio attached to a failed request is neither, and a request whose outcome is unknown keeps the run out of scoring. A cited interval counts as heard only if the harness log, never the agent's report, shows a charged, confirmed action delivering it in this run.

Reference agent. The reference agent wraps an open omni-modal model in this loop. A listen names up to five parts, each a 60 s chunk by index or a span of at most 60 s, and delivers at most 300 s; over-limit requests are truncated or rejected with feedback. Old audio is not re-sent, so the agent keeps what it heard in its evidence state. A run without a well-formed submission fails; none is fabricated.

4 The CabinBench Corpus

4.1 Synthesis pipeline

Figure 3 is read left to right; each box below is one stage, and code sets structure and timing while models supply content, so the facts that scoring needs hold by construction.

Scenario brief. Each of 200 frozen recording specifications, stratified over ten factors (Appendix A.5), yields a brief of people, weather, and trip. DeepSeek turns it into a script in a closed loop of outline, expansion, and review.

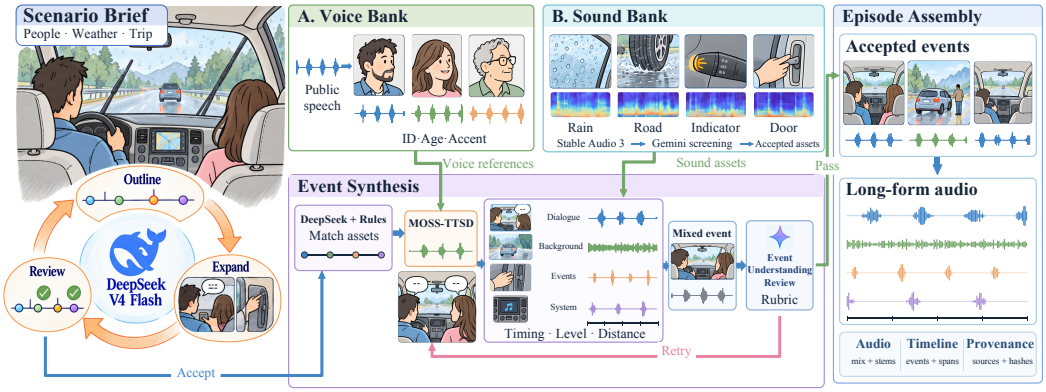


Figure 3: Construction pipeline. DeepSeek turns a scenario brief into a script. Event synthesis renders it event by event from the voice bank (A) and the sound bank (B), and events that pass the event understanding review are assembled into long-form audio.

Voice bank (A) and sound bank (B). The voice bank holds 147 voice references from public speech (Dai et al., 2025), each with an ID, age, and accent. The sound bank holds assets such as rain, road, indicator, and door, generated with Stable Audio 3 (Evans et al., 2026) and kept only after Gemini screening.

Event synthesis. Rules in code fix each script's beats, time windows, and scene slots, and DeepSeek writes only the dialogue inside them. MOSS-TTSD (Zhang et al., 2026) renders each scene's speakers jointly from their voice references, assets are matched to each event, and the mixed event places the dialogue, background, event, and system stems by timing, level, and distance relative to a single microphone position.

Event understanding review. A model that does not see the script listens to every mixed event and scores it against a rubric; the event is accepted or sent back to retry.

Episode assembly. Accepted events and assets are joined by measured duration into long-form audio, stored as audio (mix and stems), a timeline (events and spans), and provenance (sources and hashes).

4.2 Questions and Contracts

Questions are written from the script, so their evidence is known before any model listens. A model that does not hear the audio writes them against the rendered script and its timeline, each with its type, answer, and gold sentences, and an independent model answers each from the script alone (Section 4.4). Forced alignment against the dialogue stem (Gao et al., 2023) locates every gold sentence. Contracts are compiled from these records: support sets with speaker and source, counter-evidence from superseded statements, and answerability from the question type.

4.3 Counterfactual Siblings

A sibling changes only the sentence that decides a question. Because episode assembly keeps each event's stems, the dialogue segment holding that sentence is re-rendered with it removed, replaced, or moved, spliced back into the dialogue stem, and remixed with the untouched stems, so every sample outside the edited window is identical to the original. Each sibling receives its own contract: unanswerable, the new value, or the old answer with relocated evidence.

Table 2: The six task families, each isolating one behavior that long-recording understanding requires. A recording can carry questions from several families.

Family	What it tests
Sparse localization	Decisive evidence is a few seconds, among distractors.
Cross-time tracking	A plan or state evolves; asked at a time, or finally.
Multi-hop integration	The answer combines facts stated minutes apart.
Correction and conflict	A statement is later revised, and timeline order decides.
Acoustic-semantic binding	Non-speech cues or source type decide (person / phone / vehicle).
Answerability	Evidence absent, ambiguous, or conflicted: abstain justifiably.

Table 3: Main results on the `raw_listening` track, all 2,608 questions per cell. **Ans**: answer accuracy on the 2,202 answerable questions outside the two-option type, macro-averaged over five task families (Answerability is scored by Abst F_1; the two-option questions list the gold first in every case, see Section 5). **ECS**: Evidence-Constrained Success at the budget. **Hit**: every gold sentence of one support set is overlapped by a citation with the right speaker and source. **Ovl|C**: among correct answers, the share whose citations overlap a gold sentence. **Abst F$_1$**: abstention F_1 against the 202 unanswerable questions. **No sub.**: runs without a schema-valid submission at $0.25T$, counted as failures elsewhere. **Controls** answer without the recording, so only Ans applies and it does not depend on B.

	$B = 0.25T$					$B = 1.00T$				
Model	Ans	ECS	Hit	Ovl|C	Abst F$_1$	Ans	ECS	Hit	Ovl|C	No sub.
Open-source										
Qwen3-Omni-30B-A3B-Instruct	16.4	0.9	3.5	20.3	16.3	19.6	0.6	2.7	11.5	2.8
Qwen2.5-Omni-7B	0.0	0.1	0.0	–	15.3	0.0	0.0	0.0	–	31.9
Controls										
Question only, no audio	0.1	–	–	–	–	0.1	–	–	–	–
Question and metadata, no audio	0.0	–	–	–	–	0.0	–	–	–	–

4.4 Statistics and validation

The corpus comprises 200 continuous recordings of synthetic Mandarin in-cabin audio, each captured by a single fixed in-cabin microphone and stored as two channels, 138 hours in all, with 140, 52, and 8 recordings in the 30, 60, and 120 minute duration tiers. The question pool holds 2,608 questions of nine types, grouped into the six task families of Table 2, with the mapping in Appendix A.6. Every unanswerable question is `absent`. A pilot of 26 counterfactual siblings for nine questions has been rendered but not yet listened to or evaluated.

Validation. Every script passes the closed-loop check against its specification, and every rendered event the event understanding review. Of 2,617 candidate questions, the independent answerer agreed with the question writer on 2,606, disagreed on 2 that adjudication kept as paraphrases, and marked 9 as having two defensible answers, which were removed. Every contract interval was located by forced alignment without falling back to proportional timing. Two no-audio controls answer 0.1% and 0.0% of the answerable questions (Table 3); this is a shortcut diagnostic, not proof that every individual item requires listening.

5 Experiments

Setup. All runs use the `raw_listening` track: no transcript, no global preprocessing, and the budget of Section 3. Listens follow the limits of Section 3.3. Each question is run once at $B = 0.25T$ with 24 tool calls and once at $B = 1.00T$ with $\lceil T/300\,\mathrm{s} \rceil + 12$ calls, enough to hear the recording in 300 s audio views and re-listen twelve times (18 to 36 calls). Runs are capped at 48 steps. We evaluate Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b) and Qwen2.5-Omni-7B (Xu et al., 2025a), served with vLLM under the same prompts and scorer. Scores use protocol v0.1 unchanged, with sentence-level gold intervals from forced alignment.

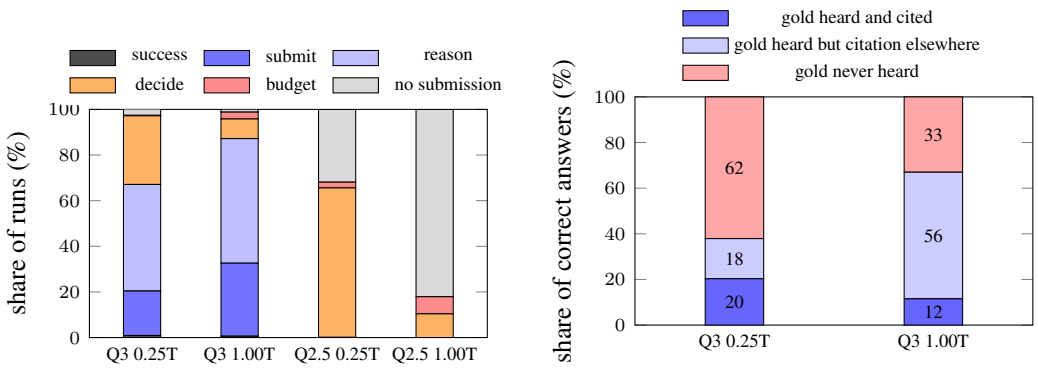


Figure 4: **Where runs fail, and what stands behind the correct answers.** Left: the first failing stage of every run (Section 3); a run counts as *success* only when it passes every gate of ECS. Right: the correct answers of Qwen3-Omni split by whether the delivery log shows a gold sentence was heard and whether a citation overlaps it. Q3 is Qwen3-Omni-30B-A3B-Instruct and Q2.5 is Qwen2.5-Omni-7B.

5.1 Main results

The main results test the paper's central question: whether the model's correct answers come from listening. Table 3 gives the headline numbers, Figure 4 the first failing stage of every run, and Table 7 (Appendix A.6) the breakdown by task family.

Answers against evidence. Correct answers far outnumber established evidence. Qwen3-Omni's answer accuracy is 16.4% at $0.25T$ and 19.6% at $1.00T$, while its ECS is 0.9% and 0.6%, and Qwen2.5-Omni's ECS is at most 0.1%. The protocol's answer comparator accepts loose matches; a strict one (exact, substring, or bare number) gives 12.5% and 14.8%. Two no-audio controls, question only and question with metadata, abstain on 99% of questions and answer 0.1% and 0.0% correctly (Table 3), consistent with little answer leakage in these controls; this does not prove every item requires audio. At $0.25T$, 62% of Qwen3-Omni's correct answers never heard a gold sentence (Section 5.3).

Citations of unheard audio. The harness catches, rather than scores, citations of audio the model never received. The harness rejected 59 submissions at $0.25T$ and 283 at $1.00T$ (2.3% and 10.9% of runs) for citing audio outside every delivered audio view. Of these, 25 and 107 had the right answer. Answer-only scoring would count them correct.

Budget and behavior. The budget changes behavior more than accuracy. An audio budget four times larger buys three points of accuracy, and the 30, 60, and 120 minute duration tiers are within five points of each other (Table 8). Behavior moves much more. With the whole recording available, Qwen3-Omni's abstention rate falls from 26.0% to 5.5%, and the second most common first failing stage shifts from *decide* (30% of runs) to *submit* (32%), while *reason* stays the largest. The agent stops abstaining and starts answering without establishing its evidence (Figure 4, left).

Task families. Task families differ in evidence more than in accuracy. At $1.00T$ four families sit between 18 and 20%. The 204 two-option questions list the gold as the first option in every case, a bias found after the runs (a first-option policy scores 100% on them), so Ans excludes them and Table 7 reports them.

A previous-generation 7B model. A previous-generation 7B model cannot sustain the protocol. Qwen2.5-Omni-7B answers no question correctly at either budget. Of its runs, 32% end without a schema-valid submission (82% at $1.00T$), and every valid submission it makes is an abstention.

5.2 LISTENING BEHAVIOR

Budget use. The agent spends a fraction of its budget. At $0.25T$, Qwen3-Omni makes a median of 6 listens and hears a median of 16% of the recording, under the 25% it may spend.

Where the agent listens. The agent rarely leaves the beginning, so the search policy binds before the budget does. Half of its listens at $0.25T$ jump to a non-adjacent position, yet only 1.1% of runs ever start a listen in the second half of the recording. At $1.00T$, 19% do, and 30% hear anything past the midpoint, almost all by sweeping from the start.

5.3 EVIDENCE BEHIND CORRECT ANSWERS

Citations and gold sentences. Most correct answers cite no gold sentence. Among Qwen3-Omni's correct answers, only 20.3% at $0.25T$ and 11.5% at $1.00T$ cite any interval overlapping a gold sentence.

Whether the evidence was heard. At $0.25T$, most correct answers never heard the gold sentence. The delivery log splits the correct answers three ways (Figure 4, right). At $0.25T$, 62% never heard a gold sentence and cite other audio they did hear, 18% heard it inside a 300 s audio view but cite elsewhere, a median 58 s away, and 20% heard it again in a short listen and cite it. At $1.00T$ they are 33%, 56%, and 12% (without the two-option questions: 57 / 19 / 24 and 35 / 51 / 15), and the misplaced citations cluster in the first five minutes of the recording, a median 357 s from the gold. In 97% of those runs the answer is not said anywhere in the audio the model received, the same rate as for its wrong answers (Appendix A.6); with the strict comparator the share is 59%. A correct answer therefore most often rests on nothing the model heard.

Why correct answers do not come from evidence. At $0.25T$ the agent hears the first five minutes and answers: 55% of its correct answers with the gold unheard are two-option questions or default answers, and when it does hear the gold, 85% of its misplaced citations stay inside the same 300 s audio view. At $1.00T$ it sweeps the recording, hears the gold, and cites without a clock, mostly in the first five minutes. The budget is not the constraint; the search policy and the citation are, and the harness records what a policy rewarded for ECS rather than for the answer would learn from.

5.4 ABSTENTION

Abstention and search. Abstention does not follow search. At $0.25T$, Qwen3-Omni abstains on 36% of the 202 unanswerable questions but also on 25% of answerable ones (precision 10.6%). At $1.00T$ it almost never abstains (recall 9.4%).

Justified abstention. Justified abstention is rarer still. In v0.1 a justified abstention needs the reason `missing_evidence`, since no coverage minimum is set. Only 4.5% and 4.0% of unanswerable questions get that reason from Qwen3-Omni.

6 CONCLUSION

We presented CabinBench, to our knowledge the first benchmark that scores the listening process behind an answer rather than the answer alone. Its corpus is synthesized under full control, so every question carries an exact evidence contract and any sentence can be edited into a counterfactual whose truth is known. Its harness lets an agent buy audio under a budget, cite what it heard, and earn credit only for receipted citations, and it releases with the protocol, the scorer, and every replayable run. Applied to current omni-modal models, it shows that most correct answers were never heard: at a quarter budget 62% of Qwen3-Omni's correct answers never received a gold sentence, and with the full recording a tenth of its submissions cited audio it was never given. Listening is now a measurement, and a target for policies to earn.

Limitations. The audio is synthetic, two open models of one family are evaluated without an oracle upper bound, and no trajectory data or listening policy trained against ECS exists yet; the harness records the receipted citations such a policy would learn from.

AI USE STATEMENT

Generative models are part of the benchmark construction pipeline itself. DeepSeek writes scene summaries and dialogue inside a code-fixed structure and reviews its drafts in the closed authoring loop (Section 4). MOSS-TTSD and Stable Audio 3 render the audio. Gemini screens sound assets, and an understanding model performs the event understanding review. A model that does not hear the audio writes the questions, an independent model answers them from the script for validation, and a forced aligner locates the gold sentences. Each model's role, version, prompts, and sampling parameters are recorded per run. Every rendered event entering the benchmark passes a review that does not see the script. AI assistants were also used in research ideation and execution, drafted and edited sections of this text under the authors' direction, ran the evaluation and analysis code, and cross-checked the literature; the authors verified every claim, number, and citation.

ETHICS STATEMENT

All audio in CabinBench is synthesized from fictional, scripted scenarios. The characters, trips, places, and events are invented. The only real speech involved is a set of short clips from a public speech-research corpus (Dai et al., 2025), used solely as timbre references for the synthetic voices. No other recordings of real people are used in CabinBench's audio, questions, or evaluation. The clips are not redistributed, and release of the synthesized audio is subject to the corpus terms. Attributes attached to voices (age group, accent) are perceptual labels for stratification, not demographic ground truth. CabinBench studies in-cabin scene understanding entirely on synthetic scenes, and we do not apply these systems to recordings of real people.

REPRODUCIBILITY STATEMENT

The quota pool is frozen with its allocation seed (Appendix A.5). Prompt text is stored in files rather than in Python string literals. Each file is loaded once per process and content-hashed, which ties each run to the exact prompts that produced it. Each model run records the deployment identifier and version, the sampling parameters, the maximum output length, and the verbatim system prompt with its SHA-256. It also records the input, output, reasoning, and audio token counts as the service reported them. Cost and delivered intervals are computed from the harness log (Section 3.3). A stored trajectory replays through the scorer to reproduce its evaluation. Failed runs and rejected actions are retained with their records, so a reported number can be traced to the runs behind it.

REFERENCES

Junyi Ao et al. SD-Eval: A benchmark dataset for spoken dialogue understanding beyond words. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

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

Yuhang Dai et al. AISHELL-5: The first open-source in-car multi-channel multi-speaker speech dataset for automatic speech diarization and recognition. In *Proceedings of INTERSPEECH*, pp. 5493–5497, 2025. doi: 10.21437/Interspeech.2025-1886.

Zach Evans, Julian D. Parker, Matthew Rice, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio 3. arXiv preprint arXiv:2605.17991, 2026.

Chaoyou Fu et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, and Shiliang Zhang. FunASR: A fundamental end-to-end speech recognition

toolkit. In *Proceedings of INTERSPEECH*, pp. 1593–1597, 2023. doi: 10.21437/Interspeech.2023-1428.

Tiantian Geng et al. LongVALE: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Sreyan Ghosh et al. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Peize He et al. AudioMarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio llms. arXiv preprint arXiv:2510.07293, 2025.

Chien-yu Huang et al. Dynamic-SUPERB: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. doi: 10.1109/ICASSP48485.2024.10448257.

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

Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, et al. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. In *International Conference on Learning Representations (ICLR)*, 2026.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, pp. 190800–190850, 2025. doi: 10.52202/085713-5729.

Chun-Yi Kuan and Hung-yi Lee. AQUA-Bench: Beyond finding answers to knowing when there are none in audio question answering. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16262–16266, 2026. doi: 10.1109/ICASSP55912.2026.11460647.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, et al. MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 22688–22697, 2026. doi: 10.1609/aaai.v40i27.39430.

Chin-Jou Li, Masao Someki, Woojeong Jin, Yashish M. Siriwardena, Tanmay Laud, Shanil Puri, and Shinji Watanabe. HARP: Agentic hybrid retrieval and analysis for long-form audio. In *IEEE Spoken Language Technology Workshop (SLT)*, 2026a. URL https://arxiv.org/abs/2609.14116. Accepted for SLT 2026; proceedings forthcoming.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, et al. OmniGAIA: Towards native omni-modal ai agents. arXiv preprint arXiv:2602.22897, 2026b. URL https://arxiv.org/abs/2602.22897.

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

Ziyang Ma et al. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, pp. 74837–74868, 2025. doi: 10.52202/085713-2252.

Hokuto Munakata, Taichi Nishimura, Shota Nakada, and Tatsuya Komatsu. Language-based audio moment retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. doi: 10.1109/ICASSP49660.2025.10890873.

Hokuto Munakata, Takehiro Imamura, Taichi Nishimura, and Tatsuya Komatsu. CASTELLA: Long audio dataset with captions and temporal boundaries. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. doi: 10.1109/ICASSP55912.2026.11464074.

Archiki Prasad et al. MeetingQA: Extractive question-answering on meeting transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. doi: 10.18653/v1/2023.acl-long.837.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. doi: 10.18653/v1/P18-2124.

S. Sakshi et al. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations (ICLR)*, 2025.

Natarajan Balaji Shankar, Alexander Johnson, Christina Chance, Hariram Veeramani, and Abeer Alwan. CORAAL QA: A dataset and framework for open domain spontaneous speech question answering from long audio files. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13371–13375, 2024. doi: 10.1109/ICASSP48485.2024.10447109.

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.

Suwon Shon et al. SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. doi: 10.18653/v1/2023.acl-long.496.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. In *International Conference on Learning Representations (ICLR)*, 2025.

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

He Wang, Pengcheng Guo, Yue Li, Ao Zhang, Jiayao Sun, Lei Xie, Wei Chen, Pan Zhou, Hui Bu, Xin Xu, Binbin Zhang, Zhuo Chen, Jian Wu, Longbiao Wang, Eng Siong Chng, and Sun Li. ICMC-ASR: The ICASSP 2024 in-car multi-channel automatic speech recognition challenge. In *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 63–64, 2024. doi: 10.1109/ICASSPW62465.2024.10627712.

Jin Xu, Zhifang Guo, Jinzheng He, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

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

Chih-Kai Yang, Neo Ho, Yen-Ting Piao, and Hung-yi Lee. SAKURA: On the multi-hop reasoning of large audio-language models based on speech and audio information. In *Proceedings of INTERSPEECH*, 2025. doi: 10.21437/Interspeech.2025-839.

Fei Yang, Xuanfan Ni, Renyi Yang, Jiahui Geng, Qing Li, Chenyang Lyu, Yichao Du, Longyue Wang, Weihua Luo, and Kaifu Zhang. LongSpeech: A scalable benchmark for transcription, translation and understanding in long speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15547–15551, 2026. doi: 10.1109/ICASSP55912.2026.11461821.

Qian Yang et al. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. doi: 10.18653/v1/2024.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.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. Re-thinking temporal search for long-form video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8579–8591, 2025. doi: 10.1109/CVPR52734.2025.00802.

Yuqian Zhang, Donghua Yu, Zhengyuan Lin, Botian Jiang, Mingshu Chen, Yaozhou Jiang, Yiwei Zhao, Yiyang Zhang, Yucheng Yuan, Hanfu Chen, Kexin Huang, Jun Zhan, Cheng Chang, Zhaoye Fei, Shimin Li, Xiaogui Yang, Qinyuan Cheng, and Xipeng Qiu. MOSS-TTSD: Text to spoken dialogue generation. arXiv preprint arXiv:2603.19739, 2026.

Ming Zhong et al. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. doi: 10.18653/v1/2021.naacl-main.472.

A Appendix

A.1 Full benchmark comparison

Table 4 is the full benchmark comparison summarized in Section 1.

Table 4: Audio and long-context understanding benchmarks. **Dur**: audio length per item, the longest reported or else the mean. **Tx**: transcript provided. **Seg**: pre-segmented before the model sees it. **Bud**: listening or retrieval budget as a controlled variable. **Int**: an evidence interval is required. **Ver**: the cited interval is verified to have been accessed. **Abs**: abstention is evaluated. **Src**: underlying audio is recorded, TTS-generated, or mixed; concatenated recordings count as recorded here. A dash means a feature is not required by the cited default evaluation, not that a method is incapable of it. †: arXiv preprint, not peer reviewed. Parentheses mark a budget not elected by the agent, either imposed by the evaluator or set as a method hyperparameter. HARP is an agentic retrieval method; * denotes accepted to SLT 2026 with proceedings forthcoming.

Benchmark	Venue	Dur	Tx	Seg	Bud	Int	Ver	Abs	Src
Short-clip audio understanding									
SD-Eval (Ao et al., 2024)	NeurIPS'24 D&B	4.3 s	–	✓	–	–	–	–	real
MMAU (Sakshi et al., 2025)	ICLR'25	10 s	–	✓	–	–	–	–	real
MMAR (Ma et al., 2025)	NeurIPS'25 D&B	30 s	–	✓	–	–	–	–	real
AIR-Bench (Yang et al., 2024)	ACL'24	clip	–	✓	–	–	–	–	real
SAKURA (Yang et al., 2025)	INTERSPEECH'25	clip	–	✓	–	–	–	–	real
AQUA-Bench (Kuan & Lee, 2026)	ICASSP'26	clip	–	✓	–	–	–	✓	real
AF2 (Ghosh et al., 2025)	ICML'25	5 min	–	✓	–	–	–	–	real
MMAU-Pro (Kumar et al., 2026)	AAAI'26	10 min	–	✓	–	–	–	–	real
Long-form audio									
SLUE-TED (Shon et al., 2023)	ACL'23	11.8 min	–	–	–	–	–	–	real
SLUE-NEL (Shon et al., 2023)	ACL'23	clip	–	✓	–	✓	–	–	real
CORAAL QA (Shankar et al., 2024)	ICASSP'24	60 min	✓	✓	–	–	–	–	real
LongSpeech (Yang et al., 2026)	ICASSP'26	10 min	–	–	–	–	–	–	mixed
AudioMarathon† (He et al., 2025)	arXiv'25	5 min	–	–	(✓)	–	–	–	real
ChronosAudio† (Luo et al., 2026)	arXiv'26	20 min	–	–	–	–	–	–	synth
TAG-Bench† (Dai et al., 2026)	arXiv'26	20 min	–	–	–	✓	–	–	real
LAT-Bench† (Shao et al., 2026)	arXiv'26	30 min	–	–	–	✓	–	–	real
VoiceGiraffe† (Ye et al., 2026)	arXiv'26	2 h	–	–	–	–	–	–	real
HARP (Li et al., 2026a)	SLT'26*	75 min	–	–	(K=3)	✓	–	–	mixed
Long video with an audio track									
Video-MME (Fu et al., 2025)	CVPR'25	60 min	opt.	–	–	–	–	–	real
LongVALE (Geng et al., 2025)	CVPR'25	video	–	✓	–	✓	–	–	real
In-car audio									
AISHELL-5 (Dai et al., 2025)	INTERSPEECH'25	ASR	–	✓	–	–	–	–	real
ICMC-ASR (Wang et al., 2024)	ICASSPW'24	ASR	–	✓	–	–	–	–	real
CabinBench	This work	2 h	–	–	✓	✓	✓	✓	synth

A.2 CabinBench at a glance

Table 5: CabinBench at a glance (design targets of the frozen quota pool). Durations are tier targets. Realized durations are assembled from measured clip lengths.

Recordings	200, stratified over ten factors
Duration tiers	30 min / 60 min / 120 min, two-channel 16 kHz
Language	Mandarin Chinese
Speakers per recording	2 to 4 cabin occupants + phone/system sources
Voice bank	147 reference voices from a public speech corpus, fixed across recordings
Answerability split	answerable / absent / ambiguous / conflicted (realized: answerable / absent)
Budgets	0.10 / 0.25 / 1.00 of recording length
Ground truth	hidden evidence contract per question

A.3 THE THREE TRACKS

`raw_listening` gives technical metadata and raw audio access. It forbids global semantic preprocessing: no free transcription, captioning, or embedding of the whole recording. Its preprocessing cost C_{pre} must be zero. `shared_index` adds one frozen, deliberately noisy index shared by all agents, so differences between agents reflect policy rather than retrieval engineering. `open_system` permits any preprocessing, provided its full cost is reported and open to audit as two separate quantities: audio seconds read and compute seconds spent.

A.4 THE ACTION SET

Table 6: The full action set. Audio seconds are charged on what a confirmed request delivered, not on what the agent requested. Repeated listening to the same interval is charged again.

Action	Effect	Semantic obs.	Charged
`get_episode_info`	technical metadata	no	no
`seek`	move a cursor, return nothing	no	no
`listen`	raw audio for an interval	yes	yes
`inspect`	run a tool on an interval	yes	yes
`search_index`	query a permitted index	yes	indirect
`compare`	directed comparison of intervals	yes	yes
`record`	update the agent's own ledger	no	no
`submit`	submit and stop	no	no

A.5 SYNTHESIS DETAILS

Frozen quota pool. The pool of 200 recording specifications is stratified over ten factors: task family, duration tier, dialogue density, speaker count and voice attributes, evidence sparsity, distractor load, correction structure, source-type mix, acoustic scene, and answerability. Its allocation seed regenerates the same 200 specifications. Because the pool is frozen before any content exists, it cannot be tuned to whichever model we evaluate.

Three-layer script authoring. Each specification becomes a trip script in three layers, splitting labor between code and model. Code fixes the structure: beat sequence, time windows, vehicle states, scene slots, and character targets come from the specification and a seed. DeepSeek writes only what the structure cannot supply, namely scene summaries and dialogue. It works in a closed loop of outline, scene-by-scene expansion, and self-review against the specification, until the script is accepted or abandoned. Keeping the model out of the structure makes the stratification enforceable: a target such as "one correction, nineteen minutes after the plan it revises" becomes a guarantee that the renderer must honor.

Voice and sound banks. Speakers come from a bank of 147 reference voices drawn from a public speech corpus and labeled with stable identifiers and perceptual attributes. Each synthetic voice keeps a fixed timbre across recordings. Non-speech audio comes from a sound bank generated with Stable Audio 3 (Evans et al., 2026) and screened by Gemini for fidelity to its description. Both banks are fixed before rendering, so recordings cite their assets rather than generating new ones.

Event synthesis. A script is rendered event by event. Each dialogue scene is synthesized jointly: MOSS-TTSD (Zhang et al., 2026) renders all of its speakers in a single pass, preserving turn-taking and speaker consistency. Each event is rendered as stems (dialogue, background, event sounds, and system sources such as the navigation voice). The stems carry explicit timing, level, and distance and are mixed to a two-channel cabin mix. Stems let the evidence contract name both *when* something was said and *what kind of source* said it.

Event understanding review. Every event is reviewed *as audio*. Without seeing the script, an understanding model transcribes the rendered clip and answers rubric questions about who spoke,

what was decided, and whether the specified content survived synthesis. Events that fail are re-rendered or dropped. This event understanding review catches rendering failures that script-level checking cannot: unintelligible dialogue, the wrong speaker, or a decisive phrase buried under the background.

Assembly and provenance. Accepted events are assembled into a continuous recording at one of three duration tiers. Assembly uses the *measured* duration of each rendered clip, not the script's nominal timing. Each recording is stored with its stems, the event timeline with exact spans, and a provenance record listing every source asset and its content hash. Contracts are compiled against this timeline, so the ground truth names seconds of the final waveform rather than script positions.

On the task families. The families classify *failure opportunities* for an evidence-seeking agent rather than question topics. Sparse localization is a control condition rather than a contribution, since needle-style retrieval over long audio has been studied (Luo et al., 2026; Dai et al., 2026; Ye et al., 2026). Here it calibrates how much harder the other families are. Correction and conflict has no direct precedent that we know of: no prior benchmark evaluates revisions separated by minutes inside one recording, where the audio timeline itself determines which statement is current.

Terminal states. A run ends in one of five states: `submitted` (a well-formed submission), `step_limit`, `model_error`, `invalid_run`, or `setup_failure`. Failures produce auditable records with stable error codes. As in the main text, a run without a submission fails and no submission is fabricated for it.

A.6 Evaluation details and full results

Question types and families. Each contract names its task family, and the nine question types map as follows: Sparse localization, `single_detail` (776) and `multi_confirm` (214); Cross-time tracking, `temporal_order` (204, two-option) and `source_time_pinned` (201); Multi-hop integration, `cross_segment` (404); Correction and conflict, `correction_final` (160); Acoustic-semantic binding, `device_source` (244) and `attribution` (203); Answerability, `absent` (202). In every `temporal_order` question the gold is the first of the two options, and in 31 of the 36 enumerated `correction_final` questions it is the second; a first-option policy therefore scores 100% on the former. The bias was found after the runs, so Ans in Table 3 excludes the two-option questions and the Answerability family, and no threshold was changed.

Harness. The agent acts in JSON, and a listen obeys the part and length limits of the main text. Parts beyond the fifth are dropped and the agent is told so on its next turn. A call over the seconds limit is recorded as rejected, charged one tool call, and returned with the reason. Every observation lists which chunks have been heard whole and which have not. The harness fills `stop_reason` from the decision, and shows enumerated answers as their options rather than as a schema. These limits, the treatment of over-limit listens, and the call count of the full budget were set after a twelve-question pilot. No scoring threshold was changed. Runs were served by vLLM on one node with six single-GPU replicas at up to 12 concurrent episodes, and a run took 22 s at $0.25T$ and 36 s at $1.00T$ on average. Failed runs are kept with their failure record and never rerun silently.

Results by family and tier. Table 7 breaks down Qwen3-Omni by task family, and Table 8 every model and budget by duration tier.

Listening statistics. From the traces of Qwen3-Omni, over runs with a valid submission (2,536 and 2,574): at $0.25T$, median 6 listens, median 16.4% (mean 15.1%) of the recording heard, 50% of listens non-adjacent to the previous one, and 1.1% of runs starting a listen in the second half. At $1.00T$, median 7 listens, 36.4% (mean 43.1%) heard, 37% non-adjacent, and 19% starting a listen in the second half (30% hear any audio past the midpoint).

Citation placement. Table 9 joins the delivery log of every Qwen3-Omni run with a correct answer to its contract. A gold sentence counts as heard if any delivered clip covers it. The offset is from the citation midpoint to the nearest gold midpoint. No correct answer cites undelivered audio, because the harness rejects such submissions before scoring.

Table 7: Qwen3-Omni-30B-A3B-Instruct by task family. **Ans**: answer accuracy; **Hit**: evidence hit; **Abst**: abstention rate. For Answerability, a correct answer is a justified abstention.

Family	n	$B = 0.25T$			$B = 1.00T$		
		Ans	Hit	Abst	Ans	Hit	Abst
Sparse localization	990	16.5	5.1	25.4	20.1	3.6	6.1
Cross-time tracking	405	34.8	6.4	21.7	41.5	4.2	4.0
Multi-hop integration	404	13.6	0.0	30.4	17.6	0.5	4.0
Correction and conflict	160	12.5	0.6	27.5	20.0	1.9	6.9
Acoustic-semantic binding	447	15.4	3.4	22.6	19.0	2.9	4.9
Answerability	202	4.5	0.0	35.6	4.0	0.0	9.4

Table 8: Results by duration tier. Columns as in Table 3, but Ans is the plain rate over all questions of the tier, including the two-option questions and Answerability, and Abst is the abstention rate.

Model	B	Tier	n	Ans	ECS@B	Hit	Abst	No sub.
Qwen3-Omni-30B-A3B-Instruct	$0.25T$	30min	1673	17.9	1.0	4.1	24.4	3.8
	$0.25T$	60min	777	16.7	0.6	2.7	29.3	0.9
	$0.25T$	120min	158	17.7	0.6	1.9	27.2	0.6
	$1.00T$	30min	1673	21.8	0.7	3.1	5.5	1.6
	$1.00T$	60min	777	20.3	0.5	2.2	6.2	0.9
	$1.00T$	120min	158	25.3	0.6	1.3	2.5	0.6
Qwen2.5-Omni-7B	$0.25T$	30min	1673	0.1	0.1	0.0	64.8	35.2
	$0.25T$	60min	777	0.1	0.1	0.0	79.7	20.3
	$0.25T$	120min	158	0.0	0.0	0.0	45.6	54.4
	$1.00T$	30min	1673	0.0	0.0	0.0	16.4	83.6
	$1.00T$	60min	777	0.0	0.0	0.0	21.0	79.0
	$1.00T$	120min	158	0.0	0.0	0.0	19.6	80.4

Table 9: Where the citations of correct answers land, Qwen3-Omni.

	$B = 0.25T$		$B = 1.00T$	
Correct answers	n	%	n	%
All	448	100	555	100
Gold heard, citation overlaps it	91	20.3	64	11.5
Gold heard, citation misses it	79	17.6	308	55.5
Gold never heard, cites other heard audio	278	62.1	183	33.0
Median offset when gold heard but missed	58 s		357 s	
Misses inside the first 300 s	–		79.8%	

Answer audibility. For the correct answers of Table 9, we check whether the normalized answer or an alias of at least three characters occurs in the transcript of any chunk the model received (a chunk counts when at least half of it was delivered). Among runs that never heard a gold sentence, the answer is audible in 9 of 278 at $0.25T$ and 16 of 183 at $1.00T$; among wrong answers with the gold unheard, in 3.2% and 5.2%. The protocol comparator accepted 89 of the 448 correct answers at $0.25T$ only through its loose branches (shared numbers, or 75% of characters in order); with exact, substring, or bare-number matching, answer accuracy is 12.5% and 14.8% (macro over the five families of Table 3) and the share of correct answers with the gold unheard is 59% and 32%. Without the two-option questions the split of Table 9 is 24 / 19 / 57 at $0.25T$ and 15 / 51 / 35 at $1.00T$.

A.7 Deferred material

The supplement accompanying the corpus release will contain: the full evidence contract schema, with the episode of Figure 1 as a worked example; the response schema the agent must fill, including claim slots and abstain reasons; the complete prompt set for the authoring loop, the screening step,

and the event understanding review; one full script with its event timeline and provenance record; the stratification factors with the frozen allocation table; lessons from construction, including where closed-loop self-review failed; and the datasheet for the corpus.