ListenCare: Encounter-Grounded Audio Question Answering for Long-Form Clinical Conversation Speech

Seongsu Bae [*1] **Chaeeun Shim** [*1] **Sungbae Park** [2] **Edward Choi** [1]

Abstract

Clinical conversation speech is a rich source of evidence for ambient clinical documentation, patient-facing summaries, and care coordination. Yet existing audio-language benchmarks rarely test whether models can answer clinically relevant questions grounded in long-form doctor–patient encounter audio. We introduce LISTENCARE, an encounter-grounded audio question answering benchmark with 4,085 four-option MCQA instances from 457 synthetic and human-recorded mock clinical encounters. The benchmark is organized around a six-capability taxonomy spanning evidence recovery, evidence status/sufficiency, attribution grounding, temporal/discourse tracking, interaction-level reasoning, and audio-specific grounding. We evaluate five open-weight large audio-language models (LALMs) in three settings: end-to-end QA over full encounter audio, an ASR-transcript cascade, and QA over an oracle evidence-localized audio crop. We find that weaker LALMs still exhibit audio–text modality gaps and difficulty retrieving evidence from long audio, whereas the strongest models substantially reduce these gaps but remain limited on interaction-level reasoning and audio-specific grounding. The project repository is available at https://github.com/baeseongsu/listencare.

[*]Equal contribution First-author contacts: Seongsu Bae <seongsu@kaist.ac.kr>, Chaeeun Shim <chaeeun@kaist.ac.kr>. [1]Kim Jaechul Graduate School of AI, KAIST, Republic of Korea [2]Seoul National University Boramae Medical Center, Republic of Korea. Correspondence to: Edward Choi <edwardchoi@kaist.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Doctor–patient encounters are increasingly treated as source material for speech-driven clinical applications such as ambient clinical documentation (Krishna et al., 2021; Yim et al., 2023; Tierney et al., 2024), patient-facing communication (Zaretsky et al., 2024; Steimetz et al., 2024), and care-transition communication (Hartman et al., 2024; Williams et al., 2025). This source material is inherently the raw speech of a long-form, multi-speaker conversation that unfolds turn by turn. Clinically relevant evidence is scattered across the interaction: who says what; whether claims are affirmed, revised, or left unresolved; how care plans are proposed or accepted; and which audible cues (*e.g.*, a sigh preceding a reply) carry meaning beyond the words themselves. A model can reliably support speech-driven clinical applications only if it can recover and interpret this scattered evidence directly from the audio.

Despite this need, how well models understand and reason over long-form clinical conversation speech remains largely untested by existing benchmarks. General audio-language benchmarks (Yang et al., 2024; Wang et al., 2024; Sakshi et al., 2024; Kumar et al., 2025; He et al., 2025; Ahia et al., 2025) evaluate audio instruction following, audio question answering, and long-context audio understanding across speech, sounds, and music, but they rarely target long-form, multi-speaker conversations in which evidence is distributed across turns. Clinical dialogue resources and ambient clinical intelligence benchmarks provide medical dialogues, mock or synthetic consultations, summaries, and note-generation targets (Zeng et al., 2020; Papadopoulos Korfiatis et al., 2022; Yim et al., 2023; Ben Abacha et al., 2023; Labrak et al., 2026), rather than evaluating what models understand from the encounter audio itself. As a result, it remains unclear whether large audio-language models (LALMs) can answer clinically relevant questions grounded in the audio of a full doctor–patient encounter.

LISTENCARE turns this gap into an encounter-grounded audio QA task. Instead of treating clinical conversation understanding as a single score, the benchmark targets six concrete capabilities: can a model recover stated evidence, determine whether it is affirmed or absent, attribute it to the right speaker or source, track state across turns, inte-

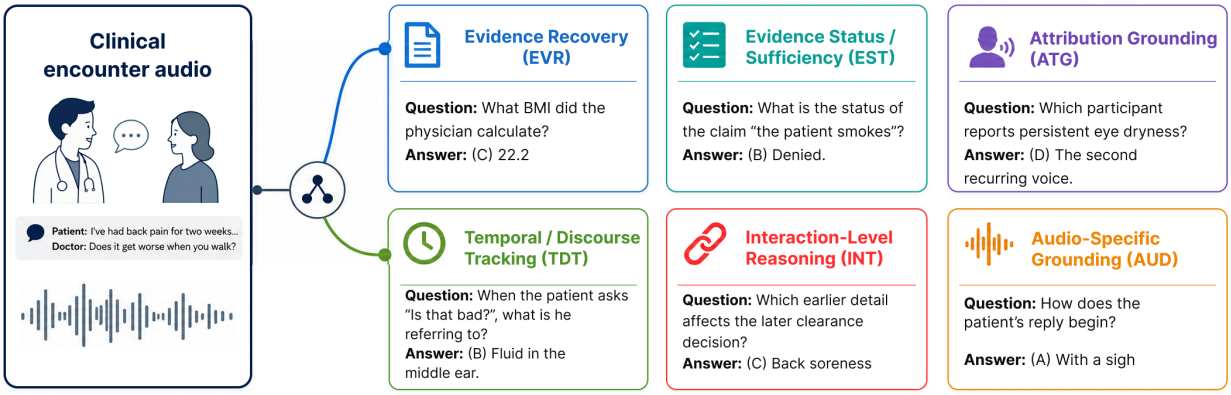

Figure 1. Overview of the LISTENCARE benchmark structure. Full clinical encounter audio is paired with MCQA questions covering six capabilities: evidence recovery (EVR), evidence status/sufficiency (EST), attribution grounding (ATG), temporal/discourse tracking (TDT), interaction-level reasoning (INT), and audio-specific grounding (AUD).

grate interaction-level rationales, and use audible cues beyond words? This makes question answering a diagnostic layer between speech perception and downstream speech-driven clinical applications. In total, LISTENCARE contains 4,085 four-option multiple-choice question-answering (MCQA) instances from 400 Synth-DoPaCo and 57 PriMock57 encounters. We use these questions not only to measure full-encounter audio QA performance, but also to diagnose whether errors arise from audio evidence access or from retrieving relevant evidence across a long encounter.

2. Related Work

Large audio-language models (LALMs) connect pretrained language models with audio encoders, enabling natural-language instruction following over speech and other audio inputs (Tang et al., 2024; Chu et al., 2023). This has led to broad audio-language benchmarks that test LALMs on understanding and reasoning over speech, non-speech sounds, and music (Yang et al., 2024; Wang et al., 2024; Sakshi et al., 2024; Kumar et al., 2025). Recent long-form audio benchmarks scale this evaluation to multi-minute or hour-scale inputs, emphasizing temporal localization and grounding, long-speech transcription and summarization, and speaker/event counting (Ahia et al., 2025; Yang et al., 2026; Shao et al., 2026). Yet these benchmarks leave open whether LALMs can answer questions grounded in evidence distributed throughout long, multi-speaker conversations.

Doctor–patient encounters are long, multi-speaker speech interactions in which clinically relevant evidence is distributed across turns. Existing clinical-conversation resources and ambient clinical intelligence benchmarks have made these encounters a practical testbed for clinical AI (Zeng et al., 2020; Papadopoulos Korfiatis et al., 2022; Yim et al., 2023; Ben Abacha et al., 2023; Labrak et al., 2026). For LALMs, evaluating doctor–patient speech raises a different question: whether they can understand and reason over encounter evidence distributed across a long, multi-speaker clinical interaction. Concurrent with our work, MedMosaic (Rajgarhia et al., 2026) assembles a large-scale medical audio QA benchmark spanning physiological sounds, synthetic clinical speech, and short and long clinical conversations. In contrast, LISTENCARE specifically targets long-form doctor–patient encounter QA, providing both larger coverage in this setting (4,085 MCQA instances over 457 full encounters, compared with MedMosaic's 1,074 long-form QA pairs) and finer-grained diagnostic evaluation through six clinical-conversation capabilities.

3. Benchmark Construction

3.1. Source Materials

Given the privacy constraints on real-patient clinical encounter audio, LISTENCARE is designed around two publicly available clinical conversation audio sources (both CC-BY-4.0) that are complementary in scale and speech realism. Synth-DoPaCo (Labrak et al., 2026) provides 8,800 synthetic doctor–patient dialogues totaling 1,329 hours, with audio, transcripts annotated with non-speech audio events (*e.g.*, sighs, keyboard typing, or ECG beeps), and reference SOAP notes. PriMock57 (Papadopoulos Korfiatis et al., 2022) provides 57 two-party mock primary-care consultations recorded with 7 practicing clinicians and 57 human participants acting as patients, with audio, manual utterance-level transcripts, and consultation notes. We use the 400-conversation Synth-DoPaCo public development split and all 57 PriMock57 consultations, yielding 457 source encounters in total.

3.2. Task Formulation

We formulate LISTENCARE as encounter-grounded audio question answering: each benchmark instance consists of full clinical encounter audio and a four-option MCQA question. Given the encounter audio x, a text question q, and four text answer choices O, the model must select the single supported option $a \in O$. Each item has exactly one correct option, and is designed to be answerable from evidence in the encounter audio rather than from clinical priors or lexical shortcuts alone. The correct option may therefore be a specific answer, or the judgment that the relevant information is insufficient, not discussed, or unclear.

3.3. Capability Taxonomy

As shown in Figure 1, we organize the benchmark around six capabilities for evaluating long clinical encounter audio understanding: recovering stated evidence (EVR), judging whether evidence is affirmed, negated, uncertain, or insufficient (EST), attributing evidence to the correct speaker or source (ATG), tracking temporal and discourse state across turns (TDT), integrating distributed turns into interaction-level goals or rationales (INT), and grounding answers in audible cues beyond lexical content (AUD). We operationalize these capabilities into 12 QA categories, where each capability is realized as one or more categories that differ in the specific question and answer form they probe. For the category-level definitions and representative MCQA templates, please refer to Appendix A.

3.4. QA Construction Pipeline

Inputs and Setup. We construct each benchmark instance from the source-provided reference transcript of the encounter audio, segmented into speaker-labeled turns. Beyond the spoken words, we additionally use non-speech sound event markers and forced-alignment timing (*i.e.*, the start and end time of each turn), which support attribution grounding (ATG) and audio-specific grounding (AUD). From these inputs, we run a category-specific construction procedure to produce a four-option MCQA instance (q, O, a) paired with the original encounter audio x and evidence e (*i.e.*, the supporting turn(s) in the encounter). We use Qwen3.6-35B-A3B as the construction LLM for LLM-assisted generation, revision, and validation steps.

Candidate Generation. While all categories share the same downstream quality control, candidate generation itself takes one of three forms. Evidence-recovery, temporal, interaction, and provenance-attribution categories generate several complete MCQA candidates per encounter. Evidence-status categories instead generate claims whose correct label is drawn from a predetermined set of statuses. The remaining attribution categories and the audio category construct their options deterministically from source annotations (*e.g.*, speaker labels or sound-event markers), using the LLM at most for question wording. In every case, the three distractors are built to be encounter-plausible yet non-equivalent to the answer: LLM-assisted categories generate them jointly with the question, whereas deterministic categories assemble them from competing source annotations (*e.g.*, other speakers or sound events).

	All	Synth	PriMock
Source encounters			
Encounters	457	400	57
Mean duration (min)	8.9	8.8	9.1
Total duration (h)	67.6	58.9	8.6
Mean speaker turns/encounter	36.8	28.7	93.8
Mean speaker-turn duration (s)	14.5	18.5	5.8
Benchmark MCQA instances			
MCQA instances	4,085	3,624	461
Evidence recovery (EVR)	366 (9.0%)	316 (8.7%)	50 (10.8%)
Evidence status/sufficiency (EST)	787 (19.3%)	692 (19.1%)	95 (20.6%)
Attribution grounding (ATG)	1,631 (39.9%)	1,441 (39.8%)	190 (41.2%)
Temporal/discourse tracking (TDT)	552 (13.5%)	481 (13.3%)	71 (15.4%)
Interaction-level reasoning (INT)	529 (12.9%)	474 (13.1%)	55 (11.9%)
Audio-specific grounding (AUD)‡	220 (5.4%)	220 (6.1%)	–

Table 1. LISTENCARE source and MCQA-instance statistics. **Synth**=Synth-DoPaCo; **PriMock**=PriMock57. Capability percentages are within-column shares of MCQA instances. ‡AUD requires annotated non-speech audio events, available only in Synth-DoPaCo.

Quality Control. Quality-control gates are applied to every candidate. Rule-based checks first reject malformed instances, invalid evidence references, and construction artifacts such as exposed turn numbers or transcript-referencing wording. LLM-based checks then handle semantic judgments: a shortcut audit tests whether the correct option is recoverable from the question and option texts alone (*i.e.*, without the audio), and a deduplication judge rejects near-duplicate questions within the same encounter. For categories that generate multiple candidates, we retain the one least solvable by this shortcut audit; candidates that fail a check may instead be revised and re-evaluated before a final accept-or-drop decision. In the released benchmark, the correct option is approximately uniformly distributed over the four answer positions (23.9% to 26.4%).

3.5. Benchmark Statistics

LISTENCARE contains 4,085 MCQA instances grounded in 457 source encounters: 400 Synth-DoPaCo encounters (mean duration 8.8 minutes) and 57 PriMock57 encounters (mean duration 9.1 minutes). Encounter durations range from 2.8 to 24.6 minutes (29.1% exceed 10 minutes), and each encounter contributes 5 to 12 MCQA instances (8.9 on average). These instances span all six capabilities, although AUD is exclusive to Synth-DoPaCo because it requires non-speech audio-event annotations not present in PriMock57. Table 1 reports the full source and capability breakdown.

4. Experimental Setup

4.1. Evaluation Protocol

We primarily evaluate *End-to-end* audio QA: the LALM receives the full encounter audio, the text question, and the answer options, and predicts an answer from the audio signal itself, without first converting the speech into an intermediate transcript. To probe the modality gap between processing the encounter as audio and as text (*i.e.*, a transcript produced by automatic speech recognition, or ASR), we also evaluate *Cascade* (ASR-to-LALM): an ASR system first transcribes the same audio, and the same LALM answers from the transcript, question, and answer options. For Cascade, we use Whisper large-v3-turbo (Radford et al., 2023) as the ASR component. Note that this ASR setup achieves 4.3% WER on Synth-DoPaCo and 15.6% WER on PriMock57 against the corresponding reference transcripts. We additionally evaluate evidence-localization effects on instances with timestamped evidence. *Local-30s* provides an oracle evidence-localized audio crop centered on aligned evidence turns, while the matched *Full* condition provides the full encounter audio for the same instances. This contrast tests whether questions answerable in localized audio fail in the Full condition. Across all conditions, we report single-answer accuracy: a prediction is correct only when it selects the supported option, and unparseable or missing predictions are counted as incorrect.

4.2. Models

We evaluate five open-weight LALMs spanning 7B to 30B parameters, selected for native support for audio inputs at roughly the 10-minute scale or beyond: Qwen3-Omni-30B-A3B-Instruct/Thinking (Xu et al., 2025), Kimi-Audio-7B (KimiTeam et al., 2025), Audio-Flamingo-3 (Ghosh et al., 2025), and Audio-Flamingo-Next (Ghosh et al., 2026). Their documented or estimated audio budgets cover the average ListenCare encounter length of 8.9 minutes, ranging from 10 minutes for Audio-Flamingo-3 and approximately 10–11 minutes for Kimi-Audio to 30 minutes for Audio-Flamingo-Next and up to 40 minutes per instance for the Qwen3-Omni variants. For each model's checkpoint, runtime, decoding, and audio-budget details, please refer to Appendix B.1.

5. Results and Analysis

5.1. Research Questions

We ask three main research questions on the ListenCare benchmark. First, we ask how well current LALMs answer encounter-grounded questions from full encounter audio, and how this compares with Cascade on the same questions (RQ1). Second, we examine where the remaining errors concentrate across the capability taxonomy introduced in Section 3.3 (RQ2). Third, we compare full encounter audio with oracle Local-30s evidence-localized crops to test whether errors reflect a failure to understand the relevant local evidence or to retrieve that evidence from long, multi-speaker audio (RQ3).

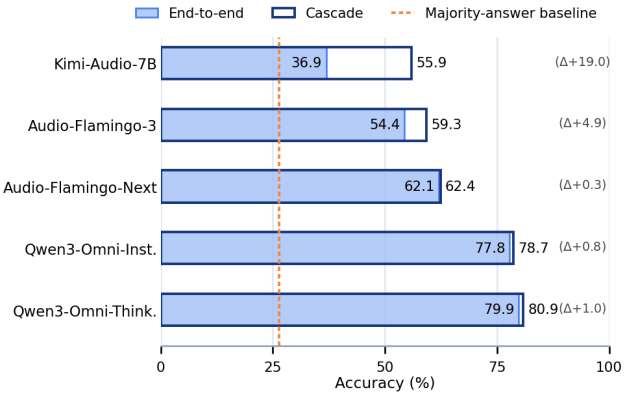

Figure 2. End-to-end and Cascade accuracy for the five primary LALMs on ListenCare. Filled bars report the End-to-end condition; outline bars report the matched Cascade condition using Whisper large-v3-turbo as ASR. The dashed orange line marks the majority-answer baseline (26.4%). Parenthetical labels show the accuracy difference between Cascade and End-to-end in percentage points (pp); positive values indicate higher Cascade accuracy.

5.2. RQ1: Encounter-Level Audio QA Performance

Figure 2 compares the End-to-end condition with the matched Cascade condition on the same encounter-grounded questions. End-to-end accuracy varies widely across the five LALMs, from 36.9% for Kimi-Audio-7B to 79.9% for Qwen3-Omni-Thinking; the top result is far above the majority-answer baseline (26.4%) but still leaves a 20.1% error rate. Across these five models, the Cascade condition raises average accuracy from 62.2% to 67.4% (+5.2 pp). The largest gain appears for Kimi-Audio, while Audio-Flamingo-Next and the two Qwen3-Omni variants nearly close the End-to-end vs. Cascade gap. The small remaining gaps for Audio-Flamingo-Next and the two Qwen3-Omni variants suggest that End-to-end QA over full clinical encounter audio is becoming feasible for the strongest current LALMs. However, the remaining Cascade gains for Kimi-Audio and Audio-Flamingo-3 indicate that End-to-end reliability remains model-dependent, consistent with broader speech-LLM modality gaps between audio and text inputs (Xiang et al., 2025).

5.3. RQ2: Capability-Level Error Sources

Table 2 shows that LALM errors are not evenly distributed across capabilities, but concentrate in interaction-level reasoning and audio-specific grounding. Average End-to-end accuracy is highest for evidence recovery (EVR; 81.7%) but

Model	EVR	EST	ATG	TDT	INT	AUD
Kimi-Audio-7B	28.7	28.0	44.5	41.5	32.4	26.4
Audio-Flamingo-3	90.4	47.9	52.2	70.7	42.9	20.5
Audio-Flamingo-Next	92.6	56.3	67.7	76.0	49.0	22.7
Qwen3-Omni-Inst.	**98.4**	75.0	81.9	84.6	55.6	**60.5**
Qwen3-Omni-Think.	**98.4**	**80.7**	**86.9**	**85.3**	**60.6**	34.7
Average	81.7	57.6	66.6	71.6	48.1	32.9

Table 2. LISTENCARE capability-level End-to-end accuracy. Rows are LALMs and columns are capabilities. Cell shading visualizes accuracy: low → mid → high. The Average row reports the mean over the five model accuracies within each capability.

Model	Full	Local-30s	Δ
Kimi-Audio-7B	37.3	64.9	+27.6
Audio-Flamingo-3	52.3	65.5	+13.2
Audio-Flamingo-Next	62.4	68.6	+6.2
Qwen3-Omni-Inst.	82.4	85.0	+2.5
Qwen3-Omni-Think.	83.6	86.2	+2.6
Matched average	63.6	74.0	+10.4

Table 3. Matched Full audio and Local-30s End-to-end accuracy on the answer-localizable Local-30s subset of 1,492 instances. We exclude items whose answer requires encounter-level voice context unavailable in the local crop. Δ is Local-30s minus Full audio accuracy, in percentage points.

drops sharply for interaction-level reasoning (INT; 48.1%) and audio-specific grounding (AUD; 32.9%). The AUD column is especially notable: Qwen3-Omni-Thinking matches or exceeds Qwen3-Omni-Instruct on the non-AUD capabilities, but drops from 60.5% to 34.7% on AUD, consistent with the Qwen3-Omni report's finding that the Thinking variant can lag behind Instruct on perception-heavy audio and music tasks (Xu et al., 2025). Appendix Table 6 shows that Cascade improves several categories, including stated-fact retrieval (EVR/SFR; +13.1 pp), source-provenance attribution (ATG/PRO; +8.9 pp), and audio-specific grounding (AUD/ACG; +15.5 pp), but has little effect on speaker-structure (ATG/SDS; +0.2 pp) or speaker-role (ATG/SPR; +0.1 pp) grounding. Together, these results suggest that transcripts mitigate some long-audio QA failures, but text-only representations do not uniformly preserve the speaker and acoustic information needed across the taxonomy.

5.4. RQ3: Evidence-Localized Audio vs. Full Audio

RQ3 tests whether errors in the Full audio condition reflect a failure to understand the answer-relevant local evidence, or a failure to find that evidence in full audio. We compare the matched Full condition with Local-30s (*i.e.*, an oracle evidence-localized crop centered on aligned evidence turns) on an answer-localizable subset of 1,492 instances, excluding items whose answer requires encounter-level voice context unavailable in the local crop (Table 3). Local-30s improves all five models, with the largest gains for Kimi-Audio (+27.6 pp) and Audio-Flamingo-3 (+13.2 pp), a smaller gain for Audio-Flamingo-Next (+6.2 pp), and modest gains for the Qwen3-Omni variants (+2.5 to +2.6 pp). Taken together, these patterns suggest that weaker LALMs are substantially retrieval-limited over long encounters, whereas stronger Qwen3-Omni models reduce, but do not eliminate, long-audio evidence-retrieval errors on answer-localizable items.

6. Discussion

LISTENCARE provides a diagnostic evaluation layer between speech recognition and downstream speech-driven clinical applications. Our experimental results reveal a more nuanced picture of long-form clinical conversation understanding in LALMs: stronger LALMs can recover much of the stated clinical evidence, while less capable models still show audio–text modality gaps and difficulty retrieving evidence from long audio. Yet even stronger models remain limited on interaction-level reasoning and audio-specific grounding. Future work should evaluate open-ended answers beyond option discrimination and connect these diagnostic errors to downstream clinical application quality.

Limitations. LISTENCARE is still limited by source realism, annotation scale, and clinical governance: most encounters are synthetic (400/457; 87.5%), and both sources are cleaner than real clinical encounters in acoustic conditions, speech patterns, and interaction dynamics. The benchmark is also English-only, and the four-option MCQA format measures discrimination among given options rather than free-form generation, which may overestimate what models can produce without candidate answers.

Ethics and Release

The public release is designed around synthetic and human-recorded mock-consultation sources, not patient-level private clinical audio. Any release of real clinical encounters would require separate governance, de-identification, and consent review. The release includes auditable question schemas, evaluation scripts, and non-sensitive artifacts where permitted.

References

Ahia, O., Bartelds, M., Ahuja, K., Gonen, H., Hofmann, V., Arora, S., Li, S. S., Puttagunta, V., Adeyemi, M., Buchireddy, C., Walls, B., Bennett, N., Watanabe, S., Smith, N. A., Tsvetkov, Y., and Kumar, S. BLAB: Brutally long audio bench. *arXiv preprint arXiv:2505.03054*, 2025. doi: 10.48550/arXiv.2505.03054. URL https://arxiv.org/abs/2505.03054.

Ben Abacha, A., Yim, W.-w., Adams, G., Snider, N., and Yetisgen, M. Overview of the MEDIQA-Chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pp. 503–513, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.clinicalnlp-1.52. URL https://aclanthology.org/2023.clinicalnlp-1.52.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., and Zhou, J. 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.

Ghosh, S., Goel, A., Kim, J., Kumar, S., Kong, Z., Lee, S.-g., Yang, C.-H. H., Duraiswami, R., Manocha, D., Valle, R., and Catanzaro, B. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 41819–41886, 2025.

Ghosh, S., Goel, A., Jayakumar, K., Koroshinadze, L., Anand, N., Kong, Z., Gururani, S., Lee, S.-g., Kim, J., Aljafari, A., et al. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026. doi: 10.48550/arXiv.2604.10905. URL https://arxiv.org/abs/2604.10905.

Hartman, V., Zhang, X., Poddar, R., et al. Developing and evaluating large language model–generated emergency medicine handoff notes. *JAMA Network Open*, 7(12): e2448723, 2024. doi: 10.1001/jamanetworkopen.2024.48723.

He, P., Wen, Z., Wang, Y., Wang, Y., Liu, X., Huang, J., Lei, Z., Gu, Z., Jin, X., Yang, J., Li, K., Liu, Z., Li, W., Wang, C., He, C., and Zhang, L. AudioMarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio LLMs. *arXiv preprint arXiv:2510.07293*, 2025. URL https://arxiv.org/abs/2510.07293v1. Version 1.

KimiTeam, Ding, D., Ju, Z., Leng, Y., Liu, S., Liu, T., Shang, Z., Shen, K., Song, W., Tan, X., et al. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*, 2025. doi: 10.48550/arXiv.2504.18425. URL https://arxiv.org/abs/2504.18425.

Krishna, K., Khosla, S., Bigham, J., and Lipton, Z. C. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4958–4972. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.384. URL https://aclanthology.org/2021.acl-long.384/.

Kumar, S., Sedlacek, S., Lokegaonkar, V., Lopez, F., Yu, W., Anand, N., Ryu, H., Chen, L., Plicka, M., Hlavacek, M., Ellingwood, W. F., Udupa, S., Hou, S., Ferner, A., Barahona, S., Bolanos, C., Rahi, S., Herrera-Alarcon, L., Dixit, S., Patil, S., Deshmukh, S., Koroshinadze, L., Liu, Y., Garcia Perera, L. P., Zanou, E., Stafylakis, T., Chung, J. S., Harwath, D., Zhang, C., Manocha, D., Lozano-Diez, A., Kesiraju, S., Ghosh, S., and Duraiswami, R. MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*, 2025. doi: 10.48550/arXiv.2508.13992. URL https://arxiv.org/abs/2508.13992.

Labrak, Y., Grunert, D., Baroudi, S., Chun, J., Cyrta, P., Burdisso, S., Hassoon, A., Liu, D., Rothschild, A., Van Deusen, R., Motlicek, P., Perrault, A., Marxer, R., and Schaaf, T. Generating synthetic doctor-patient conversations for long-form audio summarization. *arXiv preprint arXiv:2604.06138*, 2026. doi: 10.48550/arXiv.2604.06138. URL https://arxiv.org/abs/2604.06138. Submitted for review at Interspeech 2026.

Papadopoulos Korfiatis, A., Moramarco, F., Sarac, R., and Savkov, A. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 588–598, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.65. URL https://aclanthology.org/2022.acl-short.65/.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Rajgarhia, H., Ojha, S., Shaik, A., Pothanapalli, A., Lokesh, R., Mukherji, A., and Desikan, P. Medmosaic: A challenging large scale benchmark of diverse medical audio. *arXiv preprint arXiv:2605.00969*, 2026. URL https://arxiv.org/abs/2605.00969.

Sakshi, S., Tyagi, U., Kumar, S., Seth, A., Selvakumar, R., Nieto, O., Duraiswami, R., Ghosh, S., and

Manocha, D. MMAU: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024. URL https://arxiv.org/abs/2410.19168.

Shao, M., Su, H., Tian, W., Mu, B., Lin, Z., Fan, L., Luo, Z., Luan, J., and Xie, L. Listening with time: Precise temporal awareness for long-form audio understanding. *arXiv preprint arXiv:2604.22245*, 2026. doi: 10.48550/arXiv.2604.22245. URL https://arxiv.org/abs/2604.22245.

Steimetz, E., Minkowitz, J., Gabutan, E. C., et al. Use of artificial intelligence chatbots in interpretation of pathology reports. *JAMA Network Open*, 7(5):e2412767, 2024. doi: 10.1001/jamanetworkopen.2024.12767.

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., and Zhang, C. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Tierney, A. A., Gayre, G., Hoberman, B., et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery*, 5(3), 2024. doi: 10.1056/CAT.23.0404.

Wang, B., Zou, X., Lin, G., Sun, S., Liu, Z., Zhang, W., Liu, Z., Aw, A. T., and Wang, N. F. C. AudioBench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*, 2024. URL https://arxiv.org/abs/2406.16020.

Williams, C. Y. K., Subramanian, C. R., Ali, S. S., et al. Physician- and large language model–generated hospital discharge summaries. *JAMA Internal Medicine*, 185(7):818–825, 2025. doi: 10.1001/jamainternmed.2025.0821.

Xiang, B., Zhao, S., Guo, T., and Zou, W. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5187–5202. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.262. URL https://aclanthology.org/2025.emnlp-main.262/.

Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. doi: 10.48550/arXiv.2509.17765. URL https://arxiv.org/abs/2509.17765.

Yang, F., Ni, X., Yang, R., Geng, J., Li, Q., Lyu, C., Du, Y., Wang, L., Luo, W., and Zhang, K. LongSpeech: A scalable benchmark for transcription, translation and understanding in long speech. *arXiv preprint arXiv:2601.13539*, 2026. doi: 10.48550/arXiv.2601.13539. URL https://arxiv.org/abs/2601.13539.

Yang, Q., Xu, J., Liu, W., Chu, Y., Jiang, Z., Zhou, X., Leng, Y., Lv, Y., Zhao, Z., Zhou, C., and Zhou, J. 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, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.109. URL https://aclanthology.org/2024.acl-long.109/.

Yim, W.-w., Fu, Y., Ben Abacha, A., Snider, N., Lin, T., and Yetisgen, M. ACI-BENCH: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586, 2023. doi: 10.1038/s41597-023-02487-3. URL https://www.nature.com/articles/s41597-023-02487-3.

Zaretsky, J., Kim, J. M., Baskharoun, S., Zhao, Y., Austrian, J., Aphinyanaphongs, Y., Gupta, R., Blecker, S. B., and Feldman, J. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Network Open*, 7(3):e240357, 2024. doi: 10.1001/jamanetworkopen.2024.0357.

Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., Zhou, M., Zeng, J., Dong, X., Zhang, R., Fang, H., Zhu, P., Chen, S., and Xie, P. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9241–9250, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.743. URL https://aclanthology.org/2020.emnlp-main.743.

Appendix Contents

A. QA Construction Taxonomy

This appendix defines the LISTENCARE QA construction taxonomy used for question construction and score slicing. The taxonomy contains six *parent capabilities* and 12 more specific *QA categories*.

A.1. Capability Taxonomy

Table 4 gives the category set used for QA construction and reporting. The parent capability is the coarse analysis axis, and the QA category is the instance-level slice. We assign stable three-letter tags to both parent capabilities and QA categories. Tables use the compact parent/category form, such as ATG/VTC, so that appendix definitions, manifest metadata, and score slices share the same identifiers.

Table 4. Capability taxonomy for LISTENCARE QA construction, with six parent capabilities and 12 QA categories.

Tag	Parent capability	QA category	What the category probes
EVR/SFR	Evidence recovery	Stated fact retrieval	Explicitly stated encounter facts, values, entities, and actions.
EST/POL	Evidence status and sufficiency	Polarity / assertion status	Whether a claim is affirmed, denied, unclear, or not discussed.
EST/ANS	Evidence status and sufficiency	Answerability / insufficient evidence	Whether the encounter contains enough evidence to answer, or whether the model should abstain.
ATG/SDS	Attribution grounding	Speaker dialogue structure	Speaker-structure probes that do not rely on clinical content or role labels: distinct-speaker counting, endpoint-speaker tracking, and participation-time dominance.
ATG/SPR	Attribution grounding	Speaker role grounding	Mapping voices or turn behavior to roles such as clinician, patient, caregiver, nurse, or interpreter.
ATG/VTC	Attribution grounding	Voice-to-content attribution	Linking a fact, concern, value, refusal, or plan to the audible speaker who produced it.
ATG/PRO	Attribution grounding	Source / provenance attribution	Distinguishing patient self-report, family history, prior clinician statements, current clinician recommendations, and other sources.
TDT/TST	Temporal and discourse tracking	Temporal / state tracking	Event currentness, sequence, unresolved details, and final supported state across turns, with minimum-turn boundary rules.
TDT/COR	Temporal and discourse tracking	Cross-turn linking / coreference	Linking pronouns, deixis, back-references, and response turns to earlier evidence across at least two turns.
INT/GIP	Interaction-level reasoning	Goal / intent / preference inference	Interaction goals, participant intent, stance, preferred outcome, or desired care trajectory.
INT/MHI	Interaction-level reasoning	Multi-hop evidence integration / rationale	Combining multiple turns to explain a plan, conflict, missing-information chain, or rationale.
AUD/ACG	Audio-specific grounding	Audio-cue grounding	Non-lexical acoustic cues, non-speech sounds, timing, prosody, cue-to-speaker attachment, or absence-control audio cues.

A.2. Representative MCQA Examples

The following examples are drawn from the final LISTENCARE benchmark. They illustrate how each QA category appears as a four-option MCQA item.

Table 5. Representative MCQA examples drawn from the final LISTENCARE benchmark and organized by the capability taxonomy.

Tag	QA category	Example question
EVR/SFR	Stated fact retrieval	**Q:** What age does the patient state? **Options:** A. Twenty-five; B. Fifty; C. Fifteen; D. Fifty-five **Answer:** A. Twenty-five.
EST/POL	Polarity / assertion status	**Q:** Is the claim "the patient smokes tobacco" AFFIRMED, DENIED, UNCLEAR, or NOT DISCUSSED in this encounter? **Options:** A. AFFIRMED; B. DENIED; C. UNCLEAR; D. NOT DISCUSSED **Answer:** B. DENIED.
EST/ANS	Answerability / insufficient evidence	**Q:** What is the exact pulse rate recorded during the physical examination? **Options:** A. The pulse rate is not addressed in the encounter; B. The patient explicitly states that the exact pulse rate is unknown; C. The exact pulse rate is explicitly stated in the encounter; D. Only a partial or approximate pulse rate is provided **Answer:** C. The exact pulse rate is explicitly stated in the encounter.
ATG/SDS	Speaker dialogue structure	**Q:** How many participant voices speak during the encounter? **Options:** A. Three speakers; B. One speaker; C. Two speakers; D. Four speakers **Answer:** C. Two speakers.

Tag	QA category	Example question
ATG/SPR	Speaker role grounding	**Q:** Which role best describes the opening voice of the encounter? **Options:** A. A caregiver or family member; B. The clinician; C. The patient; D. Cannot be determined from the encounter audio **Answer:** B. The clinician.
ATG/VTC	Voice-to-content attribution	**Q:** Which participant voice reports an allergy to penicillin? **Options:** A. A third or brief additional participant voice in the encounter audio; B. The second recurring voice in the encounter audio; C. No participant says this in the encounter audio; D. The first recurring voice in the encounter audio **Answer:** D. The first recurring voice in the encounter audio.
ATG/PRO	Source / provenance attribution	**Q:** Whose stroke is referenced as part of the family medical background? **Options:** A. The patient's brother during his stroke; B. The patient's uncle during his stroke; C. The patient during her own stroke; D. The patient's father during his stroke **Answer:** B. The patient's uncle during his stroke.
TDT/TST	Temporal / state tracking	**Q:** Does the uncertainty about which elbow is swollen get resolved later in the encounter? **Options:** A. No, the side remains ambiguous throughout the clinical assessment; B. Yes, only the right elbow is later confirmed during the physical exam; C. Yes, it is confirmed as the right elbow after the physical examination; D. Yes, the patient clarifies it is the left elbow during the history taking **Answer:** D. Yes, the patient clarifies it is the left elbow during the history taking.
TDT/COR	Cross-turn linking / coreference	**Q:** When the patient says "That sounds... sensible," what proposal is she acknowledging? **Options:** A. Assessing her joint discomfort with a topical anti-inflammatory cream; B. Reviewing her complete family history for hereditary cardiovascular conditions; C. A baseline cognitive assessment followed by laboratory blood work; D. Managing her elevated blood pressure through diet and daily walking **Answer:** C. A baseline cognitive assessment followed by laboratory blood work.
INT/GIP	Goal / intent / preference inference	**Q:** The patient accepts the proposed testing, but with what primary condition? **Options:** A. Proceed if pain is avoided; B. Proceed if results are explained plainly; C. Proceed if it is quick; D. Proceed if cause is identified **Answer:** C. Proceed if it is quick.
INT/MHI	Multi-hop evidence integration / rationale	**Q:** Which earlier topic returns later? **Options:** A. Knee stiffness; B. Family heart history; C. Sun exposure; D. Occasional ibuprofen use **Answer:** C. Sun exposure.
AUD/ACG	Audio-cue grounding	**Q:** How does the doctor's response to the patient's request begin? **Options:** A. It begins with a hiccup; B. It begins with crying; C. It begins with a sigh; D. It begins with a sniffle **Answer:** C. It begins with a sigh.

B. Benchmark Experiment

This appendix collects the model implementation details and full evaluation tables for LISTENCARE over the five LALMs evaluated in the main results.

B.1. Model and Implementation Details

We evaluate each model in the same answer-selection format: the model receives the encounter evidence view, the text question, and four answer options, and the parser extracts a single option letter. End-to-end and Local audio runs use the model's audio input path directly; Cascade runs use the same Whisper large-v3-turbo transcript for all models. Invalid or unparsable generations are counted as invalid outputs and are never imputed to a default answer.

- **Kimi-Audio-7B.** We use `moonshotai/Kimi-Audio-7B-Instruct` with a vLLM chat-audio backend, file-URL audio input, and audio-before-text prompt order. Decoding uses temperature 0.0, top-k 5, maximum 4096 generated tokens, seed 42, and the model-specific `[EOS]` stop token. Kimi-Audio does not state an explicit maximum input duration in the technical report; the released checkpoint configuration exposes 8192 positions and the report describes a 12.5 Hz audio representation, giving an audio-only upper bound of 655 s (10.9 min) before text and output overhead. A duration audit of LISTENCARE finds that 91/457 encounters (805/4,085 QA instances) exceed this audio-only bound, so Kimi-Audio should be interpreted as a shorter-context baseline on the longest encounters.
- **Audio-Flamingo-3.** We use `nvidia/audio-flamingo-3-hf` through vLLM with file-URL audio input and audio-before-text prompt order. Decoding uses temperature 0.0, maximum 4096 generated tokens, and seed 42. The model card documents a 10 min maximum audio input, making this model a shorter-context long-audio baseline.
- **Audio-Flamingo-Next.** We use `nvidia/audio-flamingo-next-hf` through the Hugging Face Transformers backend because the local vLLM image did not initialize this checkpoint reliably. Runs use text-before-audio prompt order and bfloat16 weights; decoding uses temperature 0.0, repetition penalty 1.2, and maximum 4096 generated tokens. The released processor configuration uses internal 30-second audio windows and supports up to 1800 s (30 min) of audio.
- **Qwen3-Omni-30B-A3B-Instruct.** We use `Qwen/Qwen3-Omni-30B-A3B-Instruct` through vLLM with file-URL audio input and audio-before-text prompt order. The vLLM configuration uses bfloat16 weights, tensor parallelism over two GPUs, maximum model length 32768, and up to three audio items per prompt. Decoding follows the model-card-style setting with temperature 0.6, top-p 0.95, top-k 20, maximum 4096 generated tokens, and seed 42. The Qwen3-Omni report documents audio/spoken-language evaluation up to 40 min per instance.
- **Qwen3-Omni-30B-A3B-Thinking.** We use `Qwen/Qwen3-Omni-30B-A3B-Thinking` through vLLM with the same audio transport, prompt order, context length, and decoding settings as the Instruct variant. We enable the Qwen3 reasoning parser so that final-answer extraction is applied to the parsed answer channel rather than raw reasoning text. We treat its documented input-duration budget as the same Qwen3-Omni 40 min-per-instance setting.

B.2. Category-Level Results

Category	n	End-to-end	Cascade	Gap
EVR/SFR	366	81.7	94.7	+13.1
EST/ANS	347	56.1	64.4	+8.3
EST/POL	440	58.6	65.6	+7.0
ATG/SDS	457	64.0	64.2	+0.2
ATG/SPR	457	60.9	61.1	+0.1
ATG/VTC	361	64.0	69.4	+5.4
ATG/PRO	356	80.1	89.1	+8.9
TDT/COR	284	71.9	76.9	+5.0
TDT/TST	268	71.3	77.5	+6.2
INT/GIP	263	52.7	54.1	+1.3
INT/MHI	266	43.5	48.1	+4.6
AUD/ACG	220	32.9	48.5	+15.5

Table 6. Category-level End-to-end vs. Cascade results over the five primary LALMs. End-to-end evaluates full encounter audio; Cascade evaluates the same LALM on an ASR transcript generated with Whisper large-v3-turbo. Gap is Cascade accuracy minus End-to-end accuracy in pp.

B.3. Model-Level Results

Model	End-to-end	Cascade	Gap	E2E invalid	Cascade invalid
Kimi-Audio-7B	36.9	55.9	+19.0	1 (0.0)	53 (1.3)
Audio-Flamingo-3	54.4	59.3	+4.9	0 (0.0)	0 (0.0)
Audio-Flamingo-Next	62.1	62.4	+0.3	135 (3.3)	133 (3.3)
Qwen3-Omni-Inst.	77.8	78.7	+0.8	0 (0.0)	0 (0.0)
Qwen3-Omni-Think.	79.9	80.9	+1.0	22 (0.5)	171 (4.2)
Matched average	62.2	67.4	+5.2	–	–

Table 7. Model-level RQ1 results for the five primary LALMs over the 4,085-instance LISTENCARE benchmark. Values are accuracy percentages. Invalid columns report invalid or unparsable outputs as count (percent of instances).

B.4. Model-by-Category Diagnostics

Category	Kimi	AF3	AFN	Q3-I	Q3-T
EVR/SFR	28.7	90.4	92.6	98.4	98.4
EST/ANS	26.5	47.0	57.3	67.7	82.1
EST/POL	29.1	48.6	55.1	80.7	79.5
ATG/SDS	54.3	37.9	74.0	75.1	78.6
ATG/SPR	42.2	33.0	53.6	84.2	91.5
ATG/VTC	38.0	62.3	59.8	75.9	83.9
ATG/PRO	41.6	84.8	86.2	93.5	94.4
TDT/COR	38.0	71.5	80.3	85.2	84.5
TDT/TST	45.1	69.8	71.4	84.0	86.2
INT/GIP	37.0	49.4	51.3	60.5	65.4
INT/MHI	27.8	36.5	46.6	50.8	55.8
AUD/ACG	26.4	20.5	22.7	60.5	34.7

Table 8. Model-by-category End-to-end accuracy matrix for the five primary LALMs. Values are percentages over each category. Abbreviations: AF3=Audio-Flamingo-3, AFN=Audio-Flamingo-Next, Q3=Qwen3-Omni, I=Instruct, T=Thinking.

Category	Kimi	AF3	AFN	Q3-I	Q3-T
EVR/SFR	92.3	91.8	93.1	97.8	98.6
EST/ANS	46.1	48.1	64.6	78.1	85.1
EST/POL	57.5	52.5	55.4	81.8	80.6
ATG/SDS	51.9	41.6	74.8	71.1	81.4
ATG/SPR	36.9	54.3	47.0	74.2	92.9
ATG/VTC	47.3	70.1	65.7	78.4	85.7
ATG/PRO	84.2	84.8	86.5	94.1	95.7
TDT/COR	66.1	72.2	76.4	84.2	85.5
TDT/TST	74.5	65.3	73.1	86.9	87.5
INT/GIP	49.2	47.1	48.3	63.1	62.6
INT/MHI	39.2	42.9	44.3	51.1	62.9
AUD/ACG	28.5	34.5	24.5	76.4	78.4

Table 9. Model-by-category Cascade accuracy matrix for the five primary LALMs. Each cell evaluates the same LALM on a Whisper large-v3-turbo transcript rather than the audio waveform. Values are percentages.

Category	Kimi	AF3	AFN	Q3-I	Q3-T
EVR/SFR	+63.6	+1.4	+0.5	-0.6	+0.3
EST/ANS	+19.6	+1.1	+7.3	+10.4	+3.0
EST/POL	+28.4	+3.9	+0.3	+1.1	+1.1
ATG/SDS	-2.4	+3.7	+0.8	-3.9	+2.8
ATG/SPR	-5.3	+21.3	-6.6	-10.1	+1.4
ATG/VTC	+9.3	+7.8	+5.9	+2.5	+1.8
ATG/PRO	+42.6	+0.0	+0.3	+0.6	+1.3
TDT/COR	+28.1	+0.7	-3.9	-1.0	+1.0
TDT/TST	+29.4	-4.5	+1.7	+2.9	+1.3
INT/GIP	+12.2	-2.3	-3.0	+2.6	-2.8
INT/MHI	+11.4	+6.4	-2.3	+0.3	+7.1
AUD/ACG	+2.1	+14.0	+1.8	+15.9	+43.7

Table 10. Model-by-category End-to-end vs. Cascade gap matrix for the five primary LALMs. Each cell is Cascade accuracy minus End-to-end accuracy in pp. Blue cells favor Cascade, orange cells favor End-to-end, and gray cells mark exact ties after rounding.

B.5. Local-Window Size Ablation

	30-second window			60-second window		
Model	Full@n	Local	Gap	Full@n	Local	Gap
Kimi-Audio-7B	37.3	64.9	+27.6	36.3	58.3	+22.0
Audio-Flamingo-3	52.3	65.5	+13.2	56.0	67.2	+11.3
Audio-Flamingo-Next	62.4	68.6	+6.2	65.1	71.1	+6.0
Qwen3-Omni-Inst.	82.4	85.0	+2.5	83.0	85.2	+2.2
Qwen3-Omni-Think.	83.6	86.2	+2.6	84.7	87.0	+2.3
Matched average	63.6	74.0	+10.4	65.0	73.8	+8.8

Table 11. Local-window size ablation for evidence-localized audio on answer-localizable items. We compare 30-second and 60-second oracle crops with matched Full audio predictions. Gap is Local minus Full@n in pp. Local subsets contain 1,492 instances (30-second window) and 2,157 instances (60-second window) for most rows; Qwen3-Omni-Thinking has two fewer matched predictions for the 30-second window and three fewer for the 60-second window.