SCENEBench: An Audio Understanding Benchmark Grounded in Assistive and Industrial Use Cases

Anonymous ACL submission

Abstract

Advances in large language models (LLMs) have enabled significant capabilities in audio processing, resulting in state-of-the-art models now known as Large Audio Language Models (LALMs). However, minimal work has been done to measure audio understanding beyond automatic speech recognition (ASR). This paper closes that gap by proposing a benchmark suite, SCENEBench (Spatial, Cross-lingual, Environmental, Non-speech Evaluation), that targets a broad form of audio comprehension across four real-world categories: `background sound understanding`, `noise localization`, `cross-linguistic speech understanding`, and `vocal characterizer recognition`. In addition to performance, we also measure model latency. The purpose of this benchmark suite is to assess the audio beyond just what words are said— rather, in *how* they are said and the non-speech components of the audio. Based on the needs articulated by audio understanding use-cases of accessibility technology and industrial noise monitoring, this benchmark reveals critical gaps in current LALMs. The performance in each task is quite varied, with some tasks having performance far below random chance and others with high accuracy. These results provide direction for targeted improvements in model capabilities.

1 Introduction

Remarkable strides in natural language understanding have powered a wide range of applications in search, conversation, and information retrieval. As the capabilities of these models improve, we must develop methods to evaluate them effectively.

Speech, a widely shared mode of human communication, is an extension for text-based large language models. However, understanding spoken language is not limited to transcription. Proper audio comprehension involves recognizing tone, emotion, background noise, environmental context, speaker intent, and more. These elements often coexist and can significantly impact meaning. They also enable practical systems, such as assistive devices or captioning tools that describe traffic sounds and approaching sirens for individuals with hearing impairments, as well as telehealth and well-being systems that detect coughs, or sobs in patient speech.

Companies releasing LALMs, such as GPT-4o (OpenAI, 2024) and Qwen2-Audio (Chu et al., 2024), advertise capabilities beyond ASR—for example, Alibaba states that Qwen2-Audio "can transcribe speech and identify audio info ... including spoken words, music and ambient noises" (Alibaba Cloud, 2024). However, current evaluation strategies primarily measure speech recognition, not audio understanding. That is, they *assess what words were said, not how they were said, or the non-speech components of the audio.*

In this paper, we present a new benchmark suite, SCENEBench, for evaluating a broader conception of audio understanding in LALMs, grounded in the needs of accessibility technology and industrial noise monitoring. Many prior benchmarks focus on controlled, single-modality, or clean scenarios. Our benchmark tests four categories that reflect real-world complexity: `background sound understanding`, `noise localization`, `cross-linguistic speech understanding`, and `vocal characterizers` (e.g., crying). All of these are measured along with model latency, a parallel dimension of our benchmark suite.

2 Related Works

In Section 2.1, we review and break down the scope of prior benchmarks for audio and speech understanding (Table 1). In Section 2.2, we analyze gaps through two high-impact settings with clear stakes: accessibility and industrial monitoring. We will also discuss how our benchmark suite generalizes to other contexts.

Benchmark	Type	Multi turn	Primary Focus	Clips (≈)	#Tasks	Bkgd Sound	Noise Loc.	Cross Ling.	Vocal Chars	Latency
SCENEBench	MC+FRQ	✓	Audio & speech understanding	16k	4	✓	✓	✓	✓	✓
AudioBench (Wang et al., 2025)	MC+FRQ	✗	ASR, scene & voice	100k	8	✓	✗	✗	✗	✗
MMAU (Sakshi et al., 2024)	MC	✗	Multi-task reasoning	10k	27	✓	✗	✗	✗	✗
AIR-Bench (Yang et al., 2024)	MC+FRQ	✓	Generative comprehension	21k	20	✗	✗	✗	✗	✗
MARBLE (Yuan et al., 2023)	MC	✗	Music classification	25.9k	18	✗	✗	✗	✗	✗
Clotho-AQA (Lipping et al., 2022)	FRQ	✗	Environmental QA	1.9k	1	✓	✗	✗	✗	✗
Sound Check (Agnew et al., 2024a)	Audit	✗	Dataset quality audit	3M	7	✓	✗	✗	✗	✗
SONAR (Li et al., 2024)	CLS	✗	Deepfake detection	2.2k	1	✗	✗	✗	✗	✗
CAVA (Held et al., 2025)	Task	✓	Voice-assistant behaviour	6.4k	6	✗	✗	✗	✗	✓

Table 1: Benchmark comparison with standardized size and coverage (no datasets column). *Type*: MC = multiple choice, FRQ = free response, CLS = classification, Audit = dataset audit. *Clips (≈)* reports total items when stated or derivable; *#Tasks* counts distinct task families (— where not reported). Coverage columns indicate whether each suite evaluates that dimension.

2.1 Benchmarks for Audio and Speech Understanding

Several benchmarks evaluate model capabilities on speech and audio inputs, each making design choices to hone in on their main focus.

AudioBench (Wang et al., 2024) prioritizes broad coverage and automatically gradable tasks. This design enables reproducible cross-model comparisons at a large scale and lowers annotation cost. However, it deemphasizes phenomena that are harder to capture and grade automatically. **MMAU** (S. et al., 2024) adopts a multiple choice format across diverse tasks to improve inter-annotator agreement, reduce evaluation variance and prompt sensitivity, and simplify scoring across large model suites. The MC constraint makes results comparable and stable, though it limits models' ability to demonstrate open-ended reasoning, justifications, or descriptions of subtle acoustic attributes. **AIR-Bench** (Lee et al., 2024) introduces open-ended audio Qs, but places a large emphasis on music and environmental sounds rather than detailed speech pragmatics and paralinguistics. **CAVA** (Held et al., 2025) targets voice-assistant behavior (e.g., instruction following, latency), aligning with real deployment concerns. That focus surfaces agentic performance and responsiveness under constraints, while probing less of the fine-grained acoustic understanding. Other efforts, such as **SoundCheck** (Agnew et al., 2024b) and **SONAR** (Jain et al., 2024), audit datasets or test deepfakes, serving specific goals rather than audio understanding as a whole.

Our benchmark suite is designed to complement the space of existing audio benchmarks by prioritizing the following: adding targeted evaluations for our four tasks and including free-response scoring where it is required. These are areas that are underrepresented in the prioritization of existing benchmarks. A side-by-side comparison with prior benchmarks appears in Table 1. These gaps matter most in high-stakes, real-world deployments and we will discuss them further in the next section.

2.2 Use-Case Driven Gaps

From a use-case standpoint, existing audio benchmarks are still biased toward "clean-room" speech-recognition or captioning scenarios and rarely touch the two domains where errors are most consequential: **(i) accessibility** and **(ii) industrial sound monitoring**.

Accessibility Everyday scenes feature cross-talk, traffic, devices, and emotional or whispered speech. For Deaf or hard-of-hearing (DHH) users, useful support goes beyond transcription to sound awareness (Jain et al., 2019; Bragg et al., 2016). Some examples of crucial instances requiring the surfacing of salient non-speech events include hearing for sirens and their state (approaching vs. receding), and disambiguating paralinguistic cues (fatigue, sobs, whispers) (Findlater et al., 2019; Wu and Jain, 2025; Kim et al., 2023).

Wearable and IoT systems that detect sirens and vocalizations and relay haptic/onscreen alerts have been shown to reduce risk (Chin et al., 2023; Salem et al., 2023). Conversely, background noise (cross-talk, broadband noise) can degrade ASR rather than enrich it, underscoring the need for models to understand situational context, not just foreground words (Dilmegani, 2025).

Industrial monitoring Factory floors and labs demand early detection of anomalous machine sounds that often occur under speech or ambient noise. Public datasets such as ToyADMOS and MIMII established common testbeds for acoustic anomaly detection (AAD) in machine condition monitoring (Koizumi et al., 2019; Purohit et al., 2019). More recently, MIMII-DG introduced domain shifts across machine types and recording setups to test generalization of these models which is often the key blocker in deployment (Dohi et al., 2022). Missing or misclassifying subtle cues translates directly into safety risks.

Other applications Similar requirements recur in telehealth and well-being such as, cough and respiratory-symptom screening, fatigue/sleepiness cues (Orlandic et al., 2021; Sharma et al., 2020; Schuller et al., 2011), in smart-home/IoT for domestic activity and rare-event detection such as, glass breaks (Mesaros et al., 2018, 2017), transportation and public safety, such as siren/horn awareness in urban scenes (Gemmeke et al., 2017), and AR/VR and robotics, where spatialized audio cues support navigation and interaction (Gordon et al., 2020; Chen et al., 2020).

Rationale for our benchmark. Across these settings, four failure modes appear repeatedly: important events may occur in the background; spatial/proximity cues are lost in audio understanding; transcription quality drops under multilingual spans; and paralinguistic information carried by non-speech is underspecified. SCENEBench turns each failure mode into a targeted evaluation. These four tasks are necessary because they probe salient audio-understanding capabilities tied to our use cases, and they are motivated by concrete failure patterns and potential user benefit. We exclude full spatial-audio setups and long-form clips in this release to keep the benchmark minimal, reproducible, and aligned with real-time assistive and monitoring scenarios. Taken together, these task choices yield a suite that diagnoses where current LALMs succeed and where they still fail.

3 Methods

Our benchmark suite is constructed on top of existing datasets, to which we apply task-specific transformations. In this section, we describe the data construction process for each of the four tasks, highlighting how it diverges from conventional uses of the source dataset. We evaluate leading models such as GPT-4o (OpenAI, 2024), Qwen2-Audio-7B-Instruct (Chu et al., 2024), and Gemini-2.5 (Comanici et al., 2025), we aim to reveal both the current limitations and untapped potential of LALMs in audio-first contexts.

3.1 Tasks

Our benchmark comprises four task types designed to probe distinct dimensions of audio understanding in real-world settings. In Section 3.1.1, we test `background sound understanding` by embedding environmental audio under speech and assessing models' ability to identify it. Section 3.1.2 introduces a `noise localization` task based on amplitude change, targeting scenarios like siren detection for accessibility. Section 3.1.3 examines `cross-linguistic` robustness by asking to transcribe utterances in multiple languages. Finally, Section 3.1.4 focuses on paralinguistic cues by testing whether models can recognize and label `vocal characterizers` such as laughter and whispering. In Section 3.2, we describe a parallel dimension of our benchmark, model latency, across all our tasks to measure real-time performance constraints.

3.1.1 Background Sound Understanding

The first task examines whether a model can identify background noise layered under speech. While prior work such as WSJ0-2mix (Hershey et al., 2016; Isik et al., 2016) and Libri2Mix (LibriMix) (Cosentino et al., 2020) targets speaker separation, few benchmarks probe environmental sounds under speech. We therefore overlay ESC-50 categories (Piczak, 2015) onto DailyTalk utterances (Lee et al., 2023), using two voices (higher vs. lower pitch) to create two versions of each of 2,000 clips. The background noise and speech are overlaid with original volumes, which we discuss the limitations of in Section 7. Scoring is hierarchical: We evaluate the background-noise task with three prompts: (FR1) free-response description of all audible sounds; (FR2) targeted follow-up that explicitly asks for the background sound (issued only if FR1 omitted it); and (MC1) 4-way forced choice of the background category. Scoring is hierarchical: FR1 (describe audio.) is correct if the free-response mentions any background noise category; FR2 (describe background noise.) is credited if the correct ESC-50 class is named either in FR1 (in which case FR2 receives full credit without being asked) or in the FR2 follow-up; the 4-way multiple-choice probe is administered for all clips and scored independently. Exact prompt wordings appear in Appendix A. Outputs are normalized (lowercasing, punctuation stripping) and matched to a per-class synonym list with simple negation/uncertainty rejection (e.g., "no siren," "not a dog," "unsure"). This three-tier design deliberately separates spontaneous salience (FR1) from targeted retrieval (FR2) and discriminability (MC1), revealing whether failures arise from omission, misnaming, or confusion among plausible classes.

3.1.2 Noise Localization

This task evaluates models' ability to detect dynamic volume patterns in audio, simulating spatial motion through amplitude modulation. We created a dataset from the ESC-50 environmental sound corpus by applying three distinct volume envelopes to 2,000 source audio clips, yielding 6,000 total samples. Each original sound underwent three transformations: (1) approaching sound source, where amplitude scales from 10% to 100% over the clip duration; (2) receding source, with amplitude linearly scaling from 100% to 10%; and (3) oscillating movement past the listener, where amplitude follows a sinusoidal pattern (4 complete cycles) between 20% and 100%.

Models are evaluated using two complementary prompts. First (FR1), a general description prompt asks models to describe all auditory characteristics. Second (FR2), a follow-up position prompt specifically queries about spatial relationships and movement patterns relative to the sound source. Exact prompt wordings appear in Appendix C. Responses are automatically scored as correct if they mention the appropriate motion pattern or appropriate synonyms (e.g., "approaching," "moving away," "oscillating").

3.1.3 Cross-Linguistic Sound Recognition

We evaluate multilingual span transcription by transforming DailyTalk transcripts into controlled language-mixed stimuli: for each longest turn (2,541 total), we translate contiguous spans ($\approx$30%) into one of four languages (Mandarin, Spanish, Hindi, Portuguese) via Google Translate, retain only items passing back-translation similarity > 0.9, and synthesize audio with ElevenLabs multilingual TTS. The resulting audio files, containing multilingual sentences (e.g., "I have a fifteen-day vacation 我想拥有一个 trip to England"), were presented to various LALMs for transcription evaluation (FR1), with performance measured by similarity between model transcriptions and the reference multilingual sentences. Because high-quality human recordings with natural code-switching across these languages are scarce, we adopt this synthetic route for coverage and control; we note its limits briefly here and discuss them in Section 7. As a result, this task, multilingual span transcription is a proxy behavior for code-switching.

3.1.4 Vocal Characterizers

We target non-speech vocal traits—cough, cry, laugh, sneeze, yawn, mumble, and whisper—that carry communicative cues without requiring emotion inference. We deliberately avoid direct emotion classification due to documented ethical concerns about reductive labeling and potential harm (Stark and Hoey, 2021). Our evaluation instead asks models first to briefly describe each clip (FR1), then to perform a 7-way classification over the vocal categories (MC1). The dataset aggregates publicly available repositories: NonSpeech7k for cough/cry/laugh/sneeze/yawn (W4ng1204, 2023), CapSpeech-AgentDB-Audio for mumble/whisper (OpenSound, 2024), with additional mumble from vocal_bursts_taxonomy_100_clean_wds (Krishnakalyan, 2023) and whisper from ASMR (nyuuzyou, 2022). The final set contains 4,006 clips across five reported labels (632 cough, 1,791 cry, 1,133 laugh, 236 sneeze, 214 yawn), plus mumble and whisper for the 7-way classification.

3.2 Latency as a Dimension

Beyond accuracy, we report latency for local models only; cloud/API models are excluded from timing comparisons. For each model invocation (one prompt–audio query), we record a single wall-clock duration

$$T_i = \text{time}_{\text{end}} - \text{time}_{\text{start}},$$

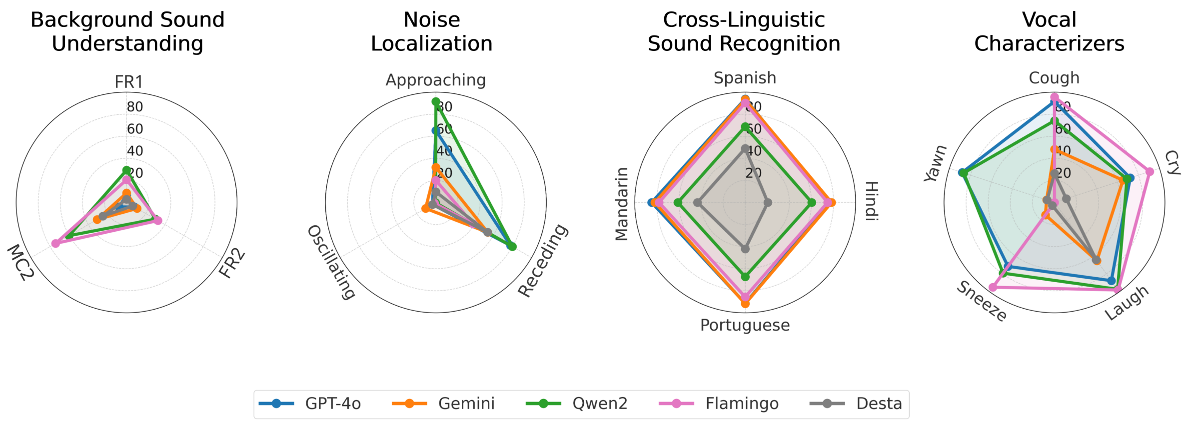

Figure 1: Summary radar charts across tasks. Each axis is a task-specific category; values are percentages (or mean similarity %). Legend shown once: GPT-4o (blue), Gemini (orange), Qwen2 (green), Flamingo (pink), Desta (gray).

where $\text{time}_{\text{start}}$ is stamped immediately before the call to the local inference runtime and time_{end} is stamped when the complete textual response is returned to our harness. We summarize T_i across all four tasks using the median and interquartile range, and we report per-task medians to show how latency varies with input content and prompt type.

3.3 Models Evaluated

We benchmark five state-of-the-art LALMs models with audio capabilities: GPT-4o (OpenAI, USA) (OpenAI, 2024), Gemini 1.5 (Google DeepMind, USA/UK) (Comanici et al., 2025), Qwen2-Audio (Alibaba DAMO Academy, China) (Chu et al., 2024), Audio-Flamingo-3 (NVIDIA, USA) (Goel et al., 2025), and DeSTA2-8B-beta (National Taiwan University + NVIDIA, Taiwan) (Lu et al., 2024). These models span a diverse range of architectures, training paradigms, and geographic origins.

We selected these five models to span the design space that is most relevant to our tasks and to balance fairness with reproducibility: they span commercial models (GPT-4o, Gemini) and open-weights models (Audio-Flamingo-3, Qwen2-Audio-7B, DeSTA2-8B-beta); and they are recent, widely used baselines that claim multilingual and non-speech capability aligned with our tasks. We exclude speech-only ASR and music-specialized models, in this paper, because they cannot run the full suite without additional components that would confound comparisons.

4 Results

We report results across four task types (summary in Figure 1), each designed to probe a distinct dimension of audio understanding. In Section 4.1, we analyze models' ability to detect background sounds layered under speech. Section 4.2 evaluates how well models can estimate the direction of background noise. Section 4.3 presents findings on transcription accuracy in multilingual contexts. Finally, Section 4.4 examines recognition of non-speech vocalizations, such as laughs and coughs, to assess affective and paralinguistic sound understanding.

4.1 Background Sound Understanding

Each clip contains foreground speech with an ESC–50 background sound. Models first give a free response (FR1: "describe the audio."). If FR1 fails to mention the background, we ask a specified follow–up (FR2: "name any specific background sound."). Finally, the model answers a four–way multiple–choice question (MC1). For scoring, if the model already got FR1 right, we count FR2 as correct even when FR2 was not asked (so FR2 is counted for all clips). "Unsure/Cannot tell" is marked incorrect. Bars show means with 95% CIs; horizontal dotted line mark chance (25% for MC1). For readability, we report key outcomes here; full statistical details appear in Appendix B. The 95% CIs are extremely tight ($\leq \sim$1.5 pp half–width at N=4000), so they are not visually distinguishable on the bars. For completeness, they are included in Appendix B.

Latency. For models with timing logs, we sum FR1 + FR2 (only if FR1 failed) + MC1. Flamingo is fast (median **2.26**s; p90 2.73s), while Desta is slow (median **15.61**s). GPT–4o and Gemini are omitted because they are API-based models.

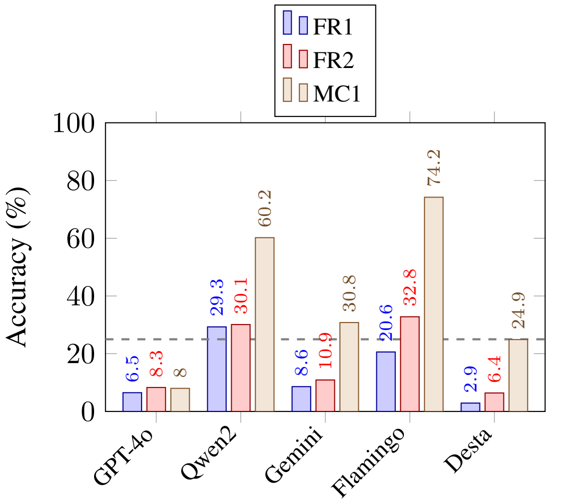

Figure 2: Ambient–sound accuracy across models (N=4000 clips). "Specific Type*" treats FR1–correct as FR2–correct. Baselines for simple chance are 25% (MC; 1 of 4).

Models seldom spontaneously mention background noise (FR1) but perform much better with an explicit prompt (FR2) or choices (MC1). Flamingo leads on FR2/MC1; Qwen2 leads FR1. Direct, specific questions markedly improve performance.

4.2 Background Noise Distance Estimation

We test whether models detect motion via amplitude envelopes applied to ESC-50 clips (Approaching ↑amp, Receding ↓amp, Oscillating sinusoid). Each clip gets two prompts: FR1 (General) free-text description; FR2 (Direction) explicit location/motion query. A response is correct if it names the ground-truth motion.

FR1: General Description FR1 remains challenging for all models. Per-class, Receding is consistently the easiest in FR1 (e.g., Gemini 27.4%), while Oscillating remains near floor for all ($\leq$10%).

Model	Correct	Accuracy (%)	95% CI
Gemini	981	16.35	[15.4, 17.3]
GPT-4o	658	10.97	[10.20, 11.78]
Flamingo	481	8.02	[7.49, 8.58]
Desta	362	6.03	[5.51, 6.65]
Qwen	447	7.45	[6.81, 8.14]

Table 2: Motion FR1 (general description) — accuracy with Wilson 95% CIs; $N = 6000$ per model.

FR2: Direction When we allow the "FR1-already correct auto-correct on FR2" rule and ask direction explicitly, scores jump for the best models:

Model	Correct	Accuracy (%)	95% CI
Qwen	3434	57.23	[55.98, 58.48]
GPT-4o	2938	48.97	[47.70, 50.23]
Gemini	1934	32.23	[31.06, 33.43]
Desta	1344	22.40	[21.36, 23.47]
Flamingo	1209	20.15	[19.11, 21.23]

Table 3: Motion FR2 (direction) — accuracy with Wilson 95% CIs; $N = 6000$ per model. FR2 counts FR1-correct as FR2-correct by design.

Asking directly (FR2) helps: (FR2–FR1) = +38.0 pts (GPT-4o), +49.8 pts (Qwen), +15.9 (Gemini), +12.1 (Flamingo), +16.4 (Desta). Models struggle to spontaneously volunteer motion cues in free text (FR1) but can often answer when asked explicitly (FR2). Oscillation remains an unsolved regime.

4.2.1 Latency

We report "effective" per-clip latency = FR1 latency + FR2 latency if FR1 failed (0 if FR2 not asked). Logged (local) medians: Flamingo 2.32 s; Qwen 6.04 s; Desta 14.53 s. API models (GPT-4o, Gemini) were not timed in this run. A full latency table is provided in Appendix D.

4.3 Cross-Linguistic Evaluation

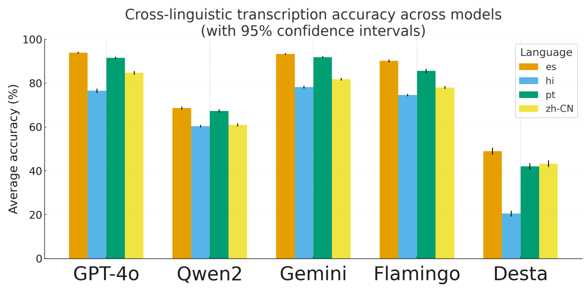

Figure 3: Cross-linguistic transcription accuracy (mean ± 95% CI) across models and languages. Languages: Spanish (es), Hindi (hi), Portuguese (pt), Mandarin Chinese (zh-CN). Per-language clip counts are N_{es}=1010, N_{hi}=1034, N_{pt}=1052, $N_{zh\text{-}CN}$=884. Error bars are normal-approximation CIs over per-clip similarity scores.

We measure transcription *similarity* (0–1; reported as %) on multilingual DailyTalk clips (construction in Section 3.1.3). The 95% CIs on per-language means are very narrow (typically < 1 percentage point), so they can be hard to discern on the bars in Figure 3. For completeness, the exact CI ranges for every model × language are reported in Appendix F.

GPT–4o and Gemini trade the lead by language (Spanish/Portuguese near-ties; Hindi → Gemini; Mandarin → GPT–4o). Flamingo trails the leaders by a small margin on Spanish/Portuguese; Qwen2 is mid-pack; Desta is lowest. None of the models are at ceiling on Mandarin.

Local (open–weights) models only: Flamingo median **0.92**s (p90 1.23s), Qwen2 **1.41**s (p90 1.92s), Desta **4.93**s (p90 10.13s). API models (GPT–4o, Gemini) were not timed in this run.

4.4 Vocal Characterizers

Non-speech vocalizations (cough, cry, laugh, sneeze, yawn) test recognition of acoustic form without relying on linguistic content. We evaluate **five-way multiple choice** on **4,006** clips (632 cough, 1,791 cry, 1,133 laugh, 236 sneeze, 214 yawn). We report **mean ± 95% CI** to keep tables compact. For a simple chance reference, the five-way baseline is 20%.

Model	Cough	Cry	Laugh	Sneeze	Yawn
GPT-4o	91.5 ± 2.2	72.1 ± 2.1	87.6 ± 1.9	71.6 ± 5.8	87.9 ± 4.4
Gemini	48.1 ± 3.9	65.2 ± 2.2	65.3 ± 2.8	14.0 ± 4.5	7.5 ± 3.6
Qwen2	74.1 ± 3.4	69.0 ± 2.2	97.2 ± 1.0	79.2 ± 5.2	86.4 ± 4.6
Flamingo	95.6 ± 1.6	90.7 ± 1.4	98.0 ± 0.8	94.9 ± 2.9	0.0 ± 0.9
Desta	26.3 ± 3.4	11.2 ± 1.5	64.1 ± 2.8	3.4 ± 2.4	7.0 ± 3.5

Table 4: Vocal characterizer accuracy (percent). Entries are mean ± 95% CI.

Effective latency sums description + MC times (local models only): Flamingo **0.80**s (p90 0.97), Qwen2 **1.22**s (p90 1.74), Desta **6.84**s (p90 62.52). API models (GPT–4o, Gemini) were not timed in this run. Flamingo leads on cough/cry/laugh/sneeze but fails on yawn; GPT–4o and Qwen2 are strong across all five; Gemini is mixed and below chance on sneeze/yawn; Desta trails. All models exceed the 20% chance level overall.

4.5 Error Analysis

Beyond aggregate accuracy, we perform a structured error analysis to understand how models fail on each task. For each task, we drew a stratified random sample of error cases across models (10 clips per model per task; $N{=}200$ total). We then developed an error taxonomy on this set and labeled the full set of samples using this schema. Our labels include (i) *omission* (the target event is never mentioned), (ii) *over-general labels* (e.g., "traffic" instead of a specific ESC-50 class), (iii) *misattribution* (wrong specific class), (iv) *direction swap* (approach↔recede), (v) *dropping language* (dropping non-English spans), (vi) *partial re-translation* (translating only the foreign span), and (vii) *noise override* (focusing on speech content of the audio snippet and ignoring the background cues). Error types are not mutually exclusive.

Error type	Bkgd. Sound	Noise Loc.	Cross Ling.	Vocal chars.
Omission	18%	43%	–	1%
Over-general label	4%	–	–	–
Misattribution	3%	21%	–	94%
Direction swap	–	8%	–	–
Dropping language	–	–	64%	–
Partial re-translation	–	–	36%	–
Noise override	75%	–	–	–

Table 5: Distribution of error types by task (percent of error cases in the aggregated dataset). Entries are percentages conditioned on an error in that task, not on all clips. Cells marked "–" denote that the error category is not applicable to that task.

The dominant failure for `background sound` is *noise override*: in 75% of wrong free responses, models transcribe the speech but never mention the background event. Plain omissions account for 18% of errors, over-general labels (e.g., "background noise," "traffic") for 4%, and misattributions for 3%.

For `noise localization`, the main issues are omission (43%; no motion described), *oscillation collapse* (25%; periodic loud–soft patterns reduced to "volume changes"), and misattributed motion type (21%). Direction swaps (approach↔recede) contribute 8% and short-horizon listening 2%.

In `cross-lingual` transcription, 64% of errors are monolingual normalization, where non-English spans are dropped and replaced with fluent English, and 36% are partial re-translation, where only the foreign fragment is translated despite instructions to preserve code-mixing.

For `vocal characterizers`, 94% of errors are misattributions between non-speech categories (e.g., yawn vs. sigh), with the mention of vocal characterizers outside of our five labels at 5% and omissions at about 1%.

Taken together, these patterns show a bias toward foreground speech over background events, limited temporal reasoning for motion, normalization of multilingual spans, and unstable labeling of non-speech vocal cues.

5 Discussion

SCENEBench assesses a simple question: if LALMs truly understand audio, can they detect salient background events, reason about motion, preserve multilingual spans, and recognize non-speech vocalizations in realistic settings? Our results show that the answer is, at best, mixed.

Across tasks, models approach or surpass chance on most multiple-choice formats but routinely miss the phenomena that matter in assistive and industrial use cases. In `background sound understanding`, omission dominates: models almost never volunteer background events in free text and only partially recover when pushed with targeted questions or options. In `noise localization`, models show substantial gains when we explicitly ask about motion, but still collapse oscillatory envelopes and over-weight the end of the clip. In `cross-linguistic transcription`, the most common error is to normalize away the non-English spans entirely. In `vocal characterizers`, lexical content overrides paralinguistic cues even when the non-lexical event (e.g., a yawn) is the label of interest.

These patterns suggest that current LALMs are optimized for *what is said* (ASR and captioning) rather than *how it is said* or *what else is happening* in the scene. This is not surprising: most public audio benchmarks focus on clean speech, single-source environmental clips, or music classification, and are designed around tasks that are easy to grade automatically. As a result, it is possible for a model to score well on existing suites while still failing on the basic building blocks of audio understanding.

So why have these gaps been overlooked? Our analysis suggests several structural reasons why these capabilities have received less attention. First, collecting and annotating overlapping events (speech with background, motion, and paralinguistic cues) is much harder than curating clean, single-label clips; therefore, most widely used corpora simply do not contain them. Additionally, common metrics such as word-error rate, caption BLEU, or single-label accuracy reward lexical fidelity and generic scene tags, but do not penalize models for dropping sirens, collapsing motion, or normalizing away multilingual spans. SCENEBench is designed around exactly these inconvenientc̈ases, not to replace existing suites, but to make it harder to claim comprehensive audio understanding without addressing them.

To improve performance on these tasks, focus on targeted data and training objectives. For background sound understanding one method of improving existing models is exposing models to speech–noise mixtures with descriptions of the background sound. Specifically, we could include information regarding (i) the presence of a background class and (ii) name it, providing the model with hard negatives where the foreground narrative is correct but the background label is wrong, since LALMs perform well at ASR (Chu et al., 2024; Wang et al., 2025; Yang et al., 2024). For localization, train on clips with known object movement; add objectives that classify approach/recede/oscillate, and ensure that the model integrates evidence over time rather than over-weighting the final seconds. For multilingual spans, instruction- and preference-tuning on span-annotated transcripts, using contrastive pairs where the only difference is whether to keep or translate the foreign segment. For paralinguistics, fine-tune or introduce data that includes short non-lexical events embedded in speech. Across tasks, use multi-task fine-tuning that combines these objectives with standard ASR/captioning, ensuring lexical fidelity while the model also learns to attend to the features of audio.

To track whether these targeted changes actually help in practice, our benchmark provides the corresponding evaluation signals. In summary, the suite identifies where current models excel and where they falter, providing a concise, reproducible testbed to guide training and model design toward parsing audio, not just speech.

5.1 Future Work

To enhance coverage, there are three improvements that can be prioritized in future work:

- **Natural code-switching data.** Replace the synthetically generated data in our third task with a small, human-recorded corpus across multiple language pairs to validate (or revise) conclusions from the synthetic set and better reflect real switching behavior.
- **Realistic acoustics.** Move beyond equal-loudness overlays and synthetic motion by (i) sweeping speech–background SNRs (e.g., -10 to $+10$ dB) and (ii) adding field recordings with moving sources

(sirens/vehicles/machinery) to test approach/recede under Doppler, occlusion, and moving-listener cases.

- **Stronger baselines.** Include non–LALM pipelines to contextualize results (e.g., speech separation $\rightarrow$ ESC–50 classifier for `background sound understanding`); this sets a clear "what classical methods achieve" line for robustness comparisons and allows us to indicate whether results indicate task difficulty or model limitations.

6 Conclusion

Overall, **SCENEBench** surfaces concrete failure modes—omission, over-general labels, misattribution, short-horizon listening, and language-prior dominance—that are obscured in cleaner, single-event benchmarks. We provide task-appropriate chance baselines, report tight confidence intervals enabled by large N, and keep full statistics in Appendix B. Our hope is that these evaluations help steer model development toward the capabilities demanded in accessibility and industrial safety.

7 Limitations

Our study has several limitations. First, the mixture design relies on controlled levels of speech–noise overlays. While equal-level mixing provides experimental control, it may not reflect signal-to-noise ratio distributions encountered in the wild. Second, some of the upstream corpora contain weak or imperfect annotations. Residual label errors can influence ceiling estimates and increase per-class confusions. Third, closed-model APIs restrict fine-grained latency profiling and ablation studies. As a result, we report timer latency measures wherever possible. Finally, the multilingual speech setup, which uses a TTS pipeline with back-translation filtering, improves consistency but does not fully capture the spontaneity of real human speech.

8 Ethical Considerations

We also highlight several ethical considerations. To reduce risks associated with reductive "emotion AI," we avoid direct emotion inference and instead focus on paralinguistic events. All speech data, including synthetic mixtures, must respect licensing constraints, and any human recordings require explicit consent and data minimization. Accessibility risks must be considered carefully: in safety-critical contexts like siren detection, both missed detections and false alarms can carry distinct harms. Systems should therefore expose calibrated uncertainty and provide fallback behaviors. Finally, fairness is essential. Benchmarks should broaden their coverage of accents, languages, and recording conditions to reduce disparate error rates across user groups.

References

William Agnew, Julia Barnett, Annie Chu, Rachel Hong, Michael Feffer, Robin Netzorg, Harry H. Jiang, Ezra Awumey, and Sauvik Das. 2024a. Sound check: Auditing audio datasets. *arXiv preprint arXiv:2410.13114*.

William Agnew, Julia Barnett, Annie Chu, Rachel Hong, Michael Feffer, Robin Netzorg, Harry H. Jiang, Ezra Awumey, and Sauvik Das. 2024b. Sound-Check: Auditing audio datasets. *arXiv preprint arXiv:2410.13114*.

Alibaba Cloud. 2024. Alibaba cloud launches qwen2-audio model to analyze speech and audio. Accessed: 2025-10-06.

Danielle Bragg, Naomi Huynh, and Richard E. Ladner. 2016. A personalizable mobile sound detector app design for deaf and hard-of-hearing users. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, pages 3–13.

Changan Chen, Unnat Jain, Carl Schissler, and 1 others. 2020. Audiogoal: Audio-visual embodied exploration via semantic audio-visual graph. In *Proc. NeurIPS*.

Chiun-Li Chin, Chia-Chun Lin, Jing-Wen Wang, Wei-Cheng Chin, Yu-Hsiang Chen, Sheng-Wen Chen, and Po-Cheng Hsieh. 2023. A wearable assistant device for the hearing impaired to recognize emergency vehicle sirens with edge computing. *Sensors*, 23(17):7454.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Zhifang Guo, and 1 others. 2024. Qwen2-Audio Technical Report. *arXiv preprint*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, and 1 others. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next-Generation Agentic Capabilities. *arXiv preprint*.

Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. 2020. Librimix: An open-source dataset for generalizable speech separation. *arXiv preprint arXiv:2005.11262*. Includes Libri2Mix and Libri3Mix variants.

Cem Dilmegani. 2025. Top 4 speech recognition challenges & solutions in 2025. https://research.aimultiple.com/speech-recognition-challenges/. Accessed: 2025-07-24.

Kota Dohi and 1 others. 2022. Mimii-dg: Sound dataset for domain generalization in industrial machine anomaly detection. In *Proc. DCASE Workshop*. ArXiv:2206.05828.

Leah Findlater, Bonnie Chinh, Dhruv Jain, Jon E. Froehlich, Raja Kushalnagar, and Angela Carey Lin. 2019. Deaf and hard-of-hearing individuals' preferences for wearable and mobile sound awareness technologies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, pages 1–13, New York, NY, USA. ACM.

Jort F. Gemmeke, Daniel P. W. Ellis, and 1 others. 2017. Audioset: An ontology and human-labeled dataset for audio events. In *Proc. ICASSP*.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

Andrew Gordon, Shurjo Banerjee Majumder, and 1 others. 2020. Soundspaces: Audio-visual navigation in 3d environments. In *Proc. ECCV*.

Will Held, Michael J. Ryan, Diyi Yang, and 1 others. 2025. Comprehensive assessment for voice assistants (cava). Online. Accessed: 24 Jul 2025.

John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In *Advances in Neural Information Processing Systems (NeurIPS)*. Introduces the WSJ0-2mix evaluation setup for single-channel speech separation.

Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey. 2016. Single-channel multi-speaker separation using deep clustering. In *Proc. Interspeech*.

Dhruv Jain, Leah Findlater, and Jon E. Froehlich. 2019. Exploring sound awareness in the home for people who are deaf or hard of hearing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 94:1–94:13.

Prateek Jain, Vaibhav Joshi, Pranay Pachpute, and Ruhi Sarma. 2024. SONAR: A synthetic ai-audio detection framework and benchmark. *arXiv preprint arXiv:2407.10101*.

JooYeong Kim, Sooyeon Ahn, and Jin-Hyuk Hong. 2023. Visible nuances: A caption system to visualize paralinguistic speech cues for deaf and hard-of-hearing individuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, New York, NY, USA. ACM.

Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto. 2019. Toyadmos: A dataset of miniature-machine operating sounds for anomalous sound detection. In *Proc. IEEE WASPAA*.

Kalyan Krishnakalyan. 2023. Vocal bursts taxonomy 100 clean (wds). https://huggingface.co/datasets/krishnakalyan3/vocal_bursts_taxonomy_100_clean_wds. Accessed: 2025-09-06.

Ji-Hoon Lee, Arthur Caillon, Juan-Manuel Perez, Keisuke Koga, and Kewei Liu. 2024. AIR-Bench: Benchmarking large audio–language models via generative comprehension. *arXiv preprint arXiv:2404.12345*.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *Proc. IEEE ICASSP*.

Xiang Li, Pin-Yu Chen, and Wenqi Wei. 2024. Sonar: A synthetic ai-audio detection framework and benchmark. *arXiv preprint arXiv:2410.04324*.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. *arXiv preprint arXiv:2204.09634*.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2024. Developing instruction-following speech language model without speech instruction-tuning data. *arXiv preprint arXiv:2409.20007*.

Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, and 1 others. 2017. Dcase 2017 challenge task 2: Detection of rare sound events in real-life audio. In *Proc. DCASE Workshop*.

Annamaria Mesaros, Toni Heittola, Aleksandr Diment, and 1 others. 2018. Dcase 2018 challenge task 5: Monitoring domestic activities based on multichannel acoustics. In *Proc. DCASE Workshop*.

nyuuzyou. 2022. Asmr whisper audio collection. https://huggingface.co/datasets/nyuuzyou/asmr. Hugging Face dataset; Accessed: 2025-10-06.

OpenAI. 2024. GPT-4o System Card. https://cdn.openai.com/gpt-4o-system-card.pdf. Accessed 24 July 2025.

OpenSound. 2024. Capspeech-agentdb-audio: Mumble and whisper audio clips. https://huggingface.co/datasets/OpenSound/CapSpeech-AgentDB-Audio. Hugging Face dataset; Accessed: 2025-10-06.

Lara Orlandic, Tomislav Teijeiro, and David Atienza. 2021. The coughvid crowdsourcing dataset: A corpus for covid-19 cough classification. *Scientific Data*, 8(1):156.

Karol J. Piczak. 2015. ESC-50: Dataset for environmental sound classification. In *Proc. ACM MM*, pages 1015–1018.

Harsh Purohit, Ryo Tanabe, Kohei Ichige, Tomoya Nishida, Takashi Endo, Yuma Koizumi, Noboru Harada, Masahiro Yasuda, and Satoru Tamura. 2019. Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In *Proc. DCASE Workshop*. ArXiv:1909.09347.

Sakshi S., Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. MMAU: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Osman Salem, Ahmed Mehaoua, and Raouf Boutaba. 2023. The sight for hearing: An iot-based system to assist drivers with hearing disability. In *2023 IEEE Symposium on Computers and Communications (ISCC)*. IEEE. Available at`https://rboutaba.cs.uwaterloo.ca/Papers/Conferences/2023/SalemIoT2023.pdf`.

Björn Schuller, Stefan Steidl, Anton Batliner, and 1 others. 2011. The interspeech 2011 computational paralinguistics challenge: Intoxication, sleepiness, and age. In *Proc. INTERSPEECH*.

Nishad Sharma, P. Krishnan, and 1 others. 2020. Coswara: A database of breathing, cough, and voice sounds for covid-19 diagnosis. ArXiv:2005.10548.

Luke Stark and Jevin D. Hoey. 2021. The ethics of emotion in ai systems. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–34. Discusses risks of reductive emotion recognition and attendant harms.

W4ng1204. 2023. Nonspeech7k: Non-speech vocalization dataset (cough, cry, laugh, sneeze, yawn). `https://huggingface.co/datasets/W4ng1204/Nonspeech7k`. Hugging Face dataset; Accessed: 2025-10-06.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, Ai Ti Aw, and Nancy F. Chen. 2024. AudioBench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*. Last revised May 2025.

Liang-Yuan Wu and Dhruv Jain. 2025. Soundnarratives: Rich auditory scene descriptions to support deaf and hard of hearing people. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25)*, New York, NY, USA. ACM. Appears at ASSETS 2025; preprint available.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*. Accepted to ACL 2024.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, Ningzhi Wang, Chenghua Lin, Emmanouil Benetos, Anton Ragni, Norbert Gyenge, Roger Dannenberg, Wenhu Chen, Gus Xia, and 6 others. 2023. Marble: Music audio representation benchmark for universal evaluation. *arXiv preprint arXiv:2306.10548*. Camera-ready version for NeurIPS 2023.

A Background Sound Task Prompts

FR1 (Free-response description).

```
Please describe what you hear in this
audio file. Be as detailed as possible
about all the sounds you can identify.
```

FR2 (Targeted follow-up; asked only if FR1 omitted the background/ambient sound).

```
In your previous response, you didn't
mention the background noise. Please
specifically describe what background
noise or ambient sound you hear in this
audio.
```

MC (4-way multiple choice; administered for all clips).

```
You will be given an audio clip. What is
the background sound? Choose the correct
answer from the following options and
reply with ONLY the number (e.g., 1, 2,
3, or 4).
```

B Statistical Methods for Background Sound Task

We compute model-wise 95% CIs for proportions using the Wilson interval. FR1 (Any noise) — GPT–4o: 6.5% [5.73, 7.27], Qwen2: 29.3% [27.89, 30.71], Gemini: 8.6% [7.73, 9.47], Flamingo: 20.6% [19.35, 21.85], Desta: 2.9% [2.38, 3.42]. FR2 (Specific type*) — GPT–4o: 8.3% [7.44, 9.16], Qwen2: 30.1% [28.68, 31.52], Gemini: 10.9% [9.93, 11.87], Flamingo: 32.8% [31.35, 34.25], Desta: 6.4% [5.64, 7.16]. MC — GPT–4o:

8.0% [7.16, 8.84], Qwen2: 60.2% [58.68, 61.72], Gemini: 30.8% [29.37, 32.23], Flamingo: 74.2% [72.84, 75.56], Desta: 24.9% [23.56, 26.24].

For each model and the multiple choice question (MC1), we test H$_0$: $p = p_{\text{chance}}$ with a one-sample proportion z-test: 0.25 (MC1; 1/4). To control multiplicity across 15 tests (5 models $\times$ 3 metrics), we apply Benjamini–Hochberg FDR at α=0.05. Effect sizes are reported as absolute risk difference Δ=$p - p_{\text{chance}}$ and Cohen's $h = 2\arcsin\sqrt{p} - 2\arcsin\sqrt{p_{\text{chance}}}$. Sample size and power notes: with N=4000 clips, the minimum detectable $|\Delta|$ at 0.8 power is <2 percentage points for MC1.

Multiple choice (MC; baseline 25%). Flamingo **74.2%**, Qwen2 60.2%, Gemini 30.8%, Desta 24.9%, GPT–4o 8.0%. Flamingo, Qwen2, and Gemini are **above chance** (BH–adj. p<10^{-10}); Desta is **not different** from chance (BH–adj. p≈0.83); GPT–4o is **below chance** (BH–adj. p<10^{-10}).

C Noise Localization Task Prompts

FR1 (Free-response description).

```
Please describe what you hear in this
audio file. Be as detailed as possible
about all the sounds you can identify,
including any changes in volume or
spatial characteristics.
```

FR2 (Targeted follow-up; asked only if FR1 omitted the direction of noise).

```
How does our position change in relation
to the sound source in this audio?
Describe the spatial relationship and
any movement patterns you detect.
```

D Noise Localization Statistics and Latency

Statistical testing. For each model and question, we test H_0: $p = \frac{1}{3}$ (33.3%) using a one-sample proportion z-test; p-values are Benjamini–Hochberg–adjusted across 10 tests (5 models $\times$ 2 questions). All models are either well below chance on FR1 or well above chance on FR2, and all baseline tests are significant (BH-adjusted $p < 10^{-4}$). We then run all pairwise model comparisons (two-proportion tests, BH-adjusted within question) and report the full set: on *FR1* (general), overall accuracies are low (6–16%), but nearly all pairs differ significantly; the *only* non-significant comparison is Flamingo3 vs. Qwen2 ($\Delta = +0.57$ points, z=1.16, p_{BH}=0.245), while all others fall in the range $\Delta \in [1.4,, 9.7]$ points with $p_{\text{BH}} <$ 0.01, and the largest gap is Gemini vs. DESTA2 (Δ≈10.32 points, z=-17.92, $p_{\text{BH}} < 10^{-15}$). On *FR2* (position), every pairwise difference is large and significant; e.g., Qwen2 vs. Flamingo3 (Δ≈41.2 points, z=-47.06, $p_{\text{BH}} \approx 0$), Qwen2 vs. GPT-4o (Δ≈10.72 points, z=11.74, $p_{\text{BH}} \approx 0$), and even the smallest gap, DESTA2 vs. Flamingo3 (Δ≈3.32 points, z=4.85, $p_{\text{BH}} \approx 1.24\times10^{-6}$), remains significant. Effect sizes are reported as absolute differences Δ=$p - \frac{1}{3}$ and Cohen's h in the supplement.

Per-class notes. Direction accuracy concentrates on *Approaching/Receding*; *Oscillating* is near floor (e.g., ≤10.8%).

Latency table.

Model (local)	Median (s)	Mean (s)	p90 (s)
Flamingo	2.32	2.35	2.98
Qwen	6.04	6.54	10.25
Desta	14.53	27.96	72.16
GPT-4o	n/a	n/a	n/a
Gemini	n/a	n/a	n/a

Table 6: Motion-task effective latency (local inference). Values are per-clip medians, means, and 90th-percentiles in seconds; API models (GPT-4o, Gemini) are not available in this run.

E Cross-Linguistic Task Prompt

FR1 (Free-response description).

```
Transcribe the following audio,
preserving any code-mixing or
multilingual content. If the audio
contains both English and other
languages, keep the code-mixed style in
your transcript. Output the transcript
exactly as spoken, including any
non-English words or phrases.
```

F Cross-Linguistic Details

Statistics. We compute 95% CIs over clip-level similarity via normal approximation; for between-model comparisons per language we use two-sample t-tests on per-clip scores with Benjamini–Hochberg correction across model pairs. Effect sizes are reported as mean differences (pp) with 95% CIs.

Language (N)	Model	Mean $\pm$ 95% CI (%)
Spanish (es, $N = 1010$)	GPT–4o	**93.9 [93.4, 94.4]**
	Gemini	93.3 [92.9, 93.7]
	Flamingo	90.1 [89.5, 90.7]
	Qwen2	68.7 [68.0, 69.4]
	Desta	49.0 [47.4, 50.5]
Hindi (hi, $N = 1034$)	Gemini	**78.2 [77.5, 78.8]**
	GPT–4o	76.5 [75.6, 77.5]
	Flamingo	74.6 [74.0, 75.1]
	Qwen2	60.3 [59.8, 60.9]
	Desta	20.5 [19.1, 21.8]
Portuguese (pt, $N = 1052$)	Gemini	**91.8 [91.4, 92.3]**
	GPT–4o	91.5 [91.0, 92.0]
	Flamingo	85.6 [84.7, 86.4]
	Qwen2	67.3 [66.6, 68.1]
	Desta	42.1 [40.6, 43.6]
Mandarin (zh–CN, $N = 884$)	GPT–4o	**84.7 [83.9, 85.4]**
	Gemini	81.8 [81.2, 82.4]
	Flamingo	78.0 [77.3, 78.7]
	Qwen2	61.0 [60.2, 61.7]
	Desta	43.2 [41.8, 44.7]

Table 7: Cross–linguistic transcription (mean similarity $\pm$ 95% CI, %). All models are shown for each language; the best mean per language is bolded.

G Vocal Characterizers Task Prompt

MC1.

```
Which of the following best describes
this sound?  (A) cough (B) cry (C)
laugh (D) sneeze (E) yawn (F) mumble (G)
whisper. Answer with the letter and the
word.
```