GlobeAudio: A Multilingual Multicultural Benchmark for Naturalistic Evaluation of Large Audio-Language Models

Ryner Tan, Wenxuan Zhang
iNLP Lab, Singapore University of Technology and Design
ryner_tan@mymail.sutd.edu.sg, wxzhang@sutd.edu.sg

Abstract

Large Audio-Language Models (LALMs) integrate audio perception and language understanding within a unified framework, enabling a wide range of real-world applications. Despite recent advances, evaluation for LALMs remains heavily underspecified relative to real-world requirements: most lack true linguistic and cultural authenticity, while others fail to capture acoustic realism. To bridge this gap, we propose GLOBEAUDIO, a multilingual and multicultural benchmark designed to evaluate naturalistic audio understanding. GLOBEAUDIO consists of 5,637 multiple-choice questions across six typologically diverse languages, expertly crafted by native speakers grounded on naturally occurring audio. In order to do well, models must possess higher-level auditory reasoning skills and culturally grounded interpretation. We systematically evaluate representative closed-source and open-source LALMs, as well as cascaded ASR–LLM pipelines. Our experiments reveal substantial performance gaps under natural acoustic conditions, particularly for open-source models and low-resource languages. These findings highlight critical limitations of current LALMs and underscore the importance of naturalistic audio evaluation for future audio-language systems. GLOBEAUDIO can be found at https://huggingface.co/datasets/iNLP-Lab/GlobeAudio.

1 Introduction

Large Audio-Language Models (LALMs) are an emerging class of systems designed to comprehend, process, and generate responses directly from audio inputs (Lakhotia et al., 2021; Chu et al., 2024; Fang et al., 2025; Tang et al., 2024; Liu et al., 2025a; OpenAI et al., 2024; Liu et al., 2025b). Unlike traditional speech processing models, which typically specialize in individual tasks such as automatic speech recognition (ASR) (Radford et al., 2022; Sekoyan et al., 2025; King et al., 2025; Shakhadri et al., 2025), LALMs unify audio processing and language understanding, enabling tasks like spoken QA and speech translation for diverse applications such as virtual assistants and educational tools (Su et al., 2025; Deshmukh et al., 2023). Driven by their rapid adoption, the community has introduced evaluation suites ranging from those that consolidate and standardise traditional audio tasks (Wang et al., 2025; Heigold et al., 2025) to broader benchmarks assessing multimodal reasoning across heterogeneous inputs (Sakshi et al., 2024; Yang et al., 2024).

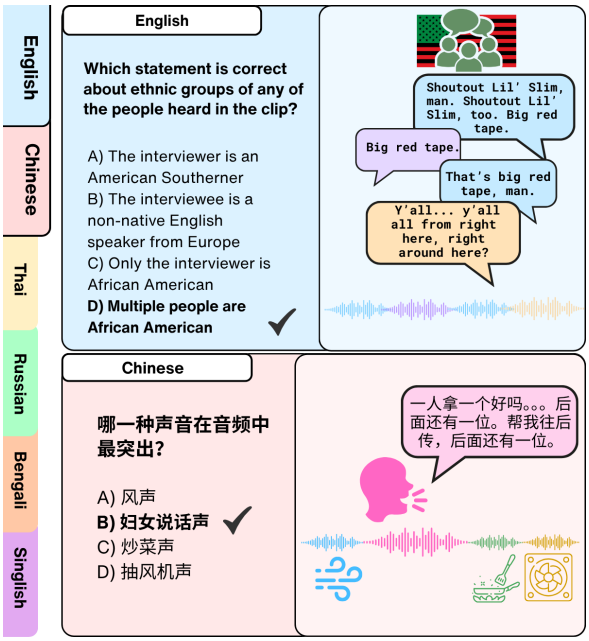

Figure 1: Examples of GLOBEAUDIO, which require models to first unpack compound auditory tasks through multiple reasoning steps.

Despite these advances, evaluation for LALMs remains vastly underspecified in several important respects relative to real-world deployment requirements. Concretely, they remain predominantly English-centric and lack true linguistic and cultural authenticity. Though some multilingual bench-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 32924–32937
October 24-29, 2026 ©2026 Association for Computational Linguistics

Dataset	Multilingual	Multicultural	Natural / Assesses Prosody	Compound Tasks
AudioBench (Wang et al., 2025)	✗	✗	✓	✗
MSEB (Heigold et al., 2025)	✓	✓	✗	✗
MMAU (Sakshi et al., 2024)	✗	✗	✓	✓
AIR-Bench (Yang et al., 2024)	✗	✗	✗	✓
Fleurs-SLU (Schmidt et al., 2025)	✓	✗	✗	✗
mSTEB (Beyene et al., 2025)	✓	✗	✗	✗
AudioMarathon (He et al., 2025)	✗	✗	✗	✓
SSEU-Bench (Yin and Choi, 2025)	✗	✗	✗	✓
GlobeAudio	✔	✔	✔	✔

Table 1: Dataset comparison of GlobeAudio with other audio benchmarks in modelling real-world scenarios.

marks (Schmidt et al., 2025; Beyene et al., 2025) are scalable and largely broaden language coverage, their reliance on automated translation compromises multicultural validity by flattening culturally grounded references.

In addition, current evaluations fail to capture acoustic realism by focusing primarily on lexical content (Sakshi et al., 2024; Schmidt et al., 2025; Beyene et al., 2025; Heigold et al., 2025) thus largely overlooking prosodic cues critical to real-world communication, or relying on semi-scripted speech (He et al., 2025; Yin and Choi, 2025; Lee et al., 2024; Conneau et al., 2022; Huang et al., 2026) which fails to capture the acoustic and sociolinguistic variability of real-world contexts. Consequently, the scope of LALM evaluation remains reduced and constrained to decomposed, atomic tasks (Schmidt et al., 2025; Heigold et al., 2025), ignoring higher-level pragmatics and cultural context inherent to the real-world.

To bridge these critical gaps summarised in Table 1, we introduce GLOBEAUDIO, a multilingual and multicultural benchmark for naturalistic audio understanding. We capture the wide sociolinguistic and acoustic variability of real-world communication by curating unscripted, naturally occurring audio clips directly from online media across diverse contexts. Consequently, rather than relying on automated translations, native speakers deeply rooted in local culture authored all multiple-choice questions with carefully designed distractors that require fine-grained audio and cultural understanding beyond transcription. Finally, we ensure the highest data fidelity by employing a rigorous two-stage cross-review and adjudication process to resolve ambiguities, yielding an exceptionally high inter-annotator consensus at **95.5%**. In total, GLOBEAUDIO covers six languages spanning high to low-resource settings and diverse language families, offering a total of 5,637 expertly crafted multiple-choice questions in total as shown in Figure 1.

Consequently, we conduct a comprehensive evaluation of representative LALMs as well as cascaded pipelines (ASR $\rightarrow$ text LLM) under controlled settings. Our results reveal significant gaps in their compound auditory reasoning capabilities, demonstrating that naturalistic audio conditions remain a major blind spot for state-of-the-art models. We discover that this gap is especially severe in open-source models which trail considerably behind their closed-source counterparts, and in low-resource languages such as Thai and Bengali. Through targeted ablations and extensive analysis, we further demonstrate the high fidelity of GLOBEAUDIO and reveal a surprising alignment behaviour: LALMs perform better when evaluated with questions and transcripts in the source language, instead of in English which is a comparatively higher resource language.

In summary, our contributions are threefold:

- We introduce GLOBEAUDIO, a multilingual, multicultural benchmark for naturalistic audio understanding, comprising 5,637 human-authored and rigorously verified MCQs across 6 languages and diverse real-world contexts.
- We extensively benchmark various state-of-the-art LALMs, spanning closed-source, open-source models across various model families, and reveal substantial performance gaps, particularly for open-source models and low-resource languages.
- Through targeted ablations and cross-lingual studies, we uncover a counter-intuitive but consistent advantage of source-language evaluation over English translation, and the linguistic–cultural grounding in shaping LALM performance.

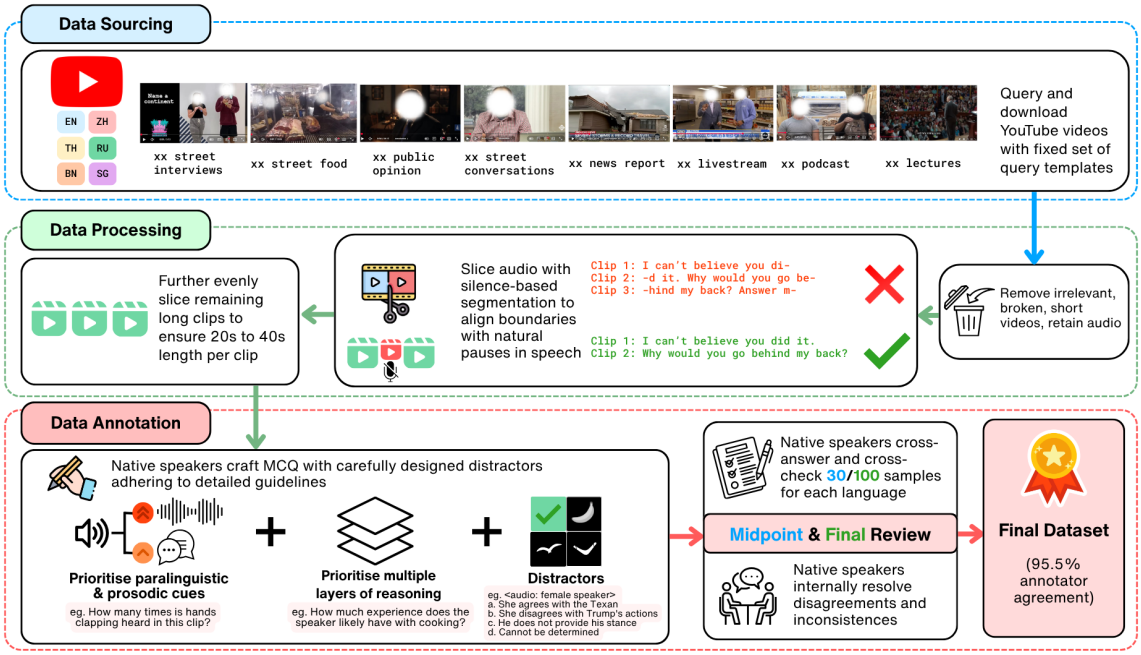

Figure 2: GLOBEAUDIO's data preparation pipeline, from curation to rigorous annotation and iterative cross-annotator quality checks, culminating in a high-fidelity dataset.

2 GLOBEAUDIO Dataset

2.1 Design Considerations

To provide a realistic reflection of real-world performance when LALMs are exposed to the acoustic and sociolinguistic conditions encountered *in-the-Wild*, we design GLOBEAUDIO with three core principles:

- *Multilingual and Multicultural Coverage*: Real-world speech understanding is inherently shaped by linguistic diversity and culturally grounded usage (Alam et al., 2025). To ensure that evaluation extends beyond English-centric or culturally homogenized settings, GLOBEAUDIO is constructed by native speakers with lived experience in the target languages and cultures.
- *In-the-Wild Audio Sources*: Spoken language in real-world environments is often characterized by background noise, prosody, overlapping speakers, informal delivery, and unconstrained recording conditions (Nagrani et al., 2020). To capture this variability, GLOBEAUDIO is built from naturally occurring audio clips drawn from online media spanning both formal and informal contexts. This design choice intentionally departs from clean or semi-scripted speech, exposing models to the realistic acoustic complexity that deployed LALMs must routinely handle.
- *Compound Task Design*: Naturalistic audio understanding demands higher-level skills like discourse tracking, pragmatic inference and culturally grounded interpretation. To this end, GLOBEAUDIO adopts compound question formulations where answering each question typically requires several layers of reasoning over the audio content, discouraging reliance on superficial cues or isolated subtasks, but instead probing holistic auditory comprehension.

2.2 Language Selection and Considerations

Guided by the above design principles, we aim to construct an evaluation set that meaningfully probes both linguistic and cultural variability in real-world speech understanding. Language selection is therefore driven by a combination of complementary criteria and was designed to maximise geographical coverage, resource availability, diversity in writing systems, and diversity in phonological characteristics.

Based on these considerations, we select six languages: **English** (United States), **Chinese** (China), **Thai** (Thailand), **Russian** (Russia), **Bengali** (India) and **Singlish** (Singapore). Collectively, this set spans multiple language families (Indo-European,

		Linguistic Properties			Dataset Statistics		
Language (Code)	**CC Size**	**Isochrony**	**Genus**	**Script**	**Clip length**	**QA length (Q/A)**	**Total**
English (en)	42.60	High, Str	Germanic	Latin	25.72	53.8 / 22.9	1,274
Russian (ru)	6.15	High, Str	Slavic	Cyrillic	25.69	37.8 / 22.2	924
Chinese (zh)	4.99	Mid, Syl	Sinitic	Hanzi	24.88	13.5 / 5.0	1,072
Thai (th)	0.37	Mid, Syl	Tai	Thai	24.82	29.3 / 13.0	1,145
Bengali (bn)	0.10	High, Syl	Indo-Aryan	Bengali	25.47	37.6 / 18.3	719
Singlish (sg)	–	Low, Syl	–	Latin	25.54	73.3 / 34.9	503
Total					**25.35**	**40.9 / 19.4**	**5,637**

Isochrony: Categorised by syllable complexity and rhythmic unit (*Str*: stress-timed, *Syl*: syllable-timed)

Table 2: Dataset Composition of GLOBEAUDIO, sorted by decreasing language resource level. We report weighted averages for clip length (in seconds) and QA length (in characters), with per-language count.

Sino-Tibetan, Tai–Kadai, and creole varieties), covers a wide range of geographic regions, and reflects substantial variation in resource availability. As shown in Table 2, where we use CommonCrawl's Size (CC Size) as a proxy for the resource level, it covers from high-resource languages such as English, to comparatively much lower-resource settings such as Thai and Bengali. The selected languages also differ markedly in script systems and spoken characteristics: the inclusion of Singlish enables the evaluation of naturally occurring code-mixing and multilingual speech within a single variety, capturing linguistic phenomena that are common in real-world audio but rarely represented in existing benchmarks.

2.3 Data Collection and Processing

We present a full, detailed visualisation of our data preparation pipeline in Figure 2, which we describe in detail below.

YouTube hosts a large and continuously updated collection of publicly accessible video content spanning a wide range of communicative contexts. We therefore source our raw data from YouTube using `yt-dlp`, an open-source audio and video downloader tool [1] on a curated set of language-specific queries designed to capture both formal and informal speech. For each language, we retrieve 15 videos with each of the following query templates: *{"xx street food", "xx street interviews", "xx street conversations", "xx public opinion", "xx news report", "xx lectures", "xx livestream", "xx podcast"}*, where x refers to the corresponding country of the selected languages.

From the retrieved videos, we discard visual content and retain only the audio streams for further processing. To obtain manageable yet contextually rich audio samples while avoiding mid-utterance truncation, we segment each full-length audio file using silence-based segmentation, aligning clip boundaries with natural pauses in speech. The silence-based audio slicer is open-sourced on GitHub.[2]. This procedure yields short audio clips with durations ranging from 20 to 40 seconds, which balances sufficient discourse context with practical annotation and evaluation constraints. Ethics and licensing considerations are reported under Section 8.

2.4 Data Annotation

Data annotation is performed exclusively by native speakers in their first language, ensuring high linguistic fidelity and culturally appropriate interpretation. For each audio clip, these annotators construct a four-option multiple-choice question (MCQ) with one correct answer and three carefully designed distractors. We adopt the MCQ format over open-ended ones since it enables scalable, unambiguous automatic evaluation while preserving difficulty. Concretely, annotators are instructed to prioritize questions that rely on information conveyed through the audio signal itself, such as and not limited to pragmatic meaning and paralinguistic cues, or culturally grounded questions. Furthermore, they are encouraged to craft ones that require multi-step reasoning over the audio, rather than direct fact retrieval. Annotators also design distractors to be semantically or acoustically confusable to discourage superficial elimination strategies. Finally, they do not annotate clips for which no meaningful question can be asked, upon our verification, as detailed in Appendix Section A.2.

To ensure high annotation quality, we conduct a two-stage quality control process consisting of a

[1] GitHub: https://github.com/yt-dlp/yt-dlp

[2] https://github.com/openvpi/audio-slicer

midpoint review and a final review. In the midpoint review, we randomly sample 30 annotated items from each annotator and perform cross-checking among native speakers within the same language to identify annotation errors and inconsistencies at an early stage. Following completion, we conduct a final review by applying stratified random sampling to select an additional 500 items (100 per language), ensuring balanced representation across annotators. These samples are independently cross-checked and aligned in the same manner as the midpoint review. Across the reviewed samples, we observe an inter-annotator consensus of **95.5%**. Disagreements are resolved through multiple rounds of discussion and corrections, resulting in a final dataset with verified and consistent gold labels as visualised in Figure 2.

2.5 Dataset Breakdown

Finally, after data processing and quality control, GLOBEAUDIO comprises a total of 5,637 multiple-choice questions spanning six languages and diverse language families. All audio clips are between 20 to 40 seconds long, with the mean clip length being 25.35s. A detailed breakdown is found in Table 2.

Figure 1 illustrates representative samples from our dataset, comprising of audio inputs and their corresponding multiple-choice questions, which require models to unpack compound auditory tasks through multiple reasoning layers. More examples can be found in Appendix Section A.3.

3 Experimental Setup

3.1 Model Selection

For comprehensive evaluation on GLOBEAUDIO, we select a diverse set of representative models spanning both open-source and closed-source LALMs, model families, architectures, and deployment trade-offs. In addition, we include cascaded Automatic Speech Recognition (ASR)–LLM pipelines as strong baselines for comparison against end-to-end LALMs.

Closed-source LALMs We consider two state-of-the-art LALMs, namely `Gemini 3.1 Pro Preview` (Google, 2026), which is a frontier multimodal model with strong cross-modal reasoning capabilities, and `GPT-4o Audio` (OpenAI et al., 2024), which achieves state-of-the-art performance in multiple audio benchmarks including translation and transcription. We also include `Gemini 2.0 Flash` (Comanici et al., 2025) as a latency and cost-optimized multimodal model, allowing for evaluation of trade-offs in real-world deployment scenarios.

Open-source LALMs We extensively evaluate seven open-source LALMs spanning a wide range of parameter scales, from smaller models like Alibaba's `Qwen2-Audio-7B-Instruct` (Chu et al., 2024) to much larger models like Xiaomi's `MiMo-V2.5` (Xiaomi, 2026). We also include `Qwen3-Omni-Flash-2025-09-15` (Xu et al., 2025), the latest high-capacity Mixture-of-Experts model which is state-of-the-art on most existing audio and audio-visual benchmarks.

Cascaded ASR–LLM models In contrast to end-to-end LALMs, cascaded models decompose audio understanding into sequential components. We evaluate two such pipelines that first transcribe audio using a Whisper-based ASR system (`whisper-1`) (Radford et al., 2022), followed by downstream reasoning with text-only input. For the second component, we select `Qwen3-235B-A22B-Instruct-2507` (Yang et al., 2025a) which has 235B parameters and `Gemini 3.1 Pro Preview` (Google, 2026), allowing for direct comparisons against their end-to-end counterparts under controlled conditions.

3.2 Settings

Inference Inference of all models is performed through API calling, with the exception of `Qwen2-Audio-7B-Instruct` which is performed on vLLM [3], an open-source high-throughput inference engine for LLMs. In the event of API timeouts or invalid responses, we perform a maximum of 3 retries to ensure reproducibility.

Prompts Across all models, we employ a unified template detailed in Appendix Section A.1. For end-to-end open- and closed-source LALMs, the audio clip is fed directly into the model. For cascaded pipelines, the ASR-generated transcript is provided as context before prompting the LLM. To avoid introducing cross-lingual confounds, all prompts are in the source language.

Evaluation Metrics As GLOBEAUDIO consists exclusively of MCQ items, model outputs are verifiable and evaluated on accuracy, comparing the predicted option label against the ground-truth.

[3] https://github.com/vllm-project/vllm

Model	Modalities	Size	Language Accuracy (%)						Avg.
			en	ru	zh	th	sg	bn	
Closed-source Models									
GPT-4o Audio (OpenAI et al., 2024)	$T, A \rightarrow T, A$	–	61.62	70.24	50.75	45.07	76.74	56.33	60.13
Gemini 2.0 Flash (Google, 2025a)	$T, A \rightarrow T$	–	64.13	77.81	55.13	60.35	74.75	68.15	66.72
Gemini 3.1 Pro (Google, 2026)	$T, A \rightarrow T$	–	**78.26**	**90.04**	**70.34**	**81.57**	**90.06**	**85.81**	**82.68**
Open-source Models									
Qwen2-Audio (Chu et al., 2024)	$T, A \rightarrow T$	7B	48.04	55.30	40.95	29.00	54.67	25.87	42.31
MERaLiON 2 (MERaLiON Team, 2024)	$T, A \rightarrow T$	10B	53.38	60.82	48.79	45.15	69.38	56.33	55.64
Audio Flamingo 3 (Goel et al., 2025)	$T, A \rightarrow T$	7B	64.99	71.65	52.38	46.81	73.36	28.93	56.35
Gemma 3n (Google, 2025b)	$T, A \rightarrow T$	8B	55.49	64.83	45.62	48.82	65.61	58.83	56.53
Voxtral (Liu et al., 2025a)	$T, A \rightarrow T$	24B	59.65	68.51	51.77	46.99	75.35	45.34	57.94
Qwen3-Omni (Xu et al., 2025)	$T, A \rightarrow T, A$	30B	71.35	82.25	63.34	65.15	80.32	70.24	72.11
MiMo-V2.5 (Xiaomi, 2026)	$T, A \rightarrow T$	310B	65.54	68.51	57.60	43.23	75.94	55.63	61.08
Cascade Systems									
Whisper + Qwen 3	$A \rightarrow T \rightarrow T$	1.5B + 30B	65.23	78.03	51.40	60.52	74.55	58.41	64.69
Whisper + Gemini 3.1 Pro	$A \rightarrow T \rightarrow T$	1.5B + –	69.78	80.84	56.81	71.79	82.50	70.51	72.04

Table 3: Main results: per-language accuracy (%) and macro-averages for closed, open-source, and cascaded systems, sorted chronologically by release date. Relevant modalities, T: Text and A: Audio are also specified.

4 Results and Analysis

4.1 Main Results

We present a comprehensive performance breakdown of all models on GLOBEAUDIO in Table 3.

Evaluation on GLOBEAUDIO reveals severe performance disparities across model families. `Gemini 3.1 Pro Preview` significantly outperforms the field at 82.68% accuracy, whereas competing LALMs cluster at an average of 58.76% and rarely surpass 70%. This stark divide underscores a systemic vulnerability in current models: while they excel on structured benchmarks, they perform poorly when met with uncontrolled acoustic variability and compound reasoning demands inherent to naturalistic audio.

Overall, closed-source models consistently demonstrate stronger performance than open-source counterparts. Notably, `Gemini 2.0 Flash` outperforms most open-source LALMs despite being optimized for latency and cost. This persistent gap underscores the need for continued community efforts to advance open and reproducible LALMs.

Comparing all open-source LALMs, `Qwen3-Omni-Flash` substantially outperforms with an average accuracy of 72.11%. This disparity may be influenced by its larger effective model capacity, as well as architectural choices such as its Mixture-of-Experts (MoE) design. Notably, the stronger closed-source models in our evaluation belong to the Gemini family, and also employ MoE-style architectures which suggests that expert routing may be beneficial for handling heterogeneous linguistic and acoustic conditions. This is further validated by `MiMo-V2.5`'s strong performance, which also employs such an architecture.

Performance further varies across languages. Higher-resource languages such as English and Russian generally yield stronger results across models, while low-resource languages such as Thai and Bengali remain more challenging, particularly for open-source LALMs and cascaded systems. Singlish, despite being a code-mixed variety, exhibits relatively strong performance across most models, with several systems achieving accuracies comparable to or higher than on English. This suggests that code-mixing alone does not necessarily degrade performance, especially when the dominant language is English.

4.2 Audio Ablations and Analysis

To systematically isolate the informational value of each modality and evaluate the models' causal reliance on context, we conduct ablations with the modality of the contextual input. Specifically, we evaluate the top-performing open-source model, `Qwen3-Omni-Flash`, across three distinct settings, while keeping the query in the source language:

- *Audio*: The model is given the raw acoustic signal to solve GLOBEAUDIO questions, serv-

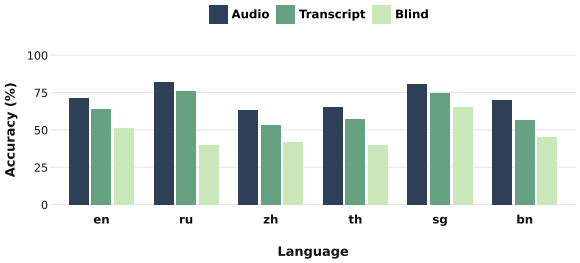

Figure 3: Accuracy of `Qwen3-Omni-Flash` on various settings, from Audio and Transcript input, to "Blind" which denotes model without audio or transcript input.

ing as our primary evaluation pipeline.

- *Transcript*: To decouple linguistic semantics from acoustic nuances, the audio input is removed and reduced to its text transcript.
- *Blind*: To detect guessing heuristics and reliance on prior world knowledge, the model is made to answer questions with zero supporting context, serving as a baseline.

Reliance on Audio When the model is subjected to the *Blind* setting, we observe a sharp performance drop across all languages, with absolute accuracy falling to an average of 47.20%, yielding a sharp average decrease of 24.91%.

This sharp degradation demonstrates the close coupling between provided auditory contexts and GLOBEAUDIO's expertly crafted questions, and that crucially, internal world knowledge and query-side shortcuts are fundamentally insufficient to perform well without its audio samples.

Embedded prosodic and acoustic cues In the *Transcript* setting, where the context is reduced strictly to text, we observe a marked and consistent performance decay across all linguistic families, resulting in an average absolute decrease of 8.69% relative to the *Audio* setting. Since the underlying semantic content remains identical between these two settings, this marked performance delta is likely attributed to the utility of the embedded prosodic, and paralinguistic cues native to our raw audio samples than to textual variance.

This validates that transcripts act as a lossy compression that flattens critical prosodic and paralinguistic cues. Furthermore, we directly demonstrate that GLOBEAUDIO rigorously assesses an LALM's capacity to parse and understand messy, non-linguistic acoustic signals present in our dataset without reliance on linguistic patterns. However, we note that this degradation represents a conservative lower bound due to the presence of culturally grounded questions which inherently lean on text-retrievable context to answer.

4.3 Cross-lingual Comprehension Analysis

To investigate LALMs' cross-lingual comprehension capabilities, we aim to assess the model's reliance on language-specific cues when processing complex auditory context.

Linguistic Fidelity To probe the depth of the linguistic cues present within GLOBEAUDIO, we evaluate `Qwen3-Omni-Flash` using original source-language questions, comparing performance against English-translated counterparts. This allows us to quantify the information loss when language-specific nuances are subjected to translation.

We observe that their performance dipped slightly when subjected to English questions, by an average of 2.11% with the exception of Bengali which performed better on English questions by 4.17%. Despite framing the question in English, a high resource language, we discover that models still generally perform better in source languages.

This is possibly because GLOBEAUDIO's human-annotated questions contain dense, language-specific nuances that facilitate precise alignment with the provided natural audio. These nuances were likely irreversibly flattened during translation, depriving the model of critical grounding cues. On the other hand, the models' robust English capabilities might have mitigated the performance gap.

Conversely, the performance gain in Bengali, a relatively much lower resource language, suggests the model's possible reliance on English grounding when native-language depth is limited. This underscores the high-resource complexity of other language sets, where the original linguistic structure consistently outperforms translated alternatives.

Contextual Nuance To validate our findings by further analysing language-specific nuance present in GLOBEAUDIO, we conduct a parallel experiment on the audio context and isolate linguistic from prosodic cues using Qwen3 LLM's cascade pipeline, which is particularly strong in multilingual understanding, and keep questions in their source language. We then compare its performance given source language transcripts with their

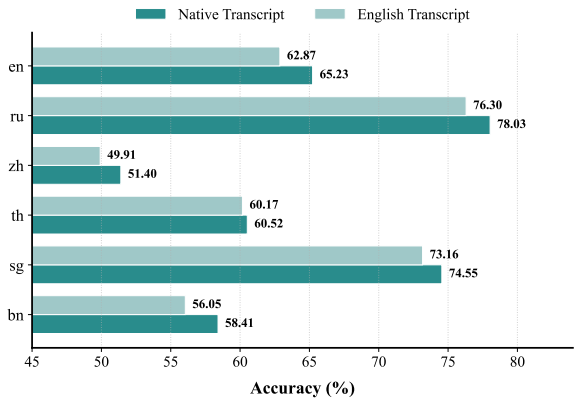

Figure 4: Analysis on varying language of transcripts in Cascade pipelines. We report average accuracy across all languages.

English-translated counterparts and visualise results in Figure 4.

Concretely, we observe a similar consistent cross-lingual performance decline which validates our previous findings, and confirms that models struggle primarily with the loss of intrinsic linguistic and cultural structures in GLOBEAUDIO. These elements are non-fungible across translations, suggesting that the performance gap is driven by a loss of semantic depth present in our dataset, rather than purely by the absence of prosodic information demonstrated by earlier ablations.

5 Related Work

Scattered Evaluation Efforts of LALMs The evaluations of LALMs has made considerable progress across diverse areas in recent years. IFEval-Audio (Gao et al., 2025b) is an audio benchmark assessing instruction-following across 6 diverse dimensions. AHa-Bench (Cheng et al., 2025) is a comprehensive benchmark for evaluating auditory hallucinations in LALMs to enhance their robustness and stability. SpeechR (Yang et al., 2025b) is a speech-based benchmark assessing LALMs' reasoning capabilities. To assess complex reasoning across speech, environmental sounds and music, MMAU (Sakshi et al., 2024) and AIR-Bench (Yang et al., 2024) was introduced. As a step towards evaluating under *in-the-Wild* conditions, WildSpeech-Bench (Zhang et al., 2025) was introduced to assess auditory capabilities through synthetic mixtures of diverse speaker types and background noise. ProsodyEval (Yang et al., 2026) is a benchmark introduced to assess ability of LALMs in capturing prosodic variations.

Universal Evaluation Efforts To unify such scattered evaluation efforts to keep up with the rising adoption of LALMs, universal benchmarks aim to provide wide coverage across a diverse range of tasks. AudioBench (Wang et al., 2025) was introduced to evaluate LALMs over 8 distinct tasks and 26 datasets, targeting 3 main aspects, namely speech understanding, audio scene understanding and voice understanding. On the other hand, MSEB (Heigold et al., 2025) was introduced to cover diverse categories of tasks, including Retrieval, Reranking, Reasoning and Classification.

Multilingual Evaluation Recognising the linguistic diversity, evaluation efforts have also begun to extend English-centric settings. Fleurs-SLU (Schmidt et al., 2025) examines 2 task types, namely speech classification across a wide range of languages. mSTEB (Beyene et al., 2025) assesses LALMs on 5 speech-based tasks and 5 text-based tasks, across a broad range of languages. Towards *in-the-Wild* evaluation, ADU-Bench (Gao et al., 2025a) is an audio dataset assessing 12 skills in 3 general scenarios, across 9 languages.

Despite these efforts, audio evaluation for LALMs remains substantially under-specified relative to real-world requirements, due to the lack of acoustic realism, and true linguistic and cultural authenticity.

6 Conclusion

In this work, we introduce GLOBEAUDIO, a comprehensive high-fidelity audio benchmark designed to realistically evaluate LALMs under real-world conditions. Amidst the recent proliferation of audio evaluation suites, GLOBEAUDIO establishes as a standardized, unified benchmark that embraces and rigorously examines LALMs' ability to handle the acoustic complexity and sociolinguistic diversity of in-the-wild contexts. Through our extensive experiments and analysis on state-of-the-art LALMs, we demonstrate that GLOBEAUDIO contains multiple key characteristics including embedded prosodic and acoustic cues, and is well-equipped to rigorously examine the lingual and cultural understanding of LALMs. Ultimately, by highlighting the current vulnerabilities of these models, we establish a foundation for the community to build inclusive, highly performant LALMs suited for complex, real-world environments.

7 Limitations

Language Coverage In our work, we perform language selection to maximise coverage of diverse language families, geographic regions, and ensuring substantial variation in resource availability. However, we acknowledge this might not fully capture all linguistic diversity or account for the full range of resource conditions across languages. As such, we plan to expand GLOBEAUDIO's language coverage to better capture global linguistic diversity.

Data distribution The resulting distribution after data sourcing is inherently shaped by platform-side filtering. Since unavailable or private videos cannot be indexed, the distribution of our obtained data might be skewed towards content that adhere to platform guidelines, potentially underrepresenting more controversial sources and media.

To mitigate the latter's impact, we (1) use the `tv_simply` player client for better reach and (2) query across a wide set of keyword templates.

8 Ethics

Privacy All audio data used in this work are derived from publicly available YouTube videos, which may contain personal information about content creators or individuals mentioned in the recordings.

To mitigate these risks, we (1) use the data exclusively for non-commercial research purposes, (2) report only aggregated statistics such as clip count and audio length. Furthermore, we do not include any metadata in our dataset to omit identifying information such as creator name, channel name or YouTube video title.

Licensing Licensing information for YouTube videos is inconsistently exposed via public metadata. We explicitly queried license metadata for each YouTube video using `yt-dlp` and report unavailable or "NA" values in the license field for many videos, indicating that such information was not exposed by YouTube.

As such, we (1) make no assumptions about reuse rights, (2) treat all content as copyrighted.

Annotators All our annotators are undergraduate students across diverse majors and were recruited as student helpers under the school. They are fully agreeable to the scope of the project and rates and will be compensated through payment detailed under the following structure:

- English, Chinese, Singlish: US$8.58 per hour
- Thai, Russian: US$10.14 per hour

We first perform annotation ourselves and provide payment corresponding to 4 mins per clip as a conservative rate.

9 Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its NRF Fellowship (Award No: NRF-NRFF18-2026-0012) and by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-016).

References

Firoj Alam, Ali Ezzat Shahroor, Md. Arid Hasan, Zien Sheikh Ali, Hunzalah Hassan Bhatti, Mohamed Bayan Kmainasi, Shammur Absar Chowdhury, Basel Mousi, Fahim Dalvi, Nadir Durrani, and Natasa Milic-Frayling. 2025. Everydaymmqa: A multilingual and multimodal framework for culturally grounded spoken visual qa. *Preprint*, arXiv:2510.06371.

Luel Hagos Beyene, Vivek Verma, Min Ma, Jesujoba O. Alabi, Fabian David Schmidt, Joyce Nakatumba-Nabende, and David Ifeoluwa Adelani. 2025. msteb: Massively multilingual evaluation of llms on speech and text tasks. *Preprint*, arXiv:2506.08400.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. 2025. AHabench: Benchmarking audio hallucinations in large audio-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, Daan Van Esch, Vera Axelrod, Simran Khanuja, Jonathan H. Clark, Orhan Firat, Michael Auli, Sebastian Ruder, Jason Riesa, and Melvin Johnson. 2022. Xtreme-s: Evaluating cross-lingual speech representations. *Preprint*, arXiv:2203.10752.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. In *Advances in Neural Information Processing Systems*, volume 36, pages 18090–18108. Curran Associates, Inc.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. LLaMA-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations*.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2025a. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4763–4784, Vienna, Austria. Association for Computational Linguistics.

Yiming Gao, Bin Wang, Chengwei Wei, Shuo Sun, and AiTi Aw. 2025b. Ifeval-audio: Benchmarking instruction-following capability in audio-based large language models. *Preprint*, arXiv:2505.16774.

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.

Google. 2025a. Gemini 2.0 Flash Model Card. `https://modelcards.withgoogle.com/assets/documents/gemini-2-flash.pdf`. [Online].

Google. 2025b. Gemma 3n Model Overview. `https://ai.google.dev/gemma/docs/gemma-3n`. [Online].

Google. 2026. Gemini 3.1 Pro Model Card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf`. [Online].

Peize He, Zichen Wen, Yubo Wang, Yuxuan Wang, Xiaoqian Liu, Jiajie Huang, Zehui Lei, Zhuangcheng Gu, Xiangqi Jin, Jiabing Yang, Kai Li, Zhifei Liu, Weijia Li, Cunxiang Wang, Conghui He, and Linfeng Zhang. 2025. Audiomarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio llms. *Preprint*, arXiv:2510.07293.

Georg Heigold, Ehsan Variani, Tom Bagby, Cyril Allauzen, Ji Ma, Shankar Kumar, and Michael Riley. 2025. Massive sound embedding benchmark (MSEB). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shuanghong Huang, Jinlei Xu, Youchao Zhou, Yanghao Zhou, Xuan Zhao, Chong Feng, and Wenxuan Zhang. 2026. Pardon? evaluating conversational repair in large audio-language models. *Preprint*, arXiv:2601.12973.

Evan King, Adam Sabra, Manjunath Kudlur, James Wang, and Pete Warden. 2025. Flavors of moonshine: Tiny specialized asr models for edge devices. *Preprint*, arXiv:2509.02523.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Beomseok Lee, Ioan Calapodescu, Marco Gaido, Matteo Negri, and Laurent Besacier. 2024. Speechmassive: A multilingual speech dataset for slu and beyond. *Preprint*, arXiv:2408.03900.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025a. Voxtral. *Preprint*, arXiv:2507.13264.

Chaoqun Liu, Mahani Aljunied, Guizhen Chen, Hou Pong Chan, Weiwen Xu, Yu Rong, and Wenxuan Zhang. 2025b. Seallms-audio: Large audio-language models for southeast asia. *Preprint*, arXiv:2511.01670.

MERaLiON Team. 2024. Meralion-audiollm: Bridging audio and language with large language models. *Preprint*, arXiv:2412.09818.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

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. *Preprint*, arXiv:2410.19168.

Fabian David Schmidt, Ivan Vulić, Goran Glavaš, and David Ifeoluwa Adelani. 2025. Fleurs-slu: A massively multilingual benchmark for spoken language understanding. *Preprint*, arXiv:2501.06117.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast. *Preprint*, arXiv:2509.14128.

Syed Abdul Gaffar Shakhadri, Kruthika KR, and Kartik Basavaraj Angadi. 2025. Samba-asr: State-of-the-art speech recognition leveraging structured state-space models. *Preprint*, arXiv:2501.02832.

Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. 2025. Audio-language models for audio-centric tasks: A survey. *Preprint*, arXiv:2501.15177.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

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. *Preprint*, arXiv:2406.16020.

Xiaomi. 2026. Xiaomi MiMo V2.5. https://mimo.xiaomi.com/mimo-v2-5/. [Online].

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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. *Preprint*, arXiv:2402.07729.

Wanqi Yang, Yanda Li, Yunchao Wei, Meng Fang, and Ling Chen. 2025b. Speechr: A benchmark for speech reasoning in large audio-language models. *Preprint*, arXiv:2508.02018.

Yifan Yang, Bing Han, Hui Wang, Long Zhou, Wei Wang, Mingyu Cui, Xu Tan, and Xie Chen. 2026. Measuring prosody diversity in zero-shot tts: A new metric, benchmark, and exploration. *Preprint*, arXiv:2509.19928.

Han Yin and Jung-Woo Choi. 2025. Can large audio language models understand audio well? speech, scene and events understanding benchmark for lalms. *Preprint*, arXiv:2509.13148.

Linhao Zhang, Jian Zhang, Bokai Lei, Chuhan Wu, Aiwei Liu, Wei Jia, and Xiao Zhou. 2025. Wildspeechbench: Benchmarking end-to-end speechllms in the wild. *Preprint*, arXiv:2506.21875.

A Appendix

A.1 Prompts used in Inference

Overview Across all experiments, we employ a unified prompting template to ensure consistency and comparability across models. Specifically, each model is prompted with the instruction: *"Please answer this multiple-choice question: <question>"*. followed by the four answer options labeled *A*, *B*, *C*, and *D*. To constrain model outputs and avoid verbose generations, we append the explicit instruction: *"Do NOT include any explanation. ONLY output a single character: A, B, C, or D."* Finally, we include the cue *"Answer:"* to elicit the model's response for evaluation.

Prompts for Open and Closed-Source Models Refer to Figure 5.

Prompts for Cascade Pipelines Refer to Figure 6.

A.2 Annotation Guidelines

A.2.1 Annotator Guidelines

Objective

- Craft 1 multiple-choice question for each corresponding audio clip which adheres to the guidelines below.
- Write this question in column 'question'.

How to Annotate

- Provide 4 corresponding options in columns 'option1', 'option2', 'option3', 'option4'. Indicate the correct answer in column 'target.

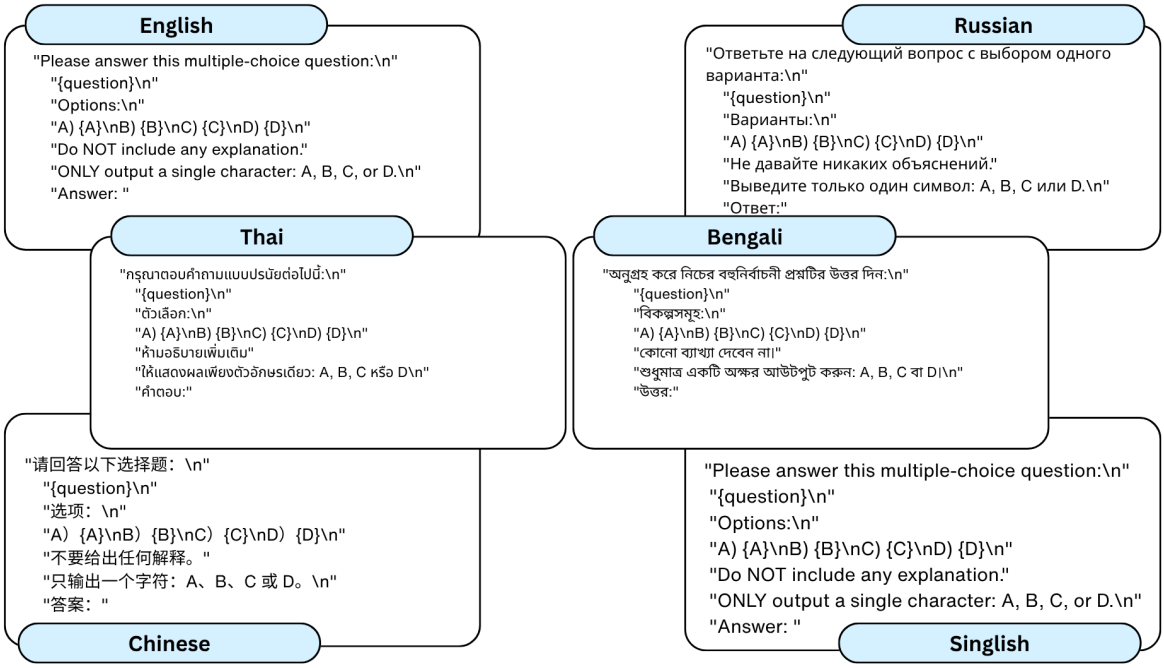

Figure 5: Prompts for Open and Closed-Source Models in various languages

- If there are no meaningful/challenging questions that can be asked about the audio clip, do not annotate and indicate NA in column 'target'.

Question guidelines

- **Please do not use AI in this task**
- Question must be longer than 5 words
- It is preferred that the question is asked about audio features, eg. such as and not limited to background noise, sentiment, mood, tone, location. This means it should be difficult to answer with just the audio's transcript.
- It is preferred that the question stimulates several layers of thinking before being able to conclude on the final answer (eg. What do you think the man is feeling? -> Man is laughing -> Talking about mental health issues -> Awkward pauses might indicate struggles -> Conclude he is feeling depressed)

Options guidelines

- It is preferred that distractor options are included to increase difficulty
- Be creative with distractors!
- You may distract with tone eg. enthusiastic tone and confuse with happiness or crying/sobbing while speaking about happy stories
- You may also distract with false options while repeating some keywords/context in the audio clip
- Or anything else that is sensible to you

A.2.2 Meta-annotator Guidelines

We use the following set of guidelines for cross-answering and cross-checking during the midpoint and final review:

- Check if the question is clear and understandable.
- Verify that the question aligns with the passage content—it should not introduce new facts or concepts.
- The answer to your question should be able to be found in the corresponding audio clip.
- Confirm that the correct answer is explicitly or implicitly supported by the passage.
- Ensure distractor options are plausible but clearly incorrect.
- Questions should require understanding, inference, or reflection, not merely factual recall (unless the task specifies fact-based questions).

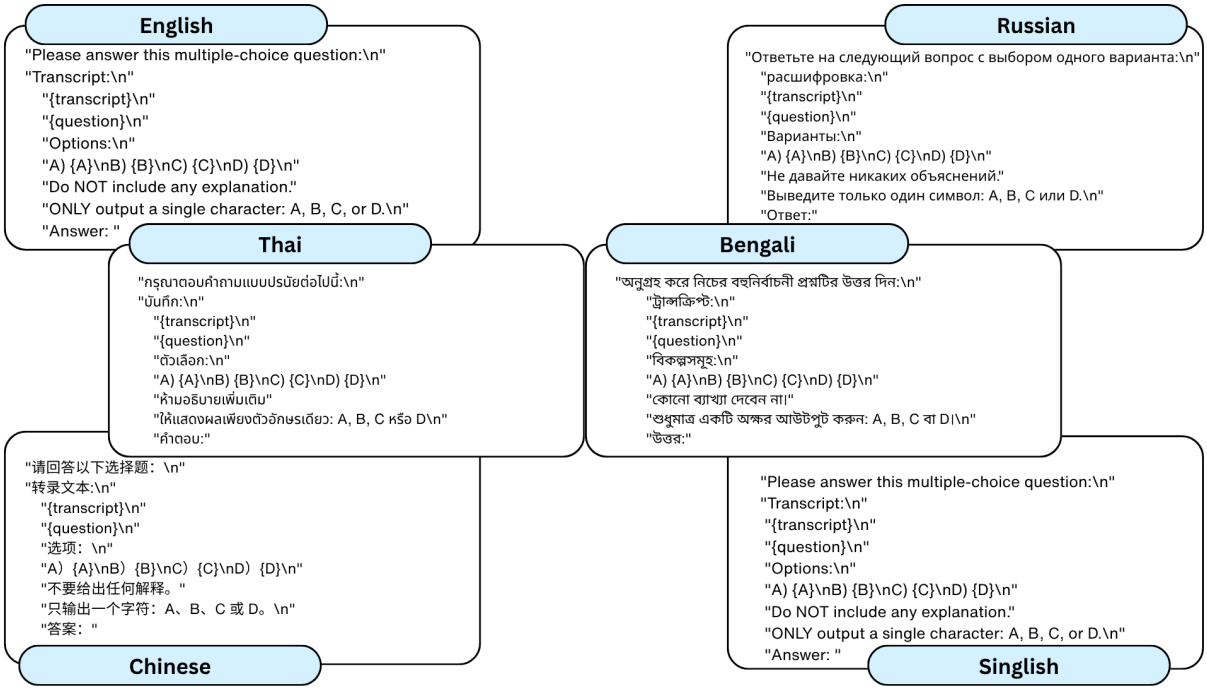

Figure 6: Prompts for Cascade pipelines in various languages

A.3 GLOBEAUDIO's Question Examples of each language

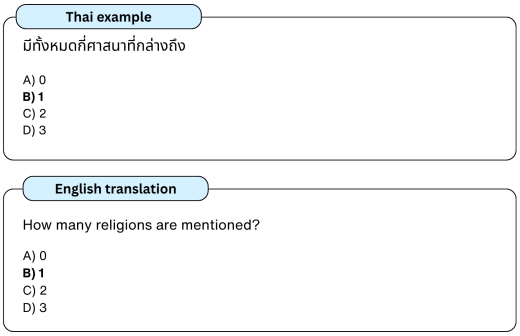

Figure 7: An example of a Thai question from GLOBEAUDIO

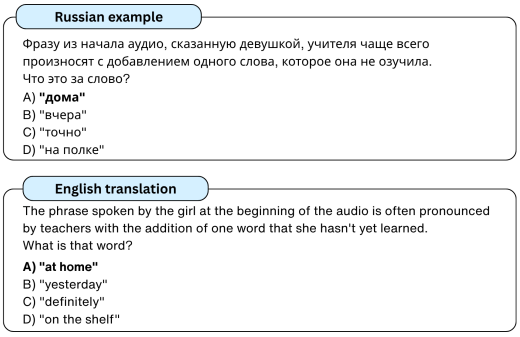

Figure 9: An example of a Russian question from GLOBEAUDIO

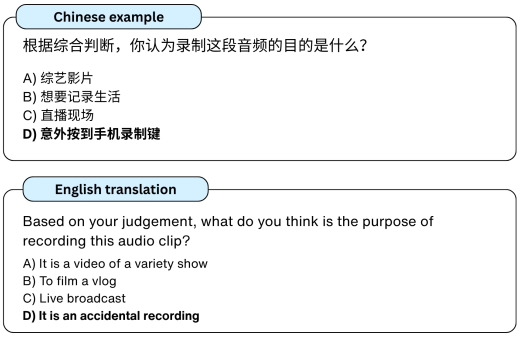

Figure 8: An example of a Chinese question from GLOBEAUDIO

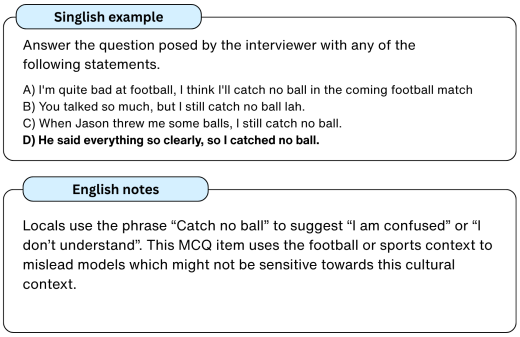

Figure 10: An example of a Singlish question from GLOBEAUDIO

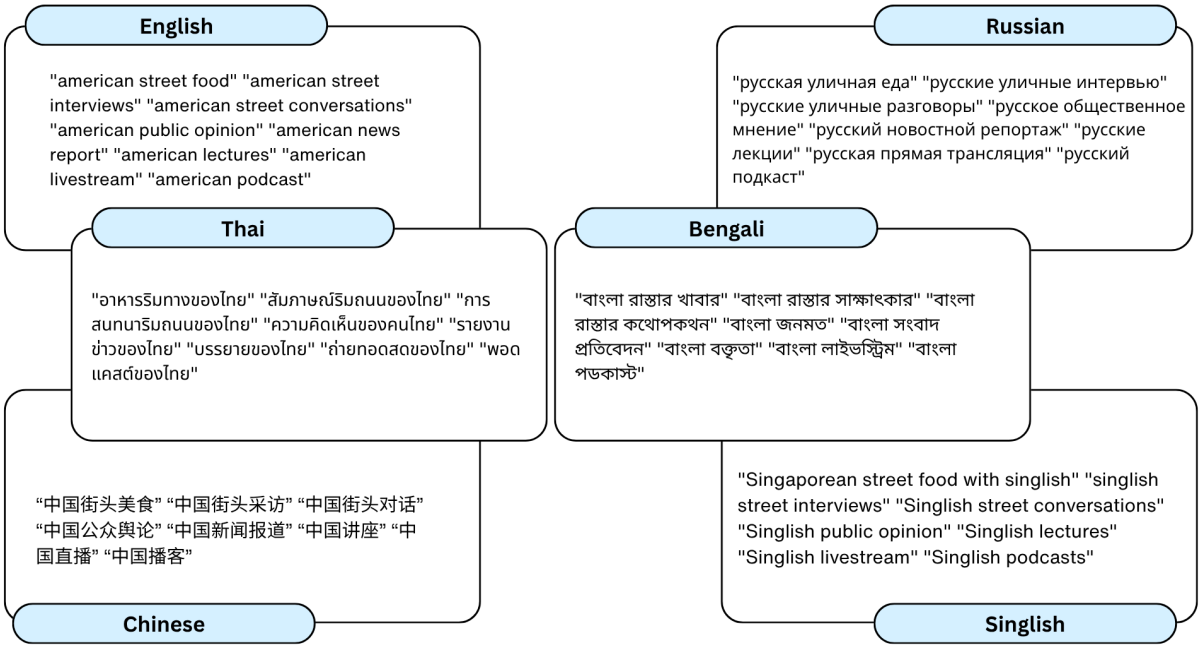

Figure 12: YouTube downloader query terms

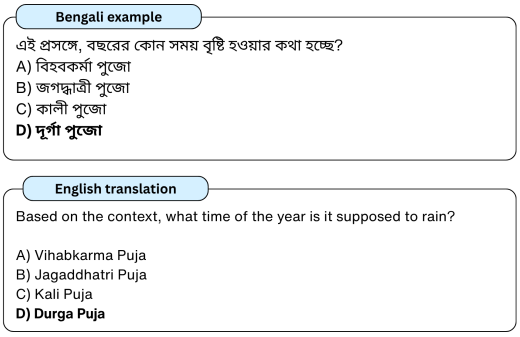

Figure 11: An example of a Bengali question from GLOBEAUDIO

A.4 `yt-dlp` Commands and Parameters

Our query terms are detailed in Figure 12.

We use the following commands on `yt-dlp` to source our audio clips from YouTube. For each query term as detailed in Figure 12, we run:

yt-dlp "ytsearch15:$term" –get-id | sed 's_^_https://www.youtube.com/watch?v=_' » <lang>_urls.txt

Then, we run this command to extract the audio from the URLs gathered in the .txt file and ensure we download the version with the best audio quality:

<lang>_urls.txt | xargs -n 1 -P 8 yt-dlp -x –extractor-args "youtube:player-client=default,-tv_simply" –audio-format mp3 –embed-metadata –add-metadata –write-info-json -f bestaudio -o "%(title)s.%(ext)s"